Center for Evaluation and Development

European Commission

Session 1: Recap of Y1 (CIE Methods) and Y2 (Data Collection)

C4ED – EUTF

October 2023

Welcome to the
Training Workshop on Counterfactual Impact Evaluation (CIE)

The material of this workshop was produced with the financial support of the European Union. Its contents are the sole responsibility of C4ED and do not necessarily reflect the views of the European Union

Introduction

Communication during the training

MUTE BUTTON

QUESTIONS

FEEDBACK

Center for Evaluation and Development

Communication during the training

MUTE BUTTON

QUESTIONS

FEEDBACK

Center for Evaluation and Development

Communication during the training

European Commission

MUTE BUTTON **QUESTIONS** **FEEDBACK**

Center for Evaluation and Development

Asking Questions

European Commission

- Please post your questions in the chat room
- Like 👍 the questions of others, so we know they are particularly relevant for you as well
- Carolin will read out all questions and we will answer these at once
- Use the longer breaks to ask more questions

Communication during the training

MUTE BUTTON

QUESTIONS

FEEDBACK

Asking Questions

- Please make suggestions
- Feel free to share your comments
- More feedback and questions (especially for the Q&A session):

Center for Evaluation and Development

Day 1 Agenda

European Commission

10:00 – 10:30	Welcome and introduction
10:30 – 11:15	Session 1: Recap of previous years – CIE Methods (Year 1) and Data Collection for CIE (Year 2)
11:15 – 11:25	Break (10 minutes)
11:25 – 12:10	Session 2: Descriptive statistics for monitoring and answering evaluation questions on effectiveness
12:10 – 12:20	Break (10 minutes)
12:20 – 12:35	Session 2 – Continued
12:35 – 13:00	Interactive Quiz
13:00 – 14:00	Lunch (60 minutes)
14:00 – 14:45	Session 3a: Statistical testing in CIE and answering evaluation questions on impact
14:45 – 15:00	Break (15 minutes)
15:05 – 15:15	Session 3a – continued
15:15 – 15:35	Session 3b: Guided walkthrough of an example t-test in Excel
15:35 – 15:50	Q&A
15:50 – 16:00	Closing Day 1

Center for Evaluation and Development

Overview of Day 1

European Commission

- First, we will briefly review the *basics of Counterfactual Impact Evaluation (CIE)*, common methods of **identifying impact**, and the importance of *high-quality data for CIE*
- The next session will focus on *basic descriptive statistics* – how to calculate, present and interpret them
- Finally, the last topic of the day will be *statistical testing*

- We will share useful external resources and case studies on CIE
- https://europa.eu/capacity4dev/

Session 1: Recap of CIE Methods and Data Collection for CIE

C4ED – EUTF

October 2023

Center for Evaluation and Development

European Commission

RECAP YEAR 1

Counterfactual Impact Evaluation (CIE) Methods

Recap Year 1 – Objectives

- Review the “why”, “what” and “how” of Counterfactual Impact Evaluation (CIE)

- Review the intuition of experimental evaluation methods → Randomized Controlled Trials

- (Briefly) Review the intuition of two important quasi-experimental methods (matching and difference-in-differences)

Center for Evaluation and Development

CIE Recap

European Commission

Why do a counterfactual impact evaluation?

- To determine whether an intervention creates an attributable, causal change in the outcome, how (the causal mechanism) and to what magnitude
- To learn which intervention strategy works best
- To help make evidence-based decisions

What is a counterfactual impact evaluation?

- **Impact**: the effect on outcomes of interest that the program/policy directly *causes* and that can be directly *attributed* to the program
- **Counterfactual**: the outcome that would have been observed/measured for program beneficiaries had they *not* received program.

→Fundamental problem: it is impossible to measure or observe the counterfactual

→program targets either receive the program or not, we cannot observe them in both scenarios at the same time

→Solution: use a control/comparison group to mimic the counterfactual

Center for Evaluation and Development

CIE Recap

European Commission

How is a CIE designed?

Goal: Mimicking the counterfactual situation with a comparison group

- The comparison group:
  - Has the same characteristics (on average) as the treatment group
  - Is not exposed to the program
  - Would react similarly to the program as the treatment group (if it were to participate)
- Based on the intervention design and context, timeline, data availability and budget, the most appropriate approach to use is selected:
  - Experimental methods
  - Quasi-experimental methods

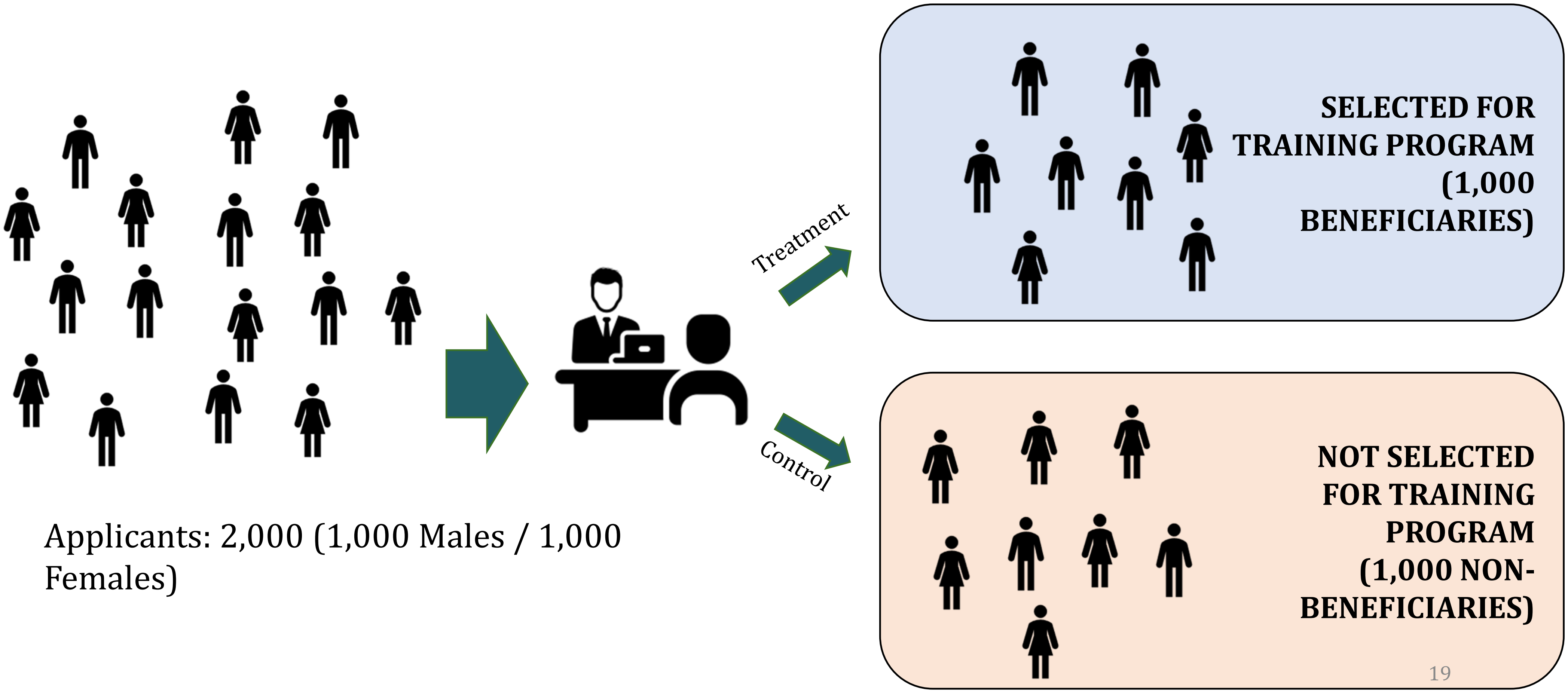
Center for Evaluation and Development

How to simulate a counterfactual

- In the following example we consider the selection for a youth vocational training program aimed to improve employment outcomes.
- Here, applicants are invited to take part in short interviews to discuss their application.
- Based on their application and interview, applicants are either selected to take part in the vocational training program, or not.

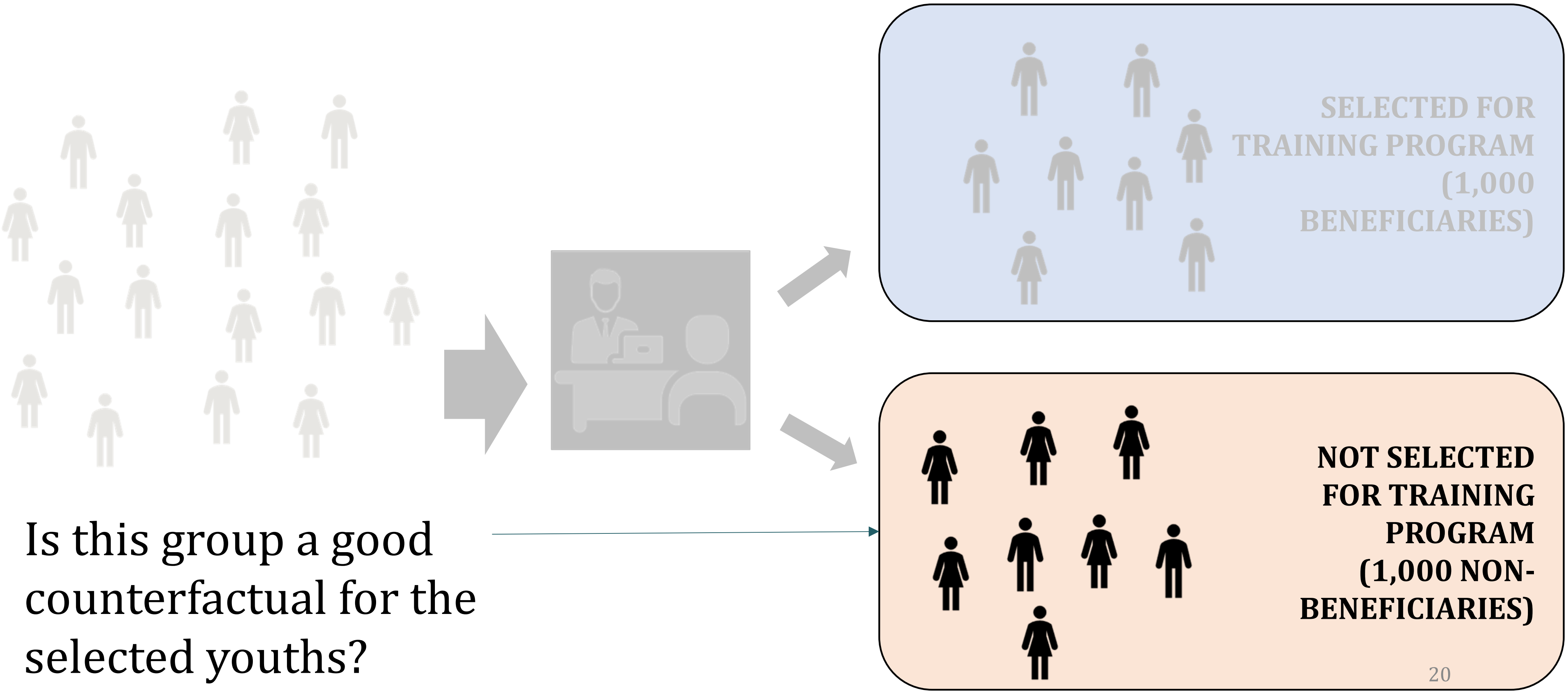
Simulating a counterfactual

Example: Selection for **youth vocational training program**



Simulating a counterfactual

Example: Selection for **youth vocational training program**



Is this group a good counterfactual for the selected youths?

Center for Evaluation and Development

Sources of differences between groups

European Commission

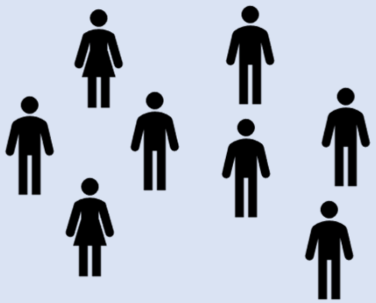
Do rejected applicants constitute a good comparison group/counterfactual?

- It is likely that ‘stronger’ applicants are chosen
  - Counterfactual is composed of applicants that are “weaker” and likely to be quite different from those that are not.
- For a program manager this may be desirable to have the most able applicants enrol in the program
- However, for an evaluator seeking to measure the impact of the program, these differences will mean that the groups are not easily comparable

- **Given the selection process, they would not represent a good counterfactual. Why?**

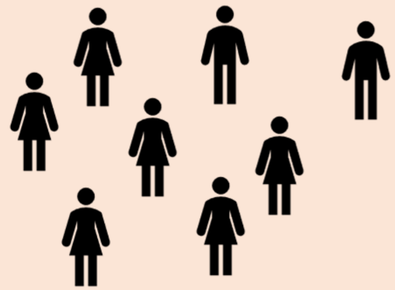
Group comparison
Observed characteristics

- Assume outcome variable of interest is employment, or income.
- Differences in outcomes between groups with different characteristics may not be *attributed* to the program with 100% certainty.



SELECTED FOR TRAINING PROGRAM

Variable	Average
Age	30
Years of schooling	10
Previous employment	60%
Parent income	5,000



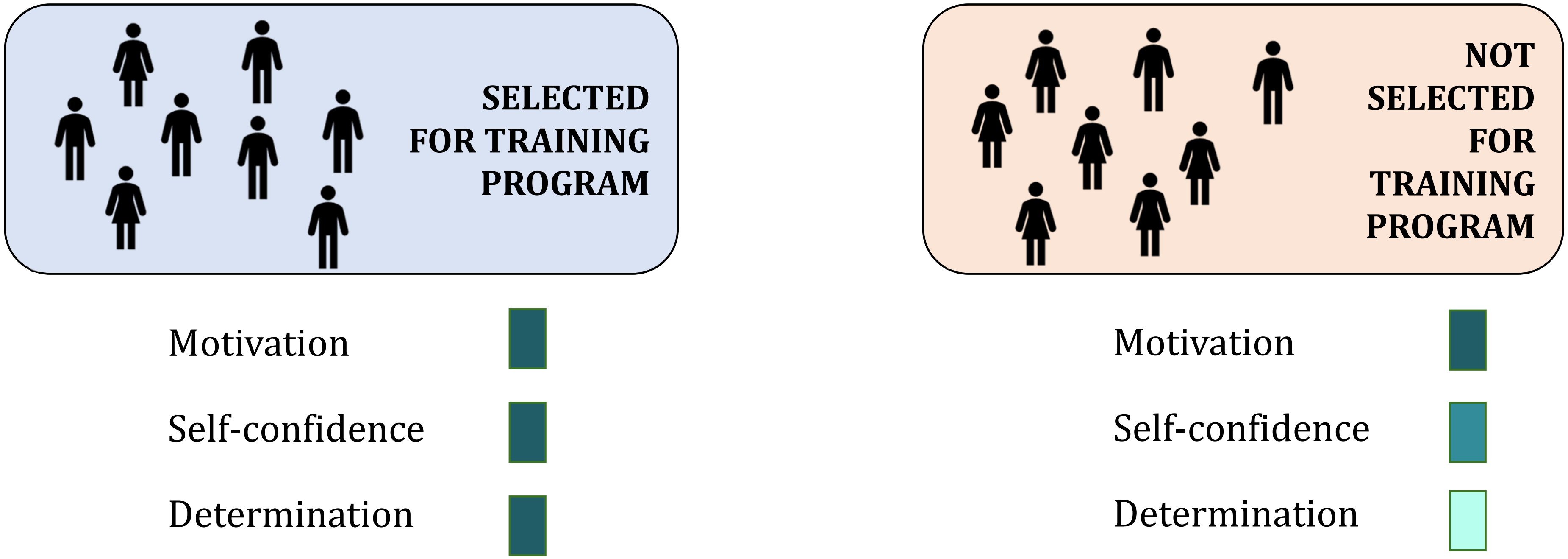
NOT SELECTED FOR TRAINING PROGRAM

Variable	Average
Age	24
Years of schooling	6
Previous employment	46%
Parent income	3,400

Center for Evaluation and Development

Group comparison
Unobserved characteristics

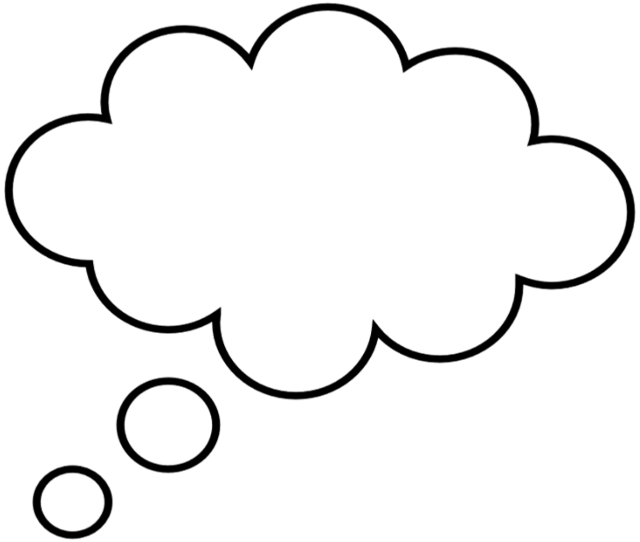
→ Even if we control for observable differences, how to account for unobservable differences?



Counterfactual Selection

- In the previous example – simply using all rejected applicants is a poor counterfactual

- How could you design the process differently to be able to use rejected applicants as a counterfactual?



Randomized Controlled Trial

Center for Evaluation and Development

Randomized Assignment

European Commission

- If the evaluation is integrated into program implementation, you could create a counterfactual by using a lottery to decide who is selected

→This is called **randomized assignment**

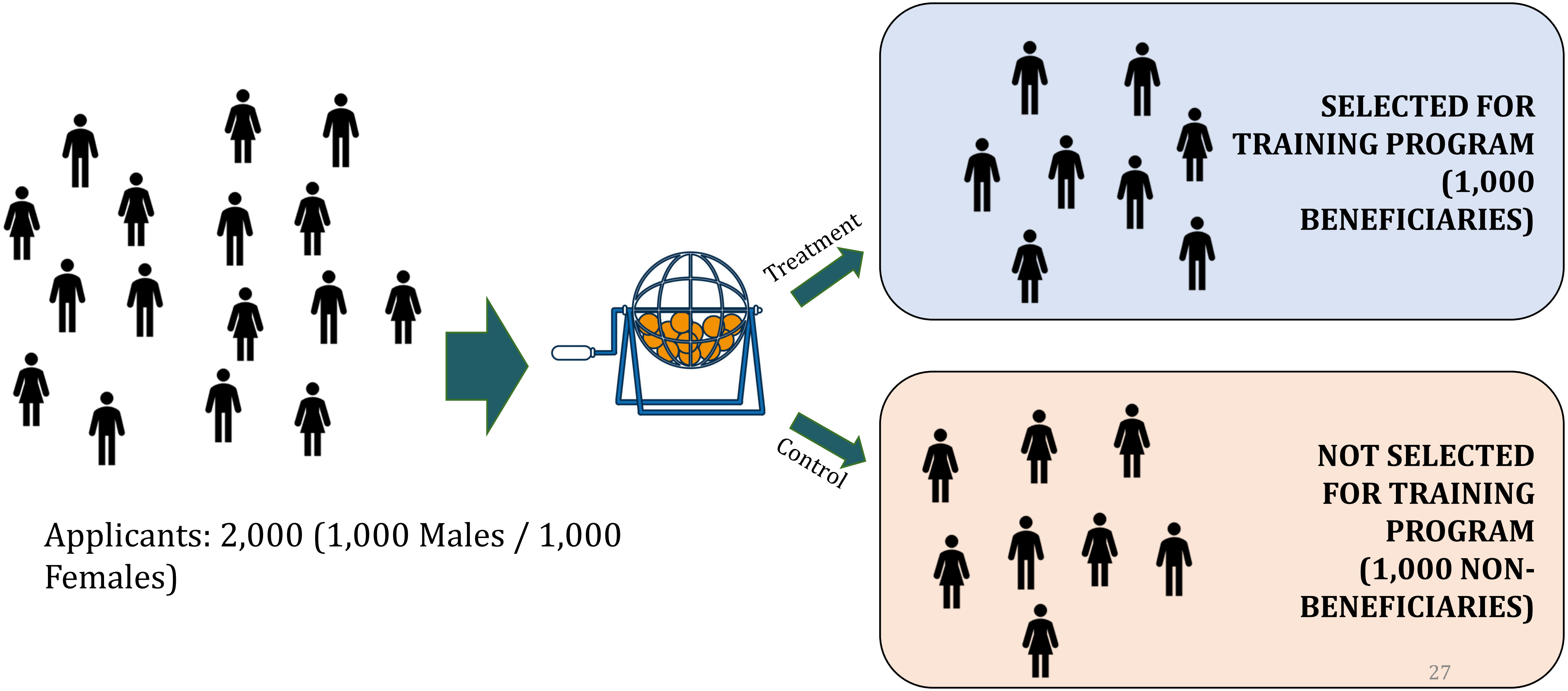
- Why does this help?

→Assuming the number of applicants is large enough, the two randomly assigned groups will be similar (on average)

Center for Evaluation and Development

Simulating a counterfactual

Example: Selection for **youth vocational training program**



Applicants: 2,000 (1,000 Males / 1,000 Females)

Center for Evaluation and Development

Group comparison – Randomized Assignment

Observed characteristics

European Commission

SELECTED FOR TRAINING PROGRAM = TREATMENT GROUP

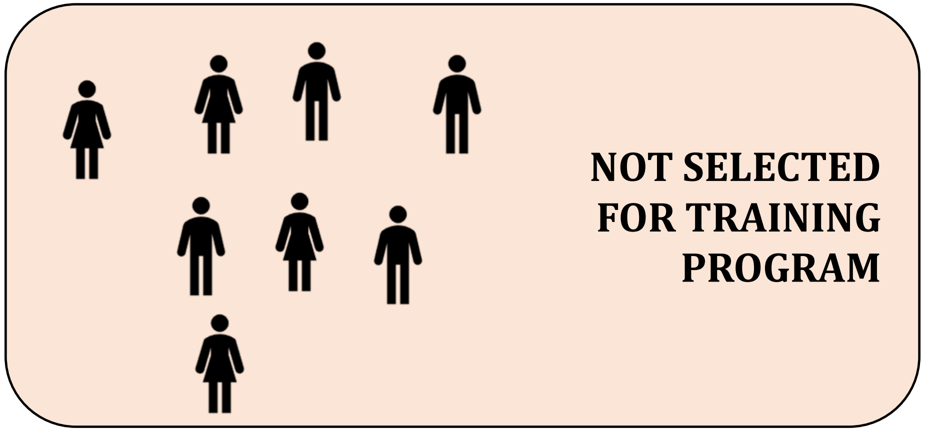
Variable	Average
Age	27
Years of schooling	8
Previous Employment	52%
Parent Income	4,300

NOT SELECTED FOR TRAINING PROGRAM = CONTROL GROUP

Variable	Average
Age	27
Years of schooling	8
Previous Employment	54%
Parent Income	4,100

Group comparison – Randomized Assignment

Unobserved characteristics



Motivation

Self-confidence

Determination

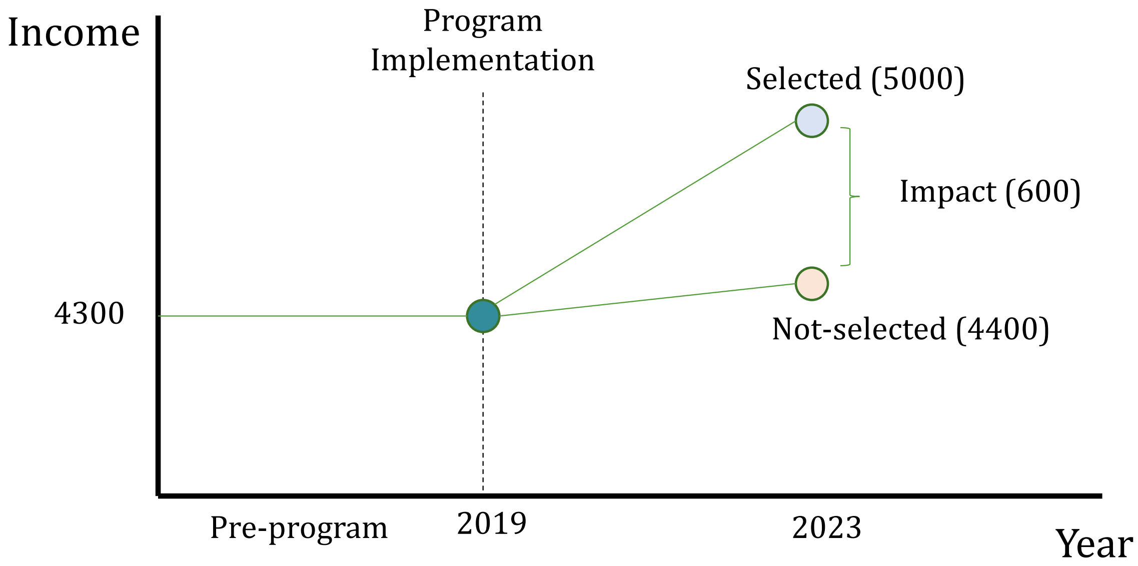
Motivation

Self-confidence

Determination

Measuring Impact in an RCT

- Because randomized assignment creates two groups that are (on average) comparable at the beginning of the program, ***impact*** can be measured simply as the difference in the outcome after the program.
- In other words, differences in outcomes between groups can be *attributed* to the program (**because** all other characteristics are similar between groups)



Center for Evaluation and Development

Randomization – What, when, how

What to randomize?

- Any aspect of the program that the implementation team *fully controls*
- Often requires creativity and a thorough knowledge of the program → plan ahead!

When to randomize?

- *Before* program starts, must be included as part of program implementation

How to randomize?

- Simple lottery
- Multiple treatment arms → can test different treatment modalities
- Phase-in → delayed treatment for part of program beneficiaries
- Encouragement → all have access to program, but some beneficiaries are actively encouraged to participate

Center for Evaluation and Development

Quasi-experimental methods

European Commission

- Random program assignment is sometimes not possible:
  - Randomization may not be socially or politically acceptable
  - Randomization may not be feasible
  - CIE is designed only after implementation starts

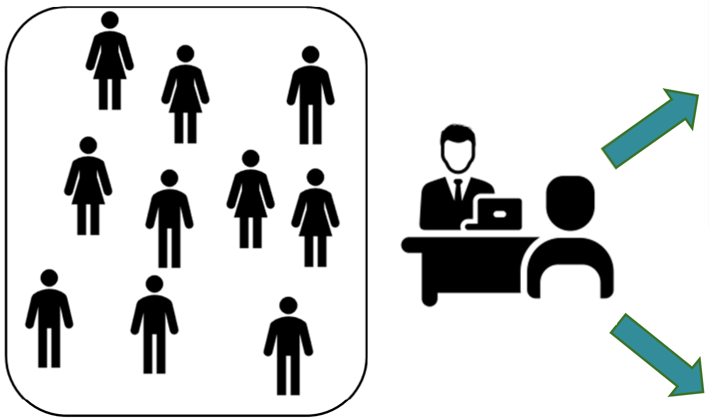
- May be possible to use **quasi-experimental methods** to construct counterfactual.
- Here we present the basic intuition of two so-called quasi-experimental methods, namely **matching** and **difference-in-differences**.

Matching

Non-random selection, the groups are different

Center for Evaluation and Development

Number of Applicants: 2,000 (1,000 Males / 1,000 Females)



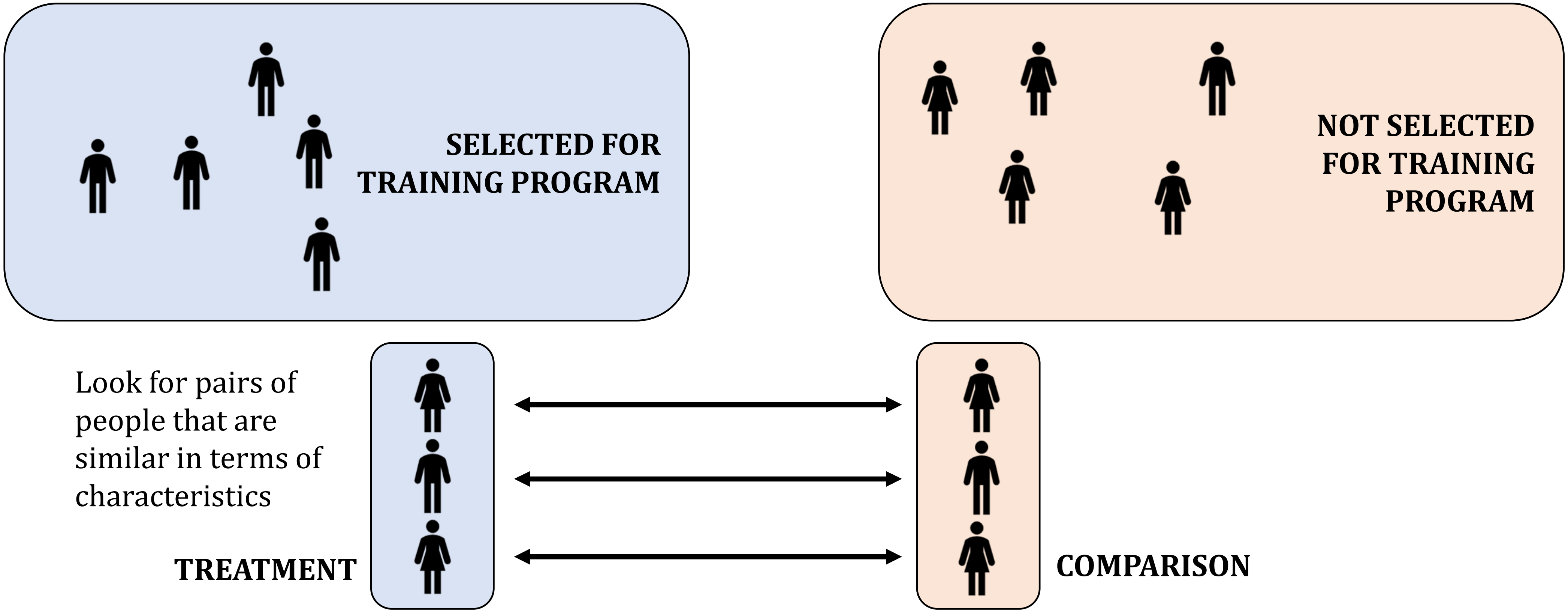
SELECTED FOR TRAINING PROGRAM (1,000 BENEFICIARIES)

Variable	Average
Age	30
Years of schooling	10
Previous employment	60%
Parent income	5,000

NOT SELECTED FOR TRAINING PROGRAM (1,000 NON-BENEFICIARIES)

Variable	Average
Age	24
Years of schooling	6
Previous employment	46%
Parent income	3,400

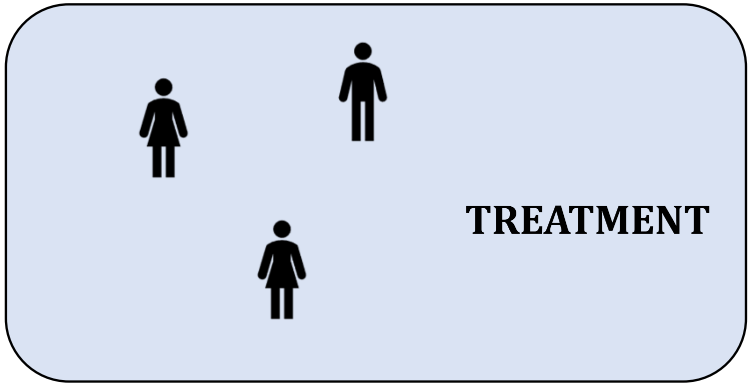
Matching



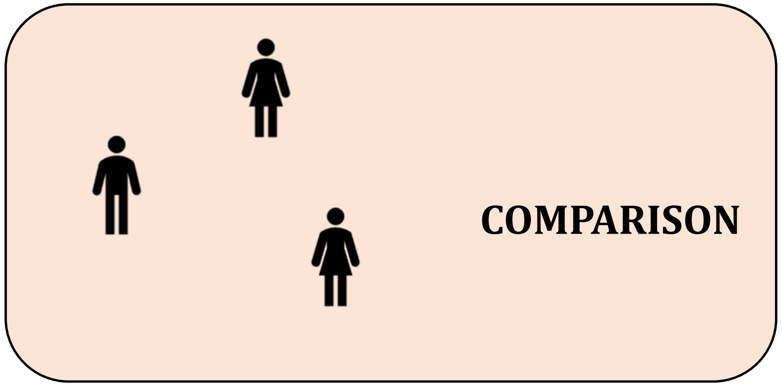
Must only consider pre-program characteristics or characteristics that do not change over time

Matching

KEY POINTS
- Matching variables must be measured *before* the program (or not change over time)
- Matching deals with **observed** characteristics only !!!



Variable	Average
Age	28
Years of schooling	8
Previous employment	54%
Parent income	4,200

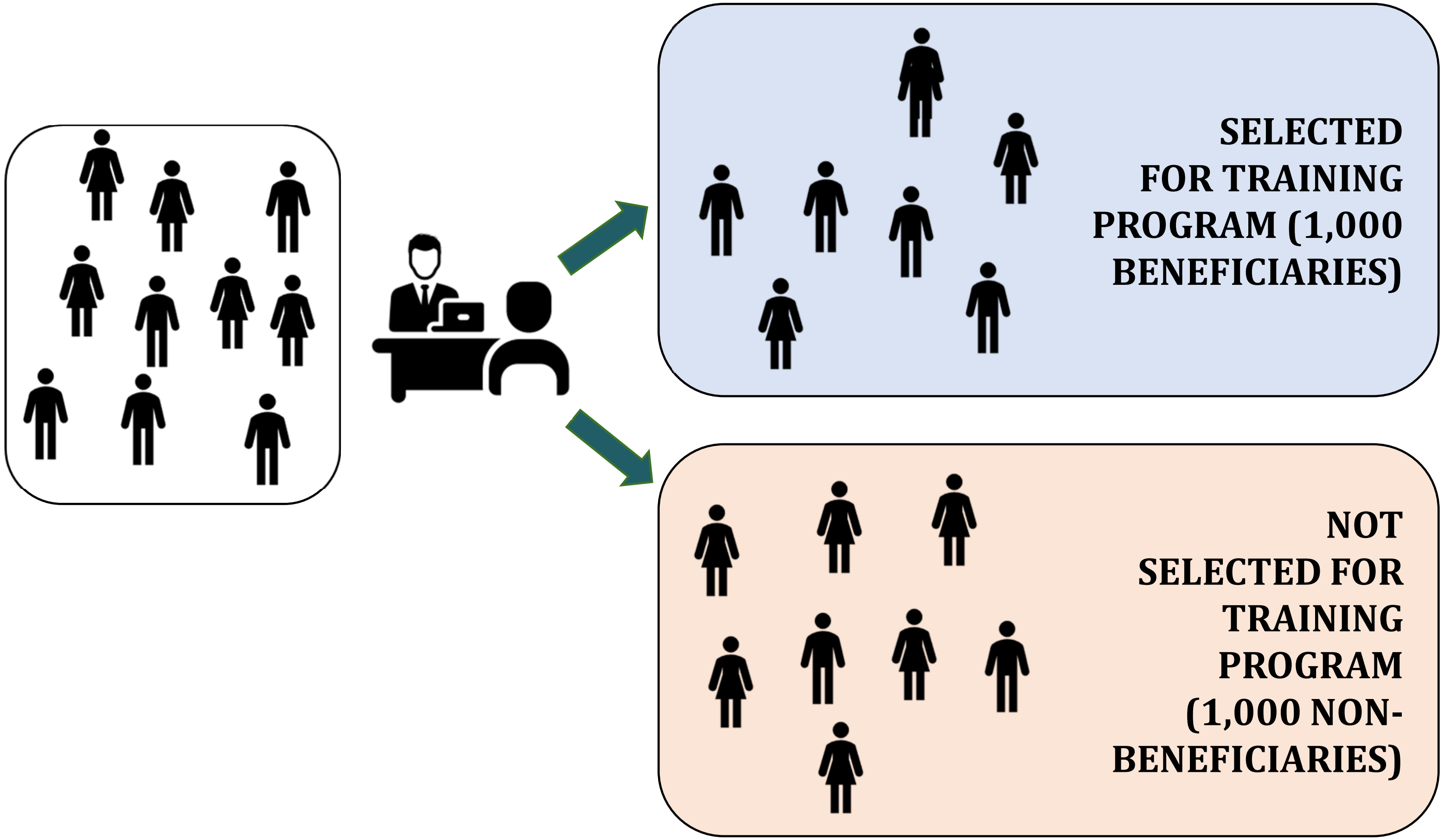


Variable	Average
Age	27
Years of schooling	7
Previous employment	53%
Parent income	4,300

Difference-in-differences

Non-random selection, the groups are different

Number of Applicants: 2,000 (1,000 Males / 1,000 Females)

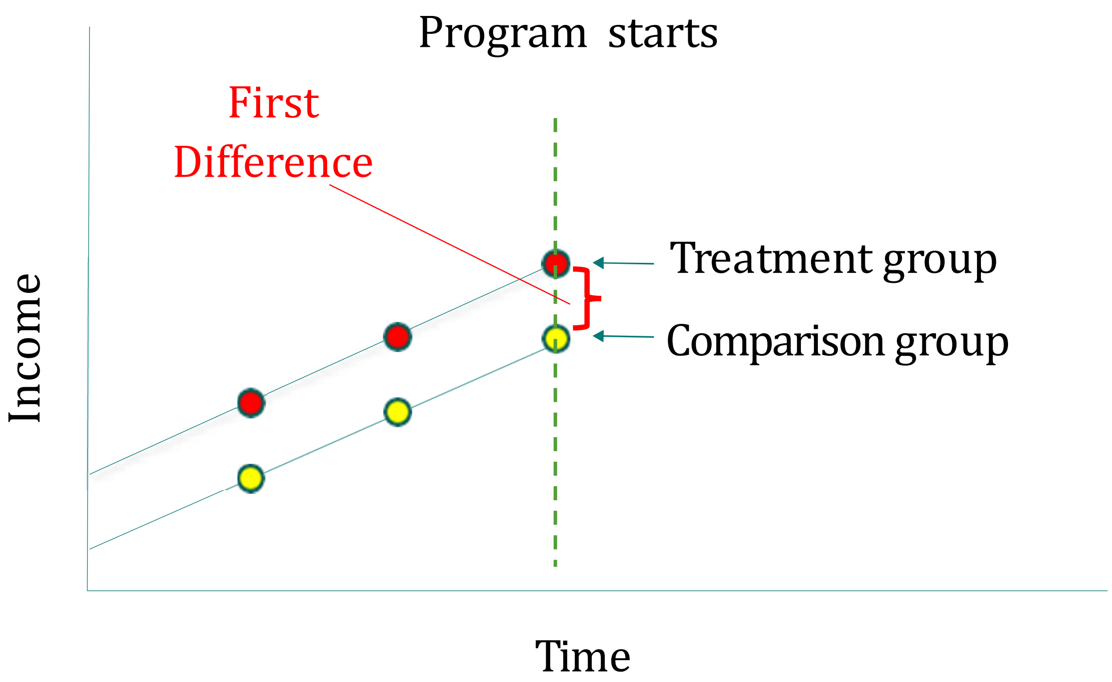


Variable	Average
Age	30
Years of schooling	10
Previous employment	60%
Parent income	5,000

Variable	Average
Age	24
Years of schooling	6
Previous employment	46%
Parent income	3,400

Difference-in-differences

- In the difference-in-differences approach, we accept that the Treatment and Comparison groups are *different*.
- IMPORTANT: this approach requires to have data on both groups *before* the program starts

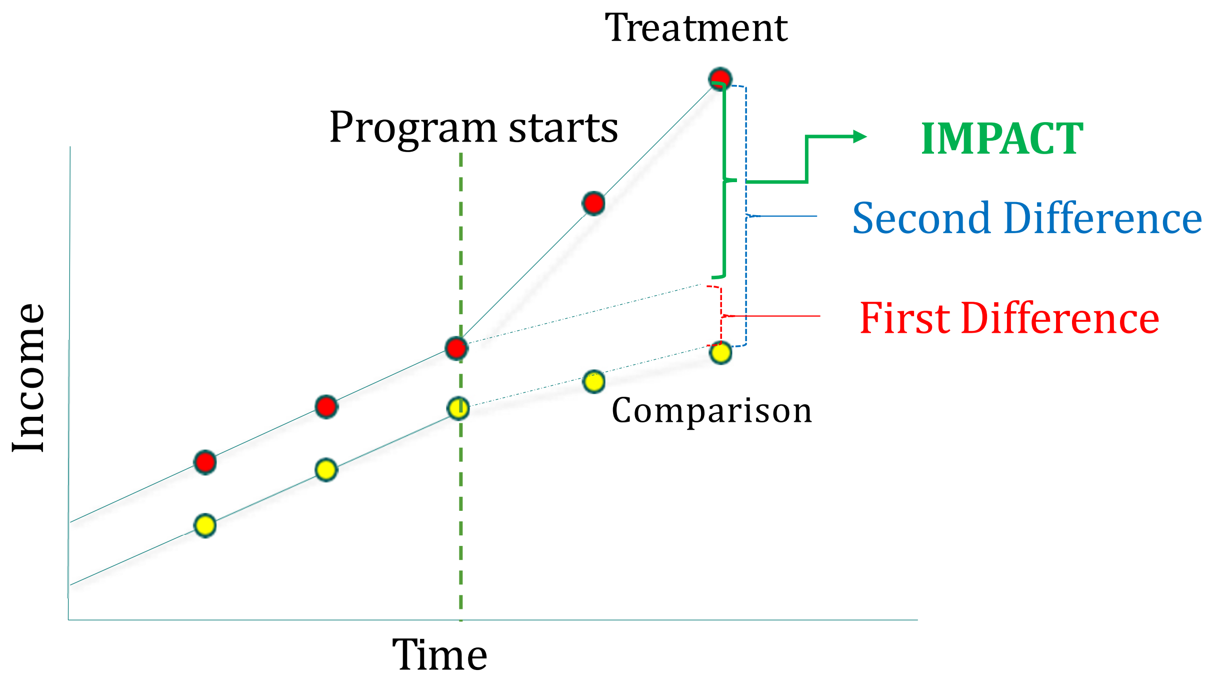


Center for Evaluation and Development

Difference-in-differences

European Commission

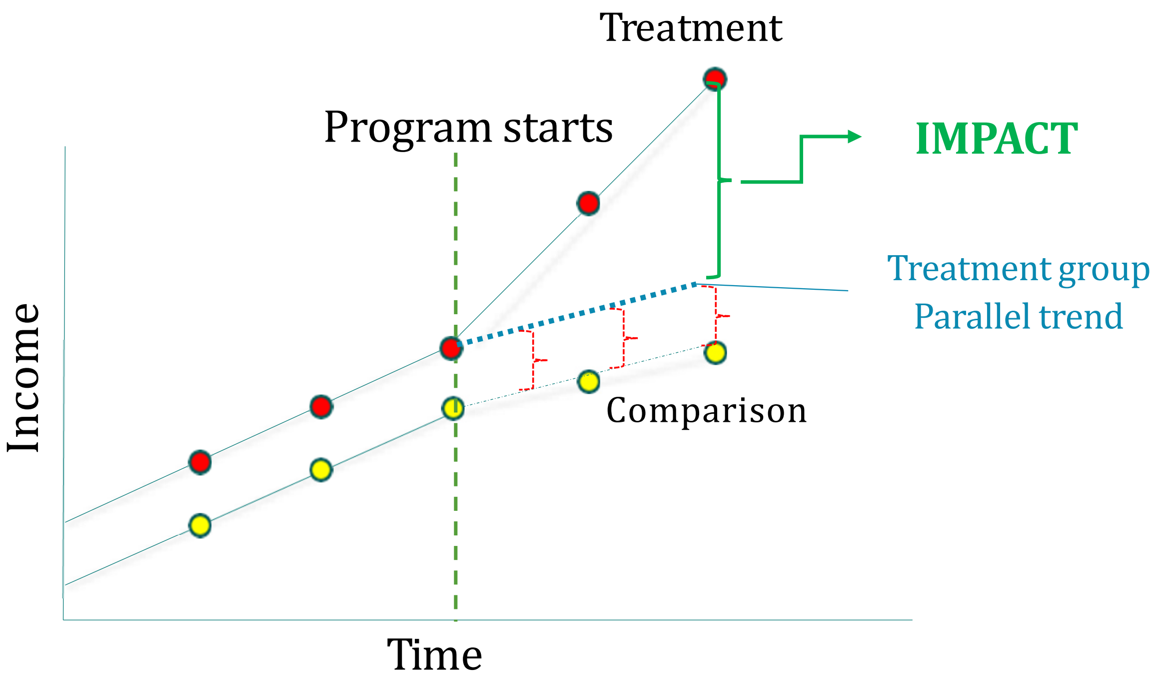
- Data on both groups must be collected at a later point in time, after the program has started.
- The difference observed after the program started is adjusted by subtracting the first difference observed *before* the program to yield the impact estimate



Difference-in-differences

- The difference-in-difference approach relies on the **parallel trends assumption**:

→ *We assume the Treatment group would have evolved similarly as the Comparison group had they not received the program (dashed blue line)*



Other CIE approaches

Other quasi-experimental approaches to conduct CIE include:

- Instrumental variable
- Regression discontinuity

CIE Resources

- Training Workshop on Counterfactual Impact Evaluation (CIE) – PowerPoint Slides

Books

- World Bank, Impact Evaluation in Practice - Second Edition (Book)

Videos

- InterAction, Introduction to Impact Evaluation
- Esther Duflo, Randomized Controlled Trials and Policy Making in Developing Countries

Podcasts

- IEU Talks Episode 2: The Power of Impact Evaluation in Development Cooperation
- Evidencing impact (parts 1+2)

RECAP YEAR 2

Data Collection (DC) for CIE

Center for Evaluation and Development

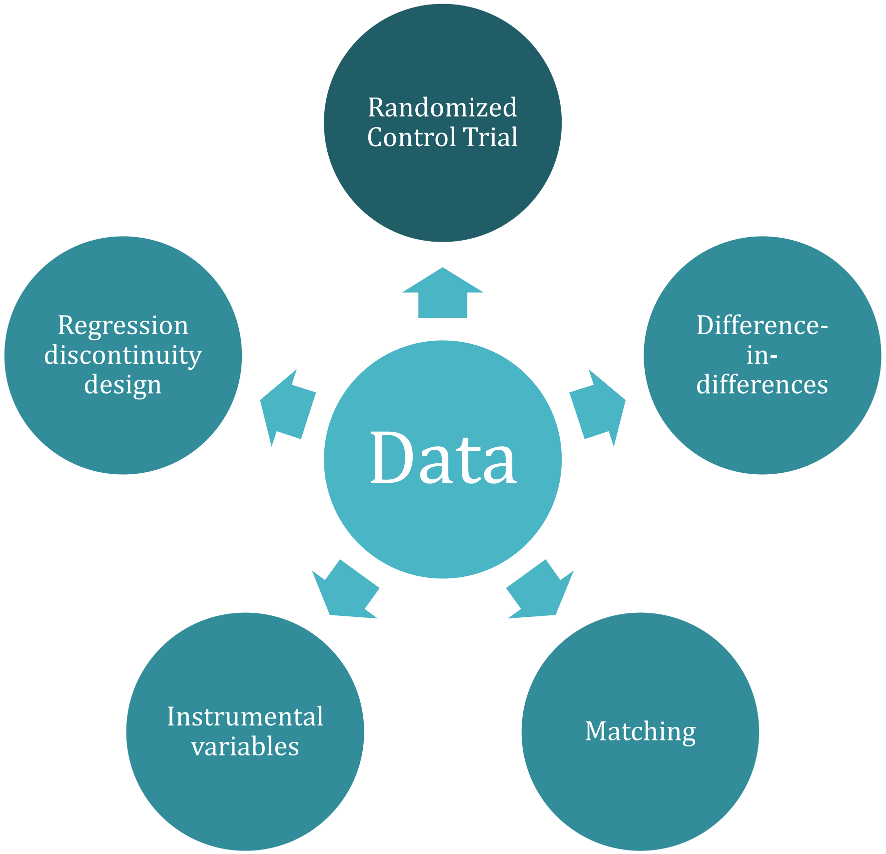
Recap Year 2 – Objectives

European Commission

- Review key aspects of conceptualizing/preparing data collection
- Review key ideas on sampling and sampling frames
- Review practical considerations such as data quality and research ethics

Center for Evaluation and Development

Relevance of data collection in CIE



Relevance of data collection in CIE

- Data collection and quality data are essential for CIE because they enable the construction of a valid and reliable counterfactual

- High quality data is essential for answering evaluation questions and measuring program impacts

“Garbage in, garbage out”

Your analysis is as good as your data.

Center for Evaluation and Development

Data for CIE

European Commission

- Data for CIE can be obtained using mixed methods:
  - Quantitative and Qualitative data
- Mixed methods allow gaining deeper insights into:
  - The experiences and motivations of beneficiaries, implementers and stakeholders
  - Program effectiveness and contextual factors that drive impacts
- Data for CIE extend beyond outcomes/impacts: monitoring data, secondary data for benchmarking and variables influencing outcomes
- Available data determines the choice of CIE approach

Center for Evaluation and Development

From Evaluation Question to Data Collection

European Commission

.....To be well prepared when designing data collection, know your research questions and objectives……

CONSTRUCT	INDICATOR	RESPONSE
Defining concepts within the evaluation question	How can we try and measure this in the real world?	Questions within a survey

Center for Evaluation and Development

Evaluation Question to Data Collection

European Commission

CONSTRUCT → INDICATOR → RESPONSE

What effects do the interventions have on livelihood in terms of **economic wellbeing** of refugee and host communities?

- Average income
- **Employment Status**
- Security in employment
- Business Ownership
- Business Performance
- Asset Ownership

1. A paid employee
2. A paid worker on household farm
3. An employer
4. Unpaid worker
5. Internship
6. None of the above

Tool development

- Tool development is critical for data collections for reliable CIE
  - Leverage existing literature and tools
  - Pretest tools (desk and field) and refine questions and responses before use in data collections
  - Important to avoid measurement error
    - Poorly designed questions and survey
    - Cognitive challenges in answering the question
    - Social desirability bias

Center for Evaluation and Development

Sampling and Sampling Frame – Definitions

European Commission

- **Target population** = the group for whom the survey data are used to derive information

- **Sampling frame** = lists or procedures used to identify all units of the target population

- **Sample** = the group of units selected from the sampling frame from which measurement will be sought

- **Respondents** = elements that are successfully measured from the sample

CIE Sample and Sampling Frame

- In most CIE, data cannot be collected from all units → focus on a *sample*
- The sampling frame ideally includes all units from the population that the evaluation is focused on → census.
- For a CIE, the sampling frame is usually a list of all the units that:
  - Received the program
  - Did not receive the program and are identified as the counterfactual group
- The sample is drawn from the sampling frame

Center for Evaluation and Development

Sampling Frame

The role of program data

- When sampling is not done properly, it leads to coverage error

→ Important to have a sampling frame that is complete, valid and reliable

- Sampling frames for program CIE often build on program data:
  - Application database and selection information
  - Information on replacements or drop-outs of the program
  - Contact information for participants

- To ensure that an evaluator can effectively use program data to build a sampling frame:
  - Integrate into a centralised monitoring system to bring all data gathered on the field into one (online) database
  - Create replicable links between various documents/data sources (i.e., Unique ID for each participant)
  - Keep as up-to-date as possible
  - Include basic quality assurance checks – no duplicates, totals match

Center for Evaluation and Development

Sampling

- What sample you draw depends on what questions you are trying to answer with your data.
- If you are trying to answer questions about the population or program quantitatively – e.g.:
  - What is the typical level of education in the participants of my program?
  - What is the rate of employment of participants after graduation
  - What is the impact of the program on income?

→ Aim to draw an unbiased sample to get the best estimates for the population

- Generally, the bigger the sample size, the better the estimate

Center for Evaluation and Development

Sampling

- Types of sampling:
  - **Probability sampling** methods reduce the possibility of bias as the possibility of someone being selected as part of the sample relies completely on chance
    - Simple random sampling; Stratified random sampling; Clustered sampling
    - Best suited approaches for quantitative part of a CIE
  - **Non-random sampling** means that the selection of the sample is not driven by chance → convenience sampling, purposive sampling (sometimes used for qualitative studies)
    - Study and sample can be guided by findings
    - Can help if the sampling frame is not clear
- Sample size is critical for increasing the chance of correctly identifying impacts

Sample Size – Key concepts

- Trade-off between sample size, the size of the effect (impact) that can be measured (MDE), and statistical precision

For a given…	*If…*	*Then…*
Sample size	MDE ↑	Statistical precision ↑
MDE	Sample size ↑	Statistical precision ↑
Statistical precision	Sample size ↑	MDE ↓

- Trade-offs are usually non-linear – e.g., if the expected impact of the program (MDE) is divided by 2, the sample size required to measure it accurately will increase by more than 2!
- Equal size of treatment and control group improves MDE and statistical precision
- Clustering matters for sample size (clustered CIE designs usually require larger sample size)

Data quality

- Non-response: Failure to obtain intended information from respondents
  - Attrition
    - E.g., Respondents that were part of the program have moved away and changed their phone number so survey teams cannot find them.
    - Teachers that have left teaching since the baseline and are therefore no longer in the population of interest for a teacher training program.
  - Refusals
  - Poor questionnaire design

- Systematic non-response of sampled respondents
  - If non-random then may lead to biased data
  - Loss of sample size and power of analysis

Center for Evaluation and Development

Data quality

European Commission

How can data quality be ensured?

- Thorough questionnaire design
- Data collection methodology
  - Fieldwork protocols
  - Training of field staff
  - Method of administering the survey
    - Pen and Paper Personal Interviews (PAPI)
    - Computer Assisted Personal Interview (CAPI)
- Data collection monitoring
  - Daily (automatized) checks to identify potentially problematic data and/or potentially poorly performing enumerators

Center for Evaluation and Development

CAPI Software

European Commission

Survey Solutions	SurveyCTO	KoboToolbox
Developed by the World Bank	Developed by Dobility using ODK Open source tool	Developed by KoBo Inc. using ODK Open source tool
Requires user to setup/have their own cloud or local server	Subscription fee	Free (with usage limits)
Extremely in-depth paradata	Range of plug-ins developed to enhance surveys	Less developed plug-ins
Simple design of complex questionnaires	Requires strong programming skills for complex questionnaires	Requires strong programming skills for complex questionnaires

Leveraging Technology for high data quality for CIE and MIS

- Leverage technology for collection of high-quality data and monitoring systems:
  - Computer Assisted Personal Interview (CAPI)
  - Computer Assisted Telephone Interview (CATI)
    - Phone surveys
  - Geographic Information Systems (GIS)
    - Sampling
    - Data quality checks

Monitoring Systems

- Monitoring evaluation is not the same as CIE

→Monitoring evaluation = Does the program/intervention work as planned?

- Similar to CIE, monitoring evaluation depends heavily on data collection
- Monitoring systems are critical for impact evaluation
- Monitoring systems provide information on available resources, outputs, and need for backstopping and correction

Center for Evaluation and Development

Data for CIE - Ethics

European Commission

- No CIE is worth risking the safety of participants in data collections
- Protection of participants must be the ultimate guiding principle in all data collections
- Obtain informed consent before collection of data
- Satisfy all ethical requirements from the relevant Ethics Board and obtain IRB approvals and permits before the start of field data collection
- Data security and protection is critical in all data collections

Center for Evaluation and Development

Resources

- Training Workshop on Counterfactual Impact Evaluation (CIE) – PowerPoint Slides
- Training Workshop on Data Collection of Micro Data in Hard-to-Reach Areas– PowerPoint Slides

Books

- World Bank, Impact Evaluation in Practice - Second Edition (Book)

Videos

- InterAction, Introduction to Impact Evaluation
- Esther Duflo, Randomized Controlled Trials and Policy Making in Developing Countries

Podcasts

- IEU Talks Episode 2: The Power of Impact Evaluation in Development Cooperation
- Evidencing impact (parts 1+2)

European Commission

Center for Evaluation and Development

European Commission

END OF SESSION 1

Center for Evaluation and Development

European Commission

Session 2: Descriptive Statistics for Monitoring and Evaluation

C4ED – EUTF

October 2023

Center for Evaluation and Development

Overview

- Year 1 focussed on intuition about CIE methods, and Year 2 on practical aspects of data collection for CIE and monitoring
- This year's workshop focusses on **what to do with the data**
- We will go over basic data analysis concepts/methods that help inform and conduct a CIE
- The aim is to get a sense of how different analyses work at an intuitive level
  - No requirement for previous knowledge of statistics


Descriptive Statistics
Statistical Testing
Regression Analysis


Descriptive Statistics
Statistical Testing
Regression Analysis

Center for Evaluation and Development

European Commission

Descriptive Statistics for Monitoring and Evaluation

Center for Evaluation and Development

Overview

- Descriptive statistics
- Categorical and continuous data
- Tabulation (univariate), cross-tabulation (bivariate)
- Measures of central tendency → mean, median, mode
- Measures of dispersion → min-max, interquartile range, variance/standard deviation
- Outliers → how to identify them and how to dealing with them (deletion, imputation, winsorization)
- Inferential statistics (intro / intuition)
- APPENDIX: Statistical distribution and Skewness

We will focus on intuition rather than technical aspects

What are descriptive statistics?

- Descriptive statistics provide an overview/summary of complex quantitative information contained in large datasets

- Even basic descriptive statistics can help:
  - Understand the key characteristics of program beneficiaries
  - Confirm whether a program goals are being met
  - Spot potentially problematic/unusual patterns in the data

Categorical data

- Information that can be recorded in terms of exclusive categories
- Gender; Employment status; Asset ownership

Continuous data

- Variables that can take any real value in the range of possible values
- In practice, continuous variables are often bounded at 0
  - Age → cannot be less than 0, cannot be infinite
  - Income; Area of land owned → cannot be less than 0

→Different types of data require different statistics

Categorical data

Univariate:

- Aggregating
- Tabulating

Bivariate:

- Cross-tabulating
- Disaggregating

Continuous data

- Measures of central tendency (mean, median, mode)
- Measures of dispersion (min-max, interquartile range, standard deviation)

Descriptive Statistics in Monitoring and Evaluation

Example – Setup

- Organization (your client) designs and implements a vocational training programme in TVET centres
- **Overarching goal**: Economically empower disadvantaged youth to engage in employment and livelihood strategies
- **Specific goals**:
    - Enrol 1,000 young people to participate in the vocational training
    - Ensure a 50:50 gender split across all participants
    - Achieve a 90% graduation rate
    - Increase monthly income of graduates by 800 units six months after completing training

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation
Example – Objective

European Commission

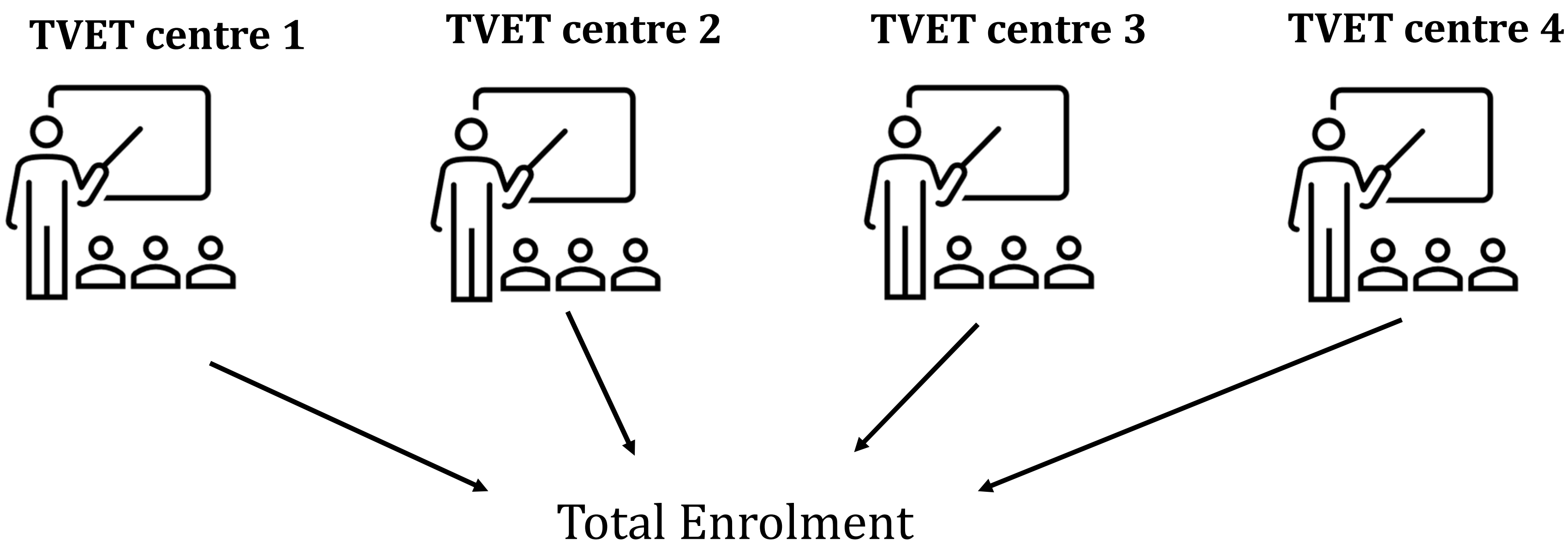
- You are provided with a monitoring data set with the following information on all participants:
  - Gender
  - TVET centre enrolled at
  - Graduation status
  - Average income of graduates six months after completing training

- In the next steps we'll go through the toolbox of descriptive statistics to see if the programme met its targets

→ In terms of OECD evaluation criteria, was it *effective*?

Descriptive Statistics in Monitoring and Evaluation
Example – Enrolment

- Aim: Enrol 1,000 young people to participate in the vocational training



Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation
Example – Aggregating / Tabulating

- Aim: Enrol 1,000 young people to participate in the vocational training

TVET Centre	Number of participants
TVET Centre 1	115
TVET Centre 2	122
TVET Centre 3	109
TVET Centre 4	116
Total	**462**

- The programme did not reach its target

  →Further investigation (possibly qualitative) warranted to understand why (Problems in program design? Or bottlenecks in implementation?)

Descriptive Statistics in Monitoring and Evaluation
Example – Aggregating / Tabulating

- Aim: Ensure a 50:50 gender split across all participants

Gender	TVET Centre 1	TVET Centre 2	TVET Centre 3	TVET Centre 4	Total	%
Male	60	79	42	51	232	50.2%
Female	55	43	67	65	230	49.8%
Total	**115**	**122**	**109**	**116**	**462**	**100%**

- Programme almost reached a perfect 50:50 gender split
  →Same cannot be said for each TVET centre though

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation
Example – Aggregating / Tabulating

European Commission

- Aim: 90% graduation rate

	Graduated		Dropped Out		Total
TVET Centre	Frequency	Percentage	Frequency	Percentage	
TVET Centre 1	108	94%	7	6%	115
TVET Centre 2	114	93%	8	7%	122
TVET Centre 3	92	84%	17	16%	109
TVET Centre 4	108	93%	8	7%	116
Total	**422**	**91%**	**40**	**9%**	**462**

In your view, is there any data point that sticks out in this table?

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation

Example – Aggregating / Tabulating

	Graduated		Dropped Out		Total
TVET Centre	Frequency	Percentage	Frequency	Percentage	
TVET Centre 1	108	94%	7	6%	115
TVET Centre 2	114	93%	8	7%	122
TVET Centre 3	92	84%	17	16%	109
TVET Centre 4	108	93%	8	7%	116
Total	**422**	**91%**	**40**	**9%**	**462**

- The overall graduation rate is over 90% → the program is reaching its target!
- **However**, the graduation rate in TVET Centre 3 seems substantially lower than in other centres

→ Let's dig a little deeper by using **cross-tabulations**

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation

Disaggregation / Cross-Tabulation

European Commission

- Cross-tabulation and disaggregation can be extremely useful when exploring data

- Can help to check some of the underlying stories behind aggregated data/aggregated statistics

→Starting point for deeper analysis

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation

Example – Disaggregating / Cross-Tabulating

European Commission

- Let's look at the average graduation rate by gender and TVET centre → cross-tabulation

	Graduation Rate	
Gender	**Male**	**Female**
TVET Centre 1	95%	93%
TVET Centre 2	94%	93%
TVET Centre 3	100%	75%
TVET Centre 4	96%	91%
Average	**96%**	**87%**

→Men have a higher graduation rate than women overall

→The difference is particularly stark in TVET Centre 3

- Aim: Average monthly income of 800 for graduates six months after completing training

→Up to now, we looked at data that could be easily summarized using frequencies and proportions

→Data such as income cannot be easily summarized that way

→We require other tools, namely **measures of central tendency**. Let’s recap what they are!

Center for Evaluation and Development

European Commission

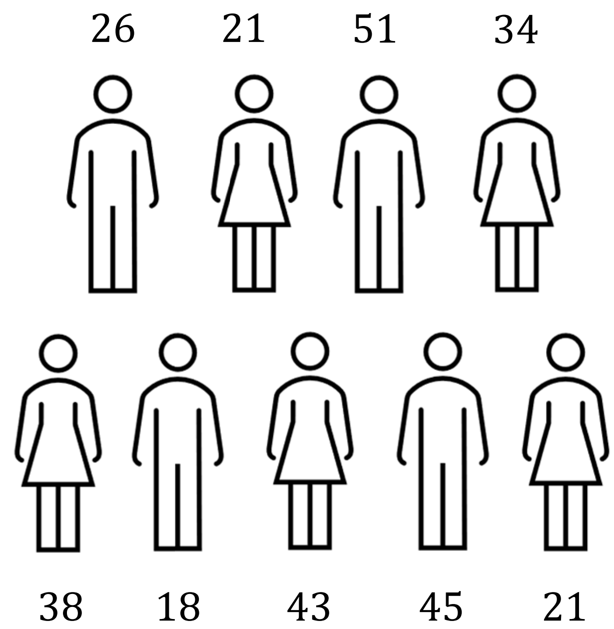
Measures of Central Tendency

Measures of central tendency
Mean

- **Mean**: The sum of all the values divided by the total number of values in the data set

- Example:
  - We have measures of age for 9 persons
  - Mean value of age:

(26+21+51+34+38+18+43+45+21)/9 = 33

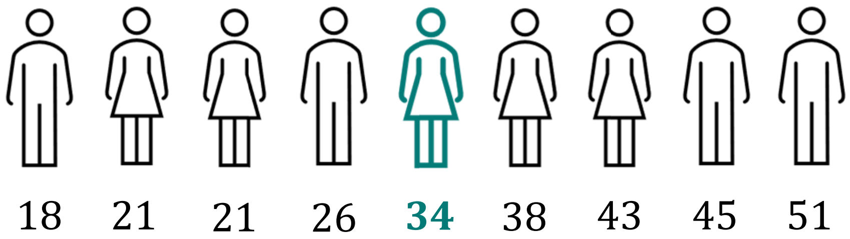


Center for Evaluation and Development

Measures of central tendency
Median

European Commission

- **Median** = For a given set of values, the median is the value that splits the set exactly in the middle, i.e., exactly half the values are below/above the median



- It is calculated by arranging the values in ascending order (from lowest to highest) and finding the value in the exact middle

- In our example → median = 34

Center for Evaluation and Development

Remark
Median and percentiles

European Commission

- The median belongs to the broader family of ***percentiles***
- Percentiles = values that split the data into given proportions
  - → E.g., the median splits the data in half, i.e., 50% of values are above the median and 50% are below
  - → The median is the 50th percentile and is sometimes denoted P50
- Similarly, we can define e.g., the 10th percentile (P10) = value such that 10% of values are below and 90% are above
- Percentiles that split the data into 4 equal-sized sets of values are called *quartiles*
  - There are 3 quartiles (P25, P50 and P75) that can be referred to as the 1st, 2nd and 3rd quartiles; the median is the 2nd quartile
- Percentiles that split the data into 10 equal-sized sets of values are called *deciles*
  - There are 9 deciles (P10, P20, P30, P40, …, P80, P90); the median is the 5th decile.

Descriptive Statistics in Monitoring and Evaluation

Example – Mean

- Aim: Average monthly income of 800 for graduates six months after completing training

[NB: impractical to write the measured income for all 462 participants, so we focus on a sample of 10 participants]

Participant	Monthly Income
Graduate 1	200
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 5	100
Graduate 6	700
Graduate 7	300
Graduate 8	0
Graduate 9	6,000
Graduate 10	200
Total	**8,580**

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation

Example – Mean

European Commission

- Aim: Average monthly income of 800 for graduates six months after completing training

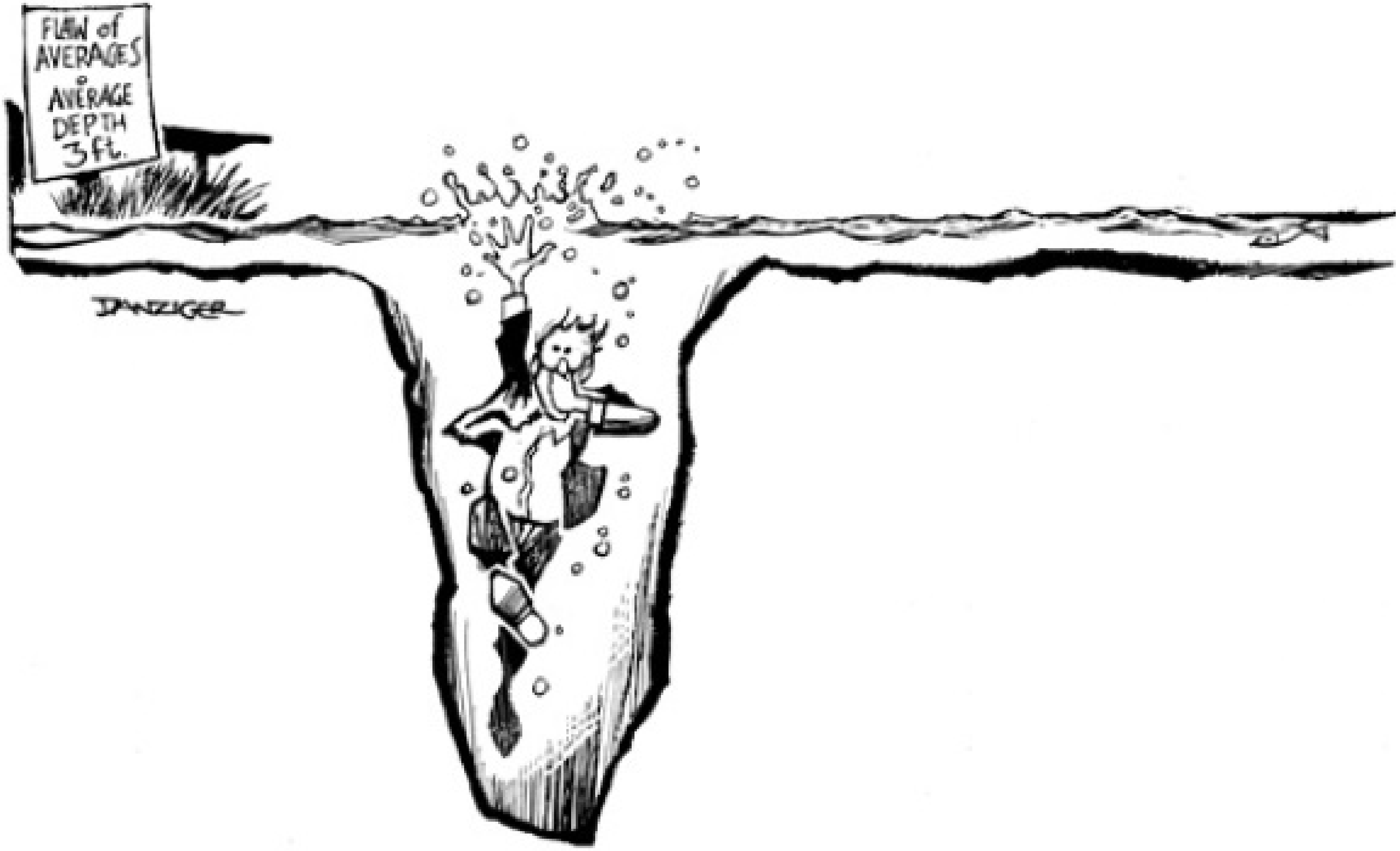
- Mean = 8,850 / 10 = 885
- Program has met its goal

→ Would you feel confident reporting that the average income of graduates is 885 per month?

Participant	Monthly Income
Graduate 1	200
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 5	100
Graduate 6	700
Graduate 7	300
Graduate 8	0
Graduate 9	6,000
Graduate 10	200
Total	**8,850**

Descriptive Statistics in Monitoring and Evaluation

The mean and outliers



Center for Evaluation and Development

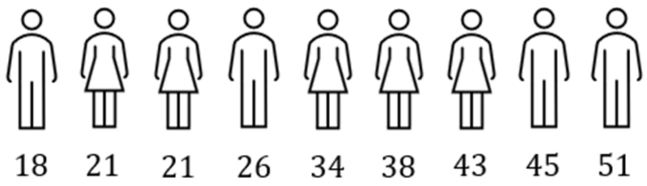
Descriptive Statistics in Monitoring and Evaluation
The mean and outliers

European Commission

- The challenge with the mean is that it can be affected by values that are very large or very small compared to the others

- In previous example, mean age was 33

→ 4 persons younger, 5 older, so the mean seems to give a good measure of central tendency



- Imagine one of the persons is very old

→ The mean is now 40

→ 6 persons younger, 3 older, so the mean may not be the best measure of central tendency in that case (we say it is "biased upwards")

- Such extreme values are called **outliers** (more on this later)

Descriptive Statistics in Monitoring and Evaluation

Skewness

- The challenge with the mean is that it can be affected by values that are very large or very small compared to the others

→Such extreme values are called **outliers** (more on this later)

- In statistics, we would say the data is **skewed**
  - **Skewness** is a measure of how symmetric a statistical distribution is – i.e., it captures whether observations are evenly spread out around the mean
  - Symmetric distributions have a skewness of 0; Asymmetric distributions can have either negative or positive skewness
  - The more asymmetric, the worse the mean is as a measure of central tendency!

[For details on skewness, please refer to the appendix.]

Skewed data

What variables can you think of that may be skewed in their distribution?

Descriptive Statistics in Monitoring and Evaluation

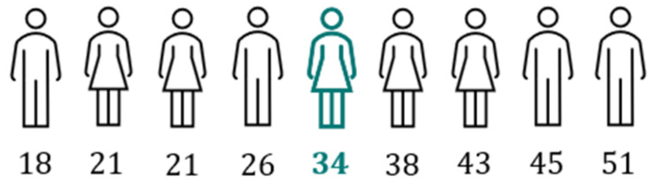
Median – Robust to outliers

- Unlike the mean, the median is less sensitive to extreme values

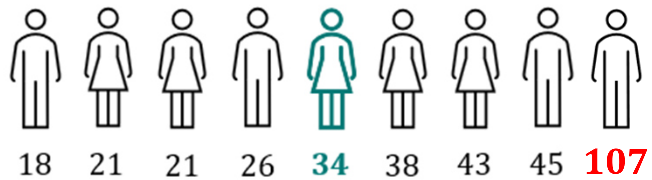
→We say the median is ***robust*** to outliers

[RECALL: The median is calculated by ranking values in ascending order (from lowest to highest) and finding the value in the exact middle]

- In previous example, median age was 34



- If the oldest person is now 107 instead of 51, the median is still 34!



Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation

Example – Median, robust to outliers

- Aim: Average monthly income of 800 for graduates six months after completing training
- Mean = 885 → "biased" due to outliers
- As an alternative you could report the median income
- Rank the values in ascending order and take the middle rank's value
  - Note: with an *even* number of observations, the median is the midpoint of the 2 middle values
- Median = (300+350)/2 = 325

→We know the median is an unbiased measure of central tendency in the presence of outliers, but is it enough to describe the data?

Participant	Monthly Income
Graduate 8	0
Graduate 5	100
Graduate 1	200
Graduate 10	200
Graduate 7	300
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 6	700
Graduate 9	6,000

Descriptive Statistics in Monitoring and Evaluation

Beyond Central Tendency

- In this example, we see the limits of looking only at measures that describe the average/central tendency
- In addition to measures of central tendency, you can look at measures of **dispersion**
  - In other words, not just what the average value/central tendency is, but how the data are dispersed/spread around it

Center for Evaluation and Development

European Commission

Measures of Dispersion

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation

Measures of dispersion – Min-max

- A simple measure of dispersion is to look at the range of values in the data, i.e., the minimum and maximum values

→We see that our measurements of monthly income range from 0 to 6,000

Participant	Monthly Income
Graduate 8	0
Graduate 5	100
Graduate 1	200
Graduate 10	200
Graduate 7	300
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 6	700
Graduate 9	6,000

Descriptive Statistics in Monitoring and Evaluation

Measures of dispersion – Variance and SD

- You can use more sophisticated measures of central dispersion such as the variance and standard deviation.
- **Variance** = measures the overall variability in the data
- **Standard deviation (SD)** = measures how far each value is from the mean on average
    - Standard Deviation = Square Root of Variance
- The higher the variance/SD, the more dispersed the data around the mean

→Variance and SD are particularly important for statistical testing (next session)

Descriptive Statistics in Monitoring and Evaluation

Example – Measures of dispersion

- Aim: Average monthly income of 800 for graduates six months after completing training

→Mean = 885

→Standard deviation = 1,810

Participant	Monthly Income
Graduate 1	200
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 5	100
Graduate 6	700
Graduate 7	300
Graduate 8	0
Graduate 9	6,000
Graduate 10	200

Descriptive Statistics in Monitoring and Evaluation

Example – Measures of dispersion

Mean = 885; Standard deviation = 1,810

- In other words, a graduate's income is on average 1,810 below or above the mean income (885)

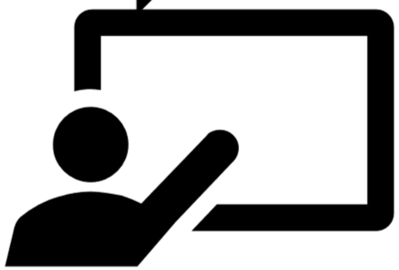
- The mean shows the programme reached its objective
- The standard deviation nuances the picture and indicates that incomes tend to be quite spread out – some far below and some far above the mean

Participant	Monthly Income
Graduate 1	200
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 5	100
Graduate 6	700
Graduate 7	300
Graduate 8	0
Graduate 9	6,000
Graduate 10	200

Descriptive Statistics in Monitoring and Evaluation

Example – SD as complement to the mean

As you can see, our program graduates earn on average 885 per month, easily beating our target of 800 per month!



Show us your standard deviations!

Participant	Monthly Income
Graduate 1	200
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 5	100
Graduate 6	700
Graduate 7	300
Graduate 8	0
Graduate 9	6,000
Graduate 10	200

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation

Example – SD as complement to the mean

- Ideally, you would have the opportunity when presenting the data (seminar, report) to contextualize the average with the measure of dispersion
  - Include standard deviation in your results
  - Discuss the results to provide context and explain

Income after 6 months		
Number of graduates	Mean	Standard Deviation
10	885	1,810

- However, people often expect a Yes/No answer to whether a target was met…

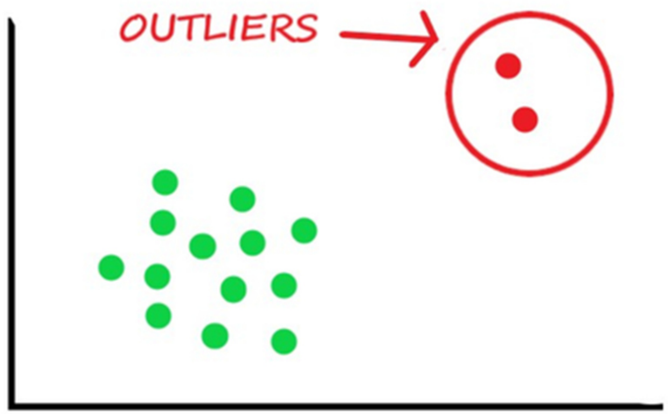
What can you do to deal with skewed data (i.e., data with outliers), so that descriptive statistics are more robust?

What can you do to deal with skewed data (i.e., data with outliers)?

- You usually deal with **outliers** when preparing the data before analysis
  - Step 1: identify outliers
  - Step 2: “process” outliers

- This can provide you with more informative/more robust descriptive statistics

- Also important for more complex analysis further down the line!



Outliers

Descriptive Statistics in Monitoring and Evaluation

Example – Outliers

- An outlier differs significantly from other observations
- In our example data, Graduate 9 could be described as an **outlier**
- No set definition to identify outliers
- Conventional rule of thumb in social sciences → outlier = value at least 2.5 to 3 standard deviations away from mean (above or below)

Participant	Monthly Income
Graduate 1	200
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 5	100
Graduate 6	700
Graduate 7	300
Graduate 8	0
Graduate 9	6,000
Graduate 10	200

Descriptive Statistics in Monitoring and Evaluation

Example – Outliers

Mean = 885; Standard deviation (SD) = 1,810

- Rule: outlier is value at least 2.5 SD (4524) below/above mean
- Consider incomes below 2.5 SD (-3,640) as outliers
  - If we assume incomes cannot be below 0 → no outliers at the lower end
- Consider incomes above 2.5 SD (5,410) as an outlier
  - Graduate 9 is an outlier

Participant	Monthly Income
Graduate 1	200
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 5	100
Graduate 6	700
Graduate 7	300
Graduate 8	0
Graduate 9	6,000
Graduate 10	200

- What can we do with outliers?

→Depends on their origin – i.e., whether they come from ***measurement errors*** or not

- What can we do with outliers?

- Outliers as measurement errors (Recall Year 2 seminar on data collection)
  - →Error in the response (intentional or not)
  - →Error in recording – e.g., accidentally entered 6,000 instead of 600
- If you are confident it is a measurement error – e.g., someone is reported as being 200 years old – you can:
  - Correct the data if possible
  - Remove this observation

- What can we do with outliers?
- If there is no evidence that outliers are due to measurement errors, 3 options:
  - Option 1: Remove
  - Option 2: Impute
  - Option 3: Winsorize

- What can we do with outliers?

➢ **Option 1: Remove**

  - Usually *case deletion* → i.e., the observation is fully removed from the analysis
  - In our example, Graduate 9 was identified as an outlier in terms of income.
  - *Case deletion* means we would completely remove Graduate 9 from the analysis, even if they are not an outlier with respect to other variables

- What can we do with outliers?

➢ **Option 2: Impute**

  - *Imputation* consists of replacing the outlier by a “representative” value
  - Could use e.g., the mean or the median, but in practice imputation methods are often more sophisticated

- What can we do with outliers?

➢ **Option 3: Winsorize**

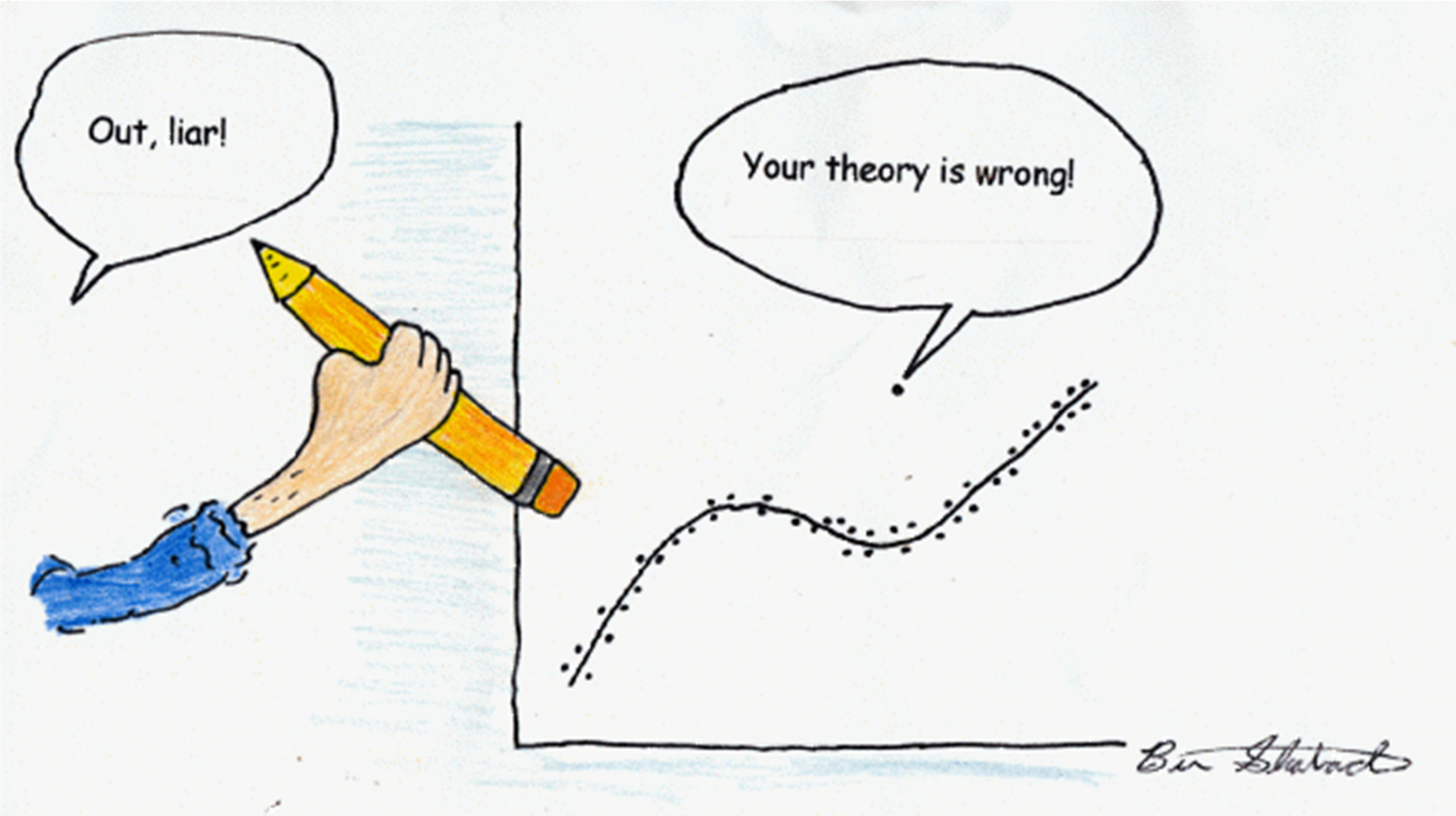
  - Set a minimum/maximum acceptable value – e.g., the 1% or 5% percentile (P1 or P5) as minimum, the 99% or 95% percentile(P99 or P95) as maximum
  - Replace values below (above) the chosen minimum (maximum) value by the minimum (maximum) value
  - This approach seeks a balance between keeping outliers and removing them.
  - Protects your data from the most extreme outliers causing issues.

- What can we do with outliers?
- If there is no evidence that outliers are due to measurement errors, 3 options:
  - ➢ Option 1: Delete
  - ➢ Option 2: Impute
  - ➢ Option 3: Winsorize

- These options should make your data easier to use and interpret
- **But** some outliers are a natural part of data and provide important information on the variance/unpredictability of the data

Center for Evaluation and Development

European Commission



Remark 1

- Some outliers bring actual information and should be kept in the data
  - ➢ Can be hard to distinguish between a "true" extreme value and an outlier due to measurement error → always a trade-off
  - ➢ E.g., with winsorization you can set a higher cut-off point (99% or 99.5% percentile) to retain some outliers
  - ➢ Other approaches exist that aim to strike a balance between stabilizing the data and retaining informative extreme values

<u>Remark 2</u>

- We said the median is robust to outliers, then why not always report the median rather than the mean?
  - Measure of dispersion when reporting the median → interquartile range
  - Statistics have been largely developed around means/averages (statistical testing, regression analysis) → descriptive statistics tend to focus on means because subsequent analysis usually focusses on means/averages

Center for Evaluation and Development

European Commission

Inferential Statistics – Intro

- If you have a strong monitoring system, you may be able to have data on **all** participants of an intervention

- However, in practice it is often the case that we don't have all information on all participants
- Then, how can we learn about the characteristics of *all* program participants?

→We can take a *sample* that is *representative* of the population
  - Please refer to Year 2 slides for information on sampling

Descriptive Statistics in Monitoring and Evaluation

Inferential Statistics – Intro

- If we use a sample, we need to estimate how ***confident*** we are that we can apply our conclusions to the whole population
  - This is referred to as our ability to make inferential statements based on our sample

- We may also use inferential statistics when estimating the level of confidence in treatment effects (i.e., measures of causal impact)

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation

European Commission

Inferential Statistics – Intro

- Imagine 2,000 people applied for our programme and 1,000 were randomly selected
  - i.e. we conducted a rigorous RCT → Any difference between the two groups should be attributable to the programme!
- The data show that those that received the vocational training earn 150 more per month than those that did not

	Number of individuals	Mean income
Received vocational training	1,000	1,200
Did not receive vocational training	1,000	1,050

- Did the programme have an *impact* on income? How confident are we that this difference reflects the *true* impact of the programme?

Center for Evaluation and Development

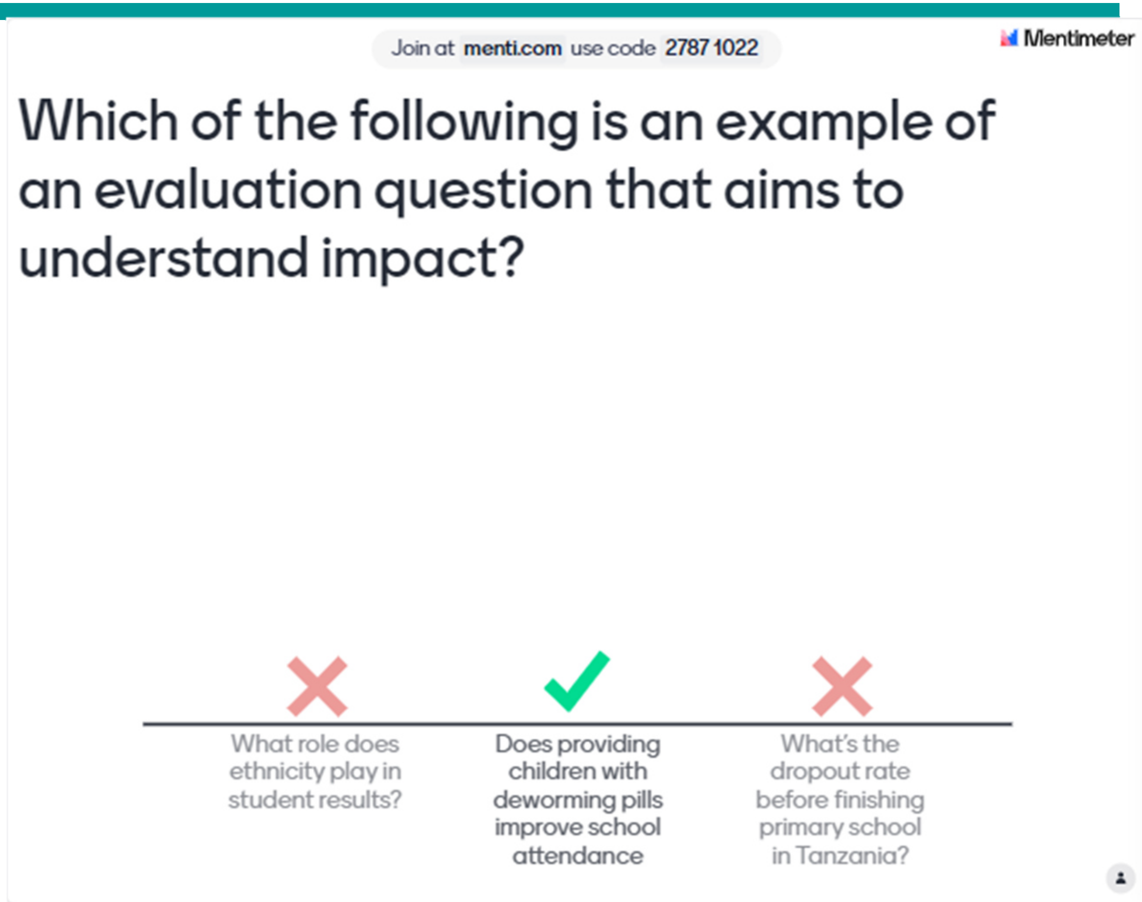
European Commission

END OF SESSION 2

Interactive Quiz

Center for Evaluation and Development

- If the interactive quiz does not show on your screen you can join in using your phone by logging in to "***menti.com***"
- Enter your name and use the code below if required to participate in the interactive quiz
    - Code: **27871022**
- Join the quiz by entering the code provided above.



Center for Evaluation and Development

European Commission

Session 2: Descriptive Statistics for Monitoring and Evaluation

Appendix

Center for Evaluation and Development

European Commission

APPENDIX – Statistical Distribution

Center for Evaluation and Development

Statistical distribution
Definition

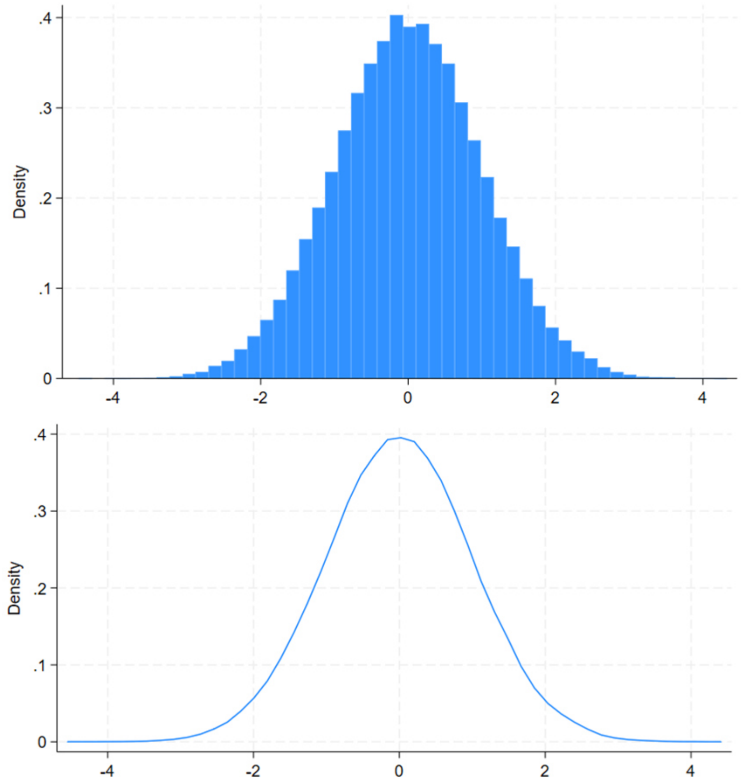
European Commission

- A continuous variable can take many possible values
  - E.g., age measured in years can take any value between 0 and 123 (the oldest person ever died at about 122.5 years)

- The **statistical distribution** (or probability distribution) of a variable basically shows what values are common and uncommon
  - E.g., if we measure age for 100 (randomly selected) people, we expect that about ¼ is < 14 y.o., about 50% is 25-65 y.o., and about 10% > 65 y.o.

Statistical distribution
Visualization

- The **statistical distribution** (or probability distribution) of a variable shows what values are common and uncommon
- Often represented by the **probability density function**:
    - Histogram (vertical bar chart) or a curve
    - X-axis = the range of possible values for the variable
    - Y-axis = the probability *density*
        - Caution: the density is NOT a probability
        - Intuition: for a given value *x*, the higher the density, the higher the likelihood/probability that the variable – when measured – takes on a value close to *x*

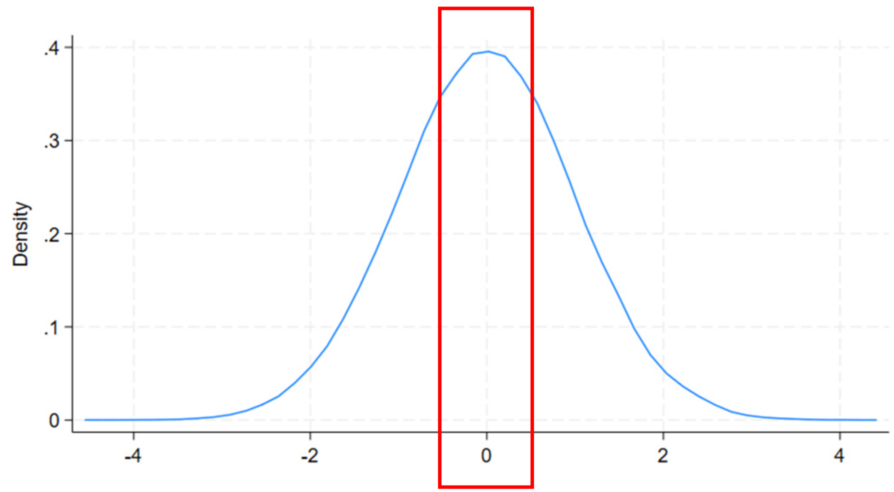


Example: Density function of a Normal distribution

Statistical distribution and central tendency

- A continuous variable can take many possible values
- Measures of ***central tendency*** help us "get a sense" of the distribution without having to browse all the different values measured for the variable

  - The higher the density, the more likely the variable to take that value
  - Focus on **central tendency** because, in most cases, the most common values tend to be around the "centre" of the distribution



APPENDIX – Skewness

Center for Evaluation and Development

Descriptive Statistics in Monitoring and Evaluation
Skewness

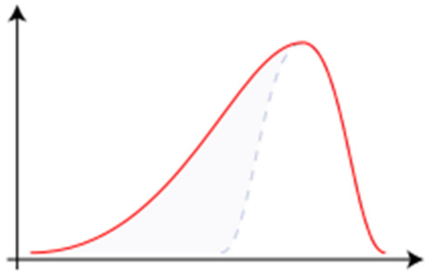
European Commission

- The challenge with the mean is that it can be affected by values that are very large or very small compared to the others

→Such extreme values are called **outliers** (more on this later)

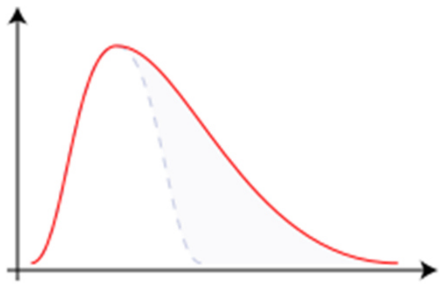
- In statistics, we would say the data is **skewed**
  - **Skewness** is a measure of how symmetric a distribution is
  - The more asymmetric, the worse the mean is as a measure of central tendency

Normal distribution (theoretical ideal), perfectly symmetric → skewness = 0



Negative Skew

Example of negative skew → low values occur more often than expected (a.k.a. left skew)



Positive Skew

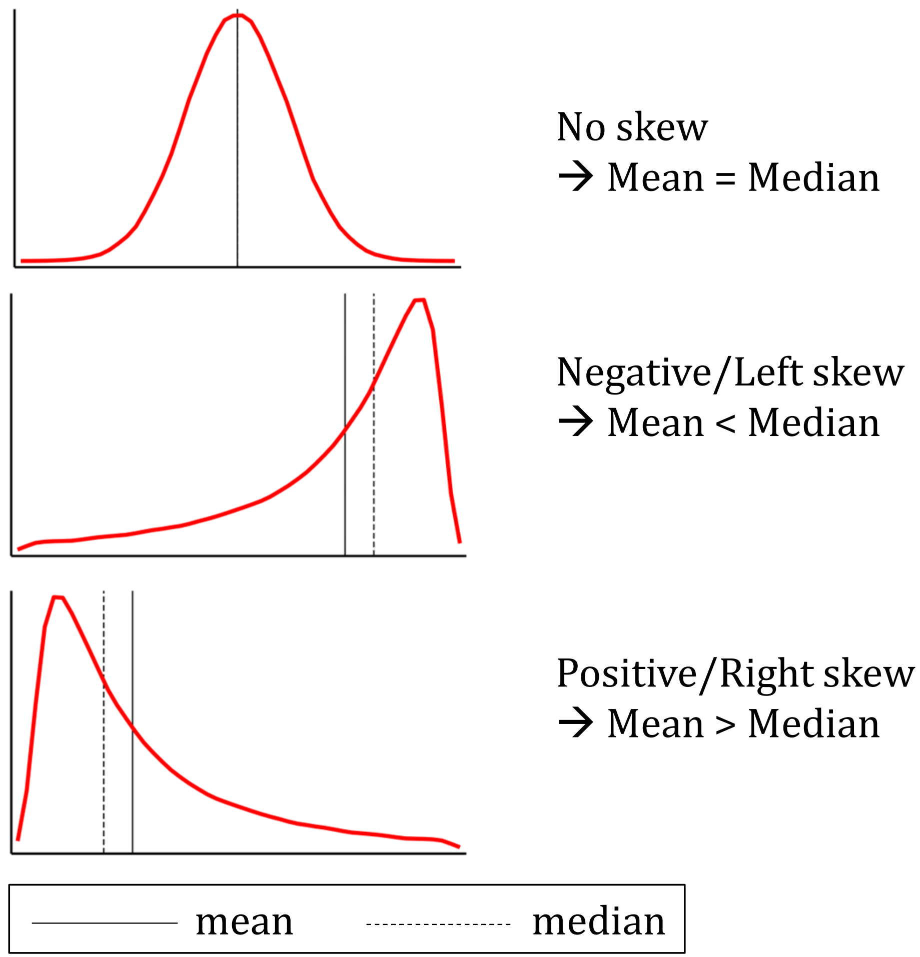
Example of positive skew → high values occur more often than expected (a.k.a. right skew)

Descriptive Statistics in Monitoring and Evaluation
Skewness, mean and median

- **Skewness** is a measure of how symmetric a distribution is

- The more asymmetric, the worse the mean is as a measure of central tendency

→The more skewed/asymmetric the data, the further apart the mean and median



Center for Evaluation and Development

European Commission

APPENDIX – Median and percentiles

Center for Evaluation and Development

Remark
Median and percentiles

European Commission

- The median belongs to the broader family of ***percentiles***
- Percentiles = values that split the data into given proportions
  - →E.g., the median splits the data in half, i.e., 50% of values are above the median and 50% are below
  - →The median is the 50^{th} percentile and is sometimes denoted P50
- Similarly, we can define e.g., the 10^{th} percentile (P10) = value such that 10% of values are below and 90% are above
- Percentiles that split the data into 4 equal-sized sets of values are called *quartiles*
  - There are 3 quartiles (P25, P50 and P75) that can be referred to as the 1^{st}, 2^{nd} and 3^{rd} quartiles; the median is the 2^{nd} quartile
- Percentiles that split the data into 10 equal-sized sets of values are called *deciles*
  - There are 9 deciles (P10, P20, P30, P40, …, P80, P90); the median is the 5^{th} decile.

Center for Evaluation and Development

Remark
Median and percentiles

European Commission

Percentile	Name	Equivalent
P1	1st percentile	Bottom percentile
P2	2nd percentile	–
…	…	…
P10	10th percentile	1st decile / Bottom decile
…	…	…
P25	25th percentile	1st quartile / Bottom quartile
…	…	…
P50	50th percentile	2nd quartile / 5th decile / Median
…	…	…
P75	75th percentile	3rd quartile / Top quartile
…	…	…
P90	90th percentile	9th decile / Top decile
…	…	…
P98	98th percentile	–
P99	99th percentile	Top percentile

Center for Evaluation and Development

European Commission

APPENDIX – Mode and interquartile range

Mode and interquartile range

The following slides present:

- Another measure of dispersion → the ***interquartile range***
  - Useful to report as a measure of dispersion with the median, because standard deviation is relevant for the mean only
- Another measure of central tendency → the ***mode***
  - Most useful for categorical variables

Descriptive Statistics in Monitoring and Evaluation

Measures of dispersion – Interquartile range

- **Interquartile range** = the range of values between the 1st and 3rd quartile, i.e., between P25 and P75

→In other words, the difference in values after removing the bottom 25% of value from the top 25% of values

- P25 = 25th percentile = 200
- P75 = 75th percentile = 550

→Interquartile range = 550 – 200 = 350

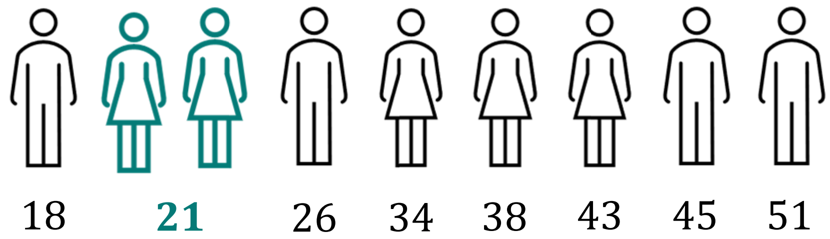
- The values in the middle (the 50% of values between P25 and P75) fall within a range of just 350 per month

Participant	Monthly Income
Graduate 8	0
Graduate 5	100
Graduate 1	200
Graduate 10	200
Graduate 7	300
Graduate 2	350
Graduate 3	400
Graduate 4	600
Graduate 6	700
Graduate 9	6,000

Percentile calculator with formulas: https://www.calculatorsoup.com/calculators/statistics/percentile-calculator.php

Measures of central tendency
Mode

- **Mode**: The value that appears most frequently in a data set
  - →Most useful for categorical variables



- In our example → mode = 21

Center for Evaluation and Development

European Commission

Session 3a: Statistical Testing in CIE

C4ED – EUTF

October 2023

Descriptive Statistics – Recap

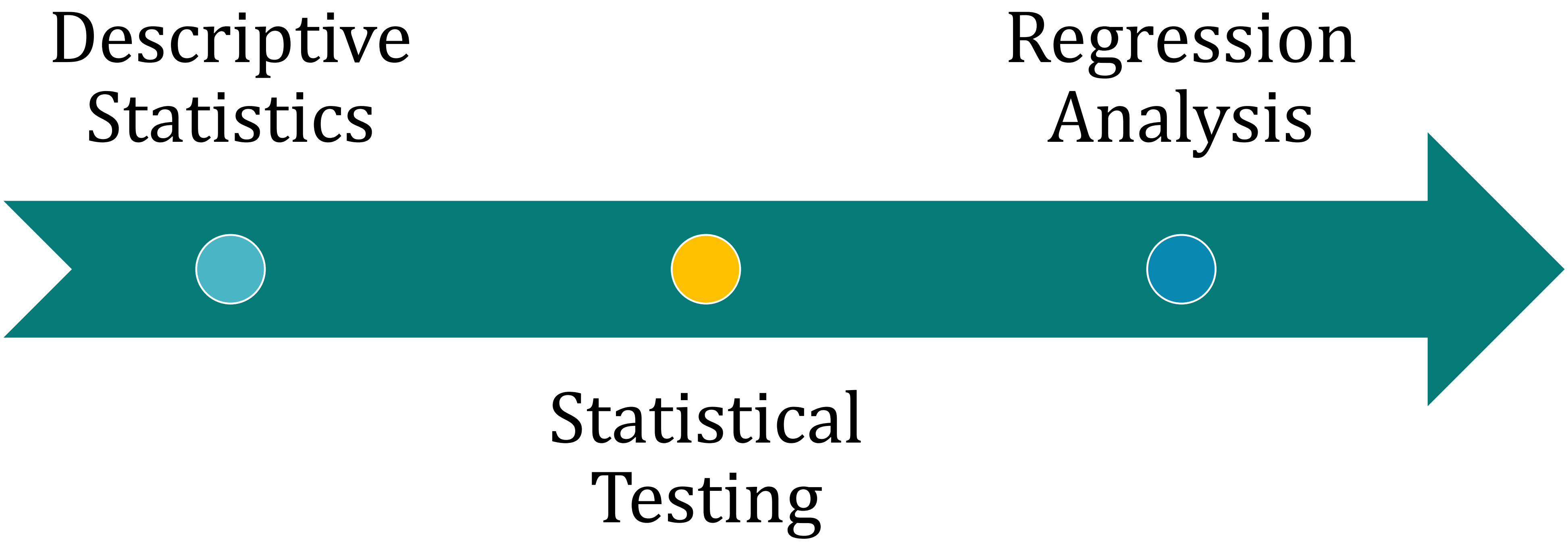
- Descriptive statistics provide a compact, high-level summary of complex data

- Categorical data → Frequencies, Proportions, Tabulation (univariate) or Cross-tabulation Continuous data → Measures of central tendency:
  - Mean → intuitive (average), but problematic with skewed data (i.e. in presence of outliers)
  - Median → robust to outliers

Descriptive Statistics – Recap (cont'd)

- In the presence of outliers (skewed data), measures of central tendency can provide a wrong picture of the distribution
  →Use **measures of dispersion** as complements
- Measures of dispersion → Min-max; Interquartile range; Variance and Standard Deviation (SD)
  - Higher variance/SD → more variability in the data
  - SD = average distance to the mean for all data points

- Variance/SD is a key ingredient of inferential statistics
  →How confident can we be that the mean in a sample represents the mean in the whole population ?
  →How confident can we be that the measured impact of a program captures the *true* impact?


Descriptive Statistics
Statistical Testing
Regression Analysis

Center for Evaluation and Development

Statistical Testing in CIE

- 2,000 people apply for program → 1,000 were randomly selected
  - Rigorous RCT → Any difference between the two groups should be attributable to the program!
- Data show the following:

	Number of individuals	Mean income
Received vocational training	1,000	1,200
Did not receive vocational training	1,000	1,050

- Did the program have an *impact* on income? How confident are we that this difference reflects the *true* impact of the program?

Statistical Testing in CIE

- We will now start getting into probability theory and more advanced statistical methods

- Statistical testing starts with a **hypothesis** that you wish to test
- The hypothesis will be guided by the research/evaluation question(s) you are interested in
- Statistical testing allows us to move past assumptions and anecdotes and see whether quantitative evidence supports our theory

Statistical Testing in CIE

Overview

Center for Evaluation and Development

We will cover the following concepts:

- Test Hypothesis and Null Hypothesis
- Sources of uncertainty in inferential statistics
  - ➢Uncertainty due to using samples → Confidence/Significance Level
  - ➢Sampling error → Standard Error
- The general steps of statistical testing
- Test the equality of means → the t-test
- Decision on a test → p-value and statistical significance
- Beyond statistical testing
- APPENDIX: Details on critical values, standard errors, t-test formulas, and the general process of statistical testing

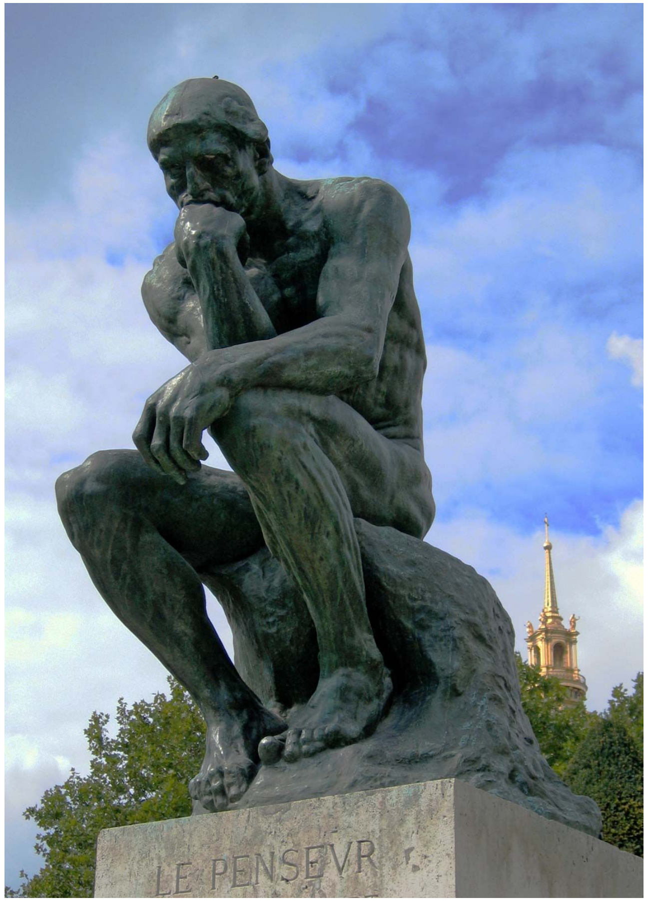
Center for Evaluation and Development

The Test Hypothesis

Statistical Testing in CIE
Hypothesis

Center for Evaluation and Development

- Best place to start with for statistical testing is thinking about a hypothesis
    - Hypothesis = a proposed explanation made as a starting point for further investigation.



Center for Evaluation and Development

Statistical Testing in CIE
The Null Hypothesis

- Strictly speaking, in statistics we do not test our hypothesis *directly*
- Formally, we formulate a **null hypothesis** (denoted H_0) and test if we can *reject* it
- The null hypothesis is usually formulated in terms of "there is no effect/no difference"
- For example:
  - ➢Our initial hypothesis: "The higher a person's education, the higher their income will be…"
  - ➢The associated Null Hypothesis: "The difference in income between those who completed primary education and those who did not is 0."

Center for Evaluation and Development

Statistical Testing in CIE
The Null Hypothesis

- Null Hypothesis: "The difference in income between those who completed primary education and those who did not is 0."

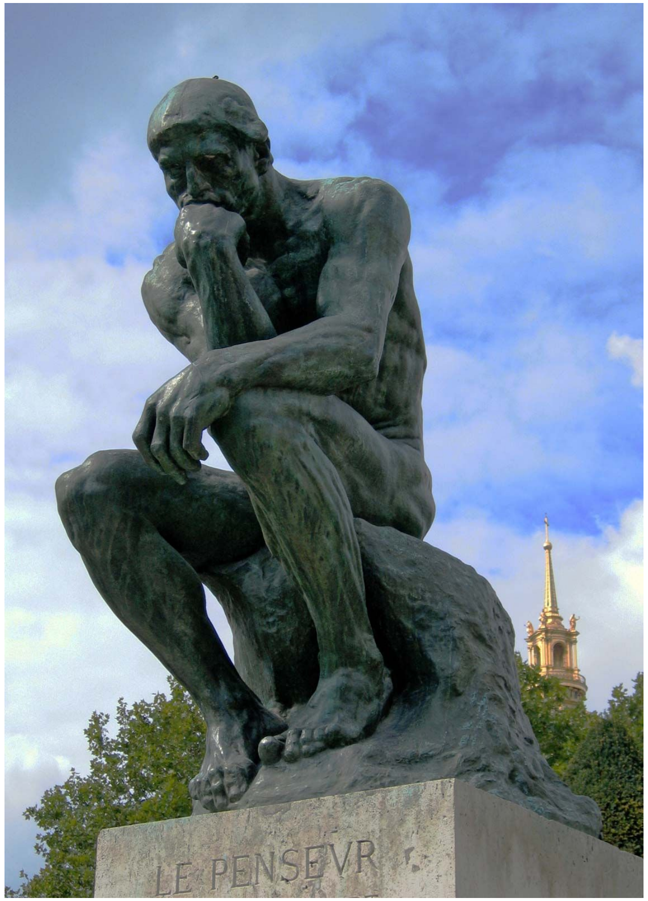
- In simple words, the statistical test tells us how *confident* we can be in *rejecting* the null hypothesis
- If we can reject the null hypothesis with sufficient confidence, we can conclude that the difference in income between the two groups is ***statistically significant*** – i.e., we are confident the observed difference reflects the truth
- While this is important from a technical point of view, for simplicity we will refer to testing the main hypothesis

Center for Evaluation and Development

Statistical Testing in CIE
Hypothesis

The higher a person's education, the higher their income will be…

- How could we go about testing this hypothesis?



Center for Evaluation and Development

Hypothesis, Population and Sample

Statistical Testing in CIE
Hypothesis, Population, Sample

- A hypothesis we want to test usually refers to a specific *population* of interest
  - ➢**Population** = complete set of all objects or persons of interest. Can be very large.
- In practice, it is very rare to have information on the whole population.

→Typically, we select a *representative* sample to learn about the population

- **Sample** = a subset of the population of interest. In inferential statistics, researchers typically use a sample to draw conclusions about the population.
  - Please refer to Year 2 training material for details on various sampling approaches

Statistical Testing in CIE

Hypothesis, Population, Sample

Center for Evaluation and Development

- <u>Hypothesis</u>: "The higher a person's education, the higher their income will be…"

→Population of interest = all the people of working age

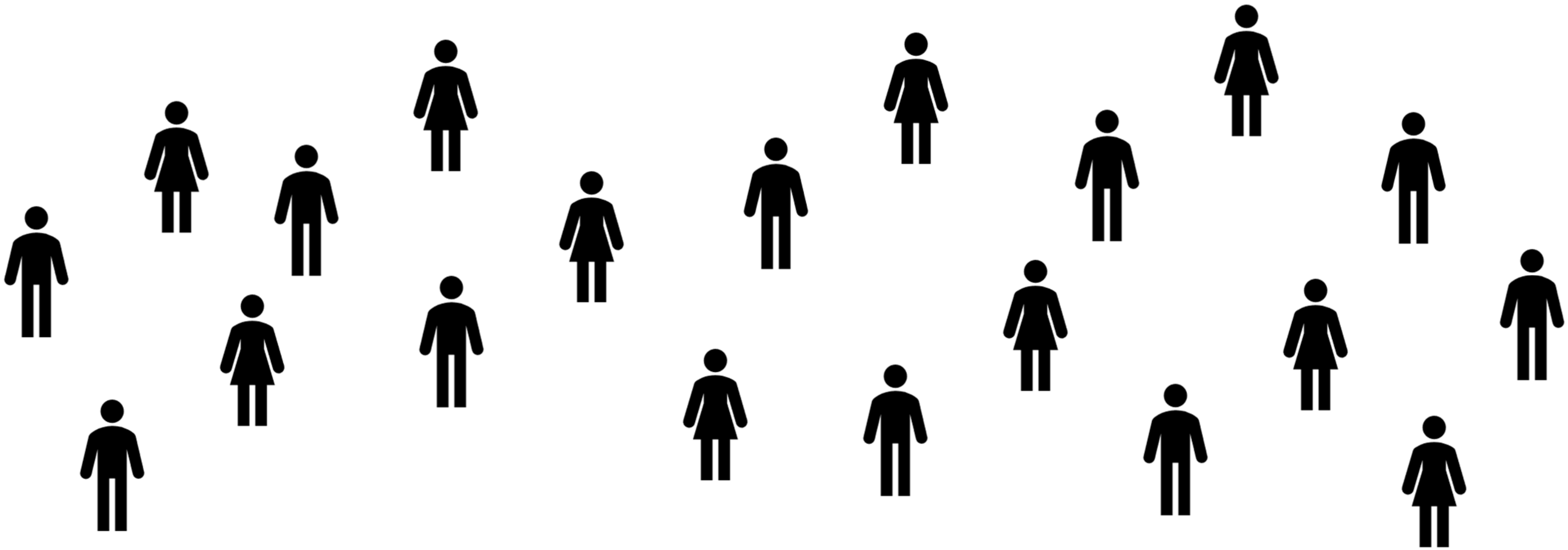
- We are unlikely to be able to gather data on education and income for the whole population of interest!

→Draw a representative sample and conduct a survey to collect the required information

Statistical Testing in CIE

Hypothesis, Population, Sample

- Assume for simplicity that this is an extremely small country of just 20 people



Statistical Testing in CIE
Hypothesis, Population, Sample

- We select a random sample of 10 people and ask them if they completed primary school – our measure of education – and their monthly income…
- Let's call it **Sample 1**

Center for Evaluation and Development

Statistical Testing in CIE

The information gathered through a survey on a selected subset of the population can be referred to as **survey data** or **sample data**

Name	Completed Primary School	Monthly Income
David	Yes	1,400
Michael	Yes	1,100
Anna	No	1,000
Monica	Yes	1,200
Emma	No	900
April	Yes	1,200
Frank	No	1,300
Daniel	Yes	1,500
Jennifer	No	800
Jodie	No	950

Center for Evaluation and Development

Statistical Testing in CIE

How to do statistical testing?

- How should we go about testing the hypothesis that people who have higher education – i.e., completed primary school – earn higher monthly income?

- Let's start by creating two groups – those who have completed primary school and those who have not – and simply compare the average monthly income between the two groups

Statistical Testing in CIE

How to do statistical testing?

Center for Evaluation and Development

Completed primary school

Name	Completed Primary School	Monthly Income
David	Yes	1,400
Michael	Yes	1,100
Monica	Yes	1,200
April	Yes	1,200
Daniel	Yes	1,500
	Average	1,280

Did not complete primary school

Name	Completed Primary School	Monthly Income
Anna	No	1,000
Emma	No	900
Frank	No	1,300
Jennifer	No	800
Jodie	No	950
	Average	990

- The sample data show the average monthly income for those who have completed primary school (1,280) is higher than for those that have not (990)
- Is this sufficient evidence to claim that your hypothesis is true?

Center for Evaluation and Development

Sources of uncertainty in inferential statistics

Center for Evaluation and Development

Statistical Testing in CIE

Uncertainty in statistical testing

People who have completed primary school earn more, on average, than those that have not

Average Monthly Income – By group

Primary Education	No primary education	*Difference*
1,280	990	*290*

- This statement is true *in our sample* – i.e., a subset of the population – but…
- …we don't know if the *averages calculated in the sample* are an accurate measure of the *true average incomes in the population*

→What if this difference exists in this specific sample only, and not in the whole population? I.e. Could this difference be due to chance?

- **Source of uncertainty #1**: trying to draw conclusions about the *whole population* based on *information from a sample*

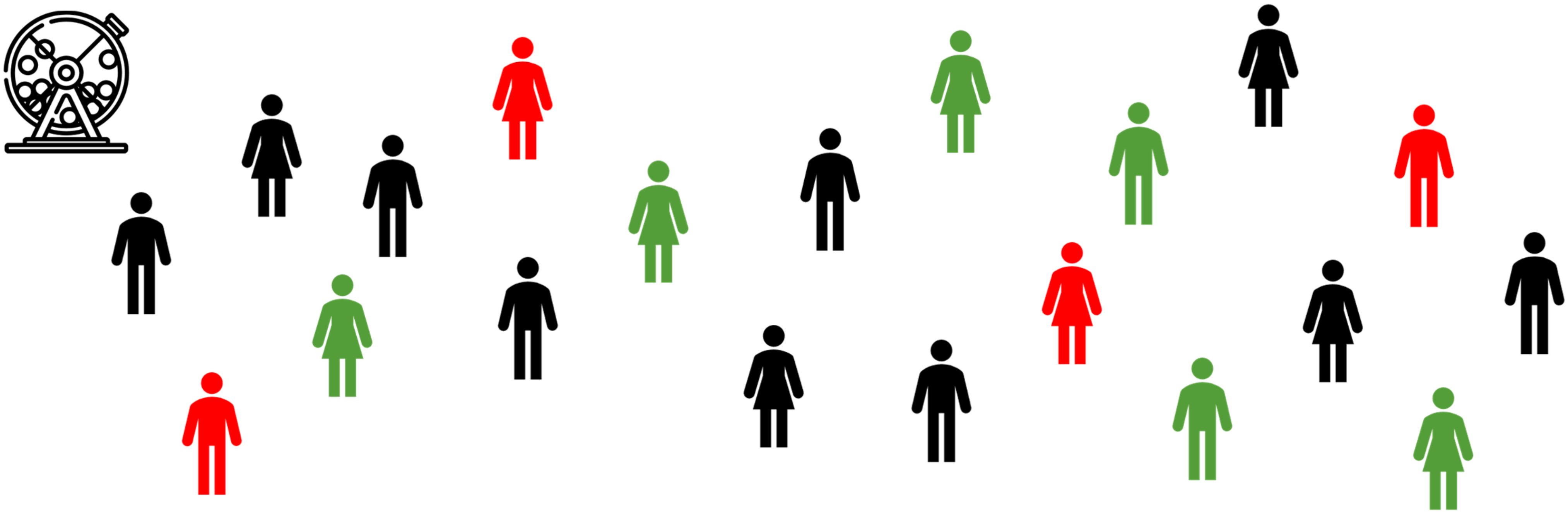
Center for Evaluation and Development

Statistical Testing in CIE
Sampling and uncertainty

- What if, by chance, we had not selected those 10 people and instead chosen another sample of 10 people from the same country?

Statistical Testing in CIE
Sampling and uncertainty

- What if, by chance, we had not selected those 10 people and instead chosen another sample of 10 people from the same country?
- Let's call it **Sample 2**



Center for Evaluation and Development

Statistical Testing in CIE
Sampling and uncertainty

The new sample data looks like this

Name	Completed Primary School	Monthly Income
John	Yes	900
Michael	Yes	1,100
Anna	No	1,000
Monica	Yes	1,200
Barbara	No	1,300
April	Yes	1,200
Frank	No	1,300
Frederic	Yes	950
Jennifer	No	800
Claire	No	1,400

Center for Evaluation and Development

Statistical Testing in CIE
Sampling and uncertainty

- Let’s compare again the average monthly income in the two groups

Completed primary school

Name	Completed Primary School	Monthly Income
John	Yes	900
Michael	Yes	1,100
Monica	Yes	1,200
April	Yes	1,200
Frederic	Yes	950
	Average	**1,070**

Did not complete primary school

Name	Completed Primary School	Monthly Income
Anna	No	1,000
Barbara	No	1,300
Frank	No	1,300
Jennifer	No	800
Claire	No	1,400
	Average	**1,160**

What has happened now?

Center for Evaluation and Development

Statistical Testing in CIE
Sampling and uncertainty

Completed primary school

Name	Completed Primary School	Monthly Income
John	Yes	900
Michael	Yes	1,100
Monica	Yes	1,200
April	Yes	1,200
Frederic	Yes	950
	Average	**1,070**

Did not complete primary school

Name	Completed Primary School	Monthly Income
Anna	No	1,000
Barbara	No	1,300
Frank	No	1,300
Jennifer	No	800
Claire	No	1,400
	Average	**1,160**

- The new sample data show the average monthly income for those who have completed primary school (1,070) is ***lower*** than for those that have not (1,160)

Statistical Testing in CIE
Sampling and uncertainty

Center for Evaluation and Development

Average Monthly Income – By group

Sample	Primary Education	No primary education	*Difference*
Sample 1	1,280	990	*290*
Sample 2	1,070	1,160	*-90*

- The two samples yield opposite conclusions!
- This example is quite extreme, but clearly makes the point:

→**Source of uncertainty #2**: *choosing a sample* generates uncertainty – this is called **sampling error**

Center for Evaluation and Development

Statistical Testing in CIE

Statistical testing and uncertainty

- To recap, there are 2 sources of uncertainty:

→**Source of uncertainty #1**: trying to draw conclusions about the *whole population* based on *information from a sample*

→**Source of uncertainty #2**: *choosing a sample* generates uncertainty – this is called **sampling error**

- The goal of inferential statistics is to account for the two types of uncertainty when testing a hypothesis
- Let's see how this works

Center for Evaluation and Development

Uncertainty in statistical testing

Center for Evaluation and Development

Statistical Testing in CIE

Uncertainty in statistical testing

- Let's use the data from Sample 1 and focus on average monthly income (irrespective of education)
- Sample mean = 1,135
- Recall our core problem here:

→ We don't know if the *average calculated in the sample* is an accurate measure of the *true average in the population*

- In other words, how close to 1,135 is the true average income in the population?
- We can answer with a **confidence interval**

Name	Monthly Income
David	1,400
Michael	1,100
Anna	1,000
Monica	1,200
Emma	900
April	1,200
Frank	1,300
Daniel	1,500
Jennifer	800
Jodie	950

Statistical Testing in CIE
Confidence Interval

- Formally, we calculate a confidence interval as:

$$\boldsymbol{Confidence\ interval\ (CI) = Sample\ mean\ \pm Z * Standard\ Error}$$

- **Confidence Interval** = it consists of calculating a lower bound and an upper bound for the *population* average based on information from a *sample*
- The CI gives a range within which we expect the "true" mean in the population to fall – with a certain *level of confidence*

- **Standard error** = a measure of how much discrepancy we expect between the mean calculated in a sample and the "true" mean in the whole population
- It accounts for uncertainty from sampling error

$$\boldsymbol{Standard\ error = \frac{Standard\ Deviation}{Square\ root\ of\ sample\ size}}$$

Statistical Testing in CIE
Confidence Interval

Center for Evaluation and Development

- Formally, we calculate a confidence interval as:

$$\boldsymbol{Confidence\ interval\ (CI) = Sample\ mean\ \pm Z * Standard\ Error}$$

- The value of ***Z*** in the formula controls the *level of confidence* we want for our range
  - ➢ It accounts for *uncertainty from using samples*
  - ➢ The most common *confidence level* in social sciences is 95% – but 90% and 99% are also common
- Let's see an example of these three statistical terms

Statistical Testing in CIE

Confidence Interval – Example

- Sample mean = 1,135;
- Standard Deviation = 226.1;
- Sample size = 10;
- SE= $226.1/\sqrt{10}$ = 71.5

- To calculate the CI for a 95% *confidence level,* we set Z=1.96 in the formula:
  - Lower bound = 994.86
  - Upper bound = 1,275.14

- **We are 95% confident that the average monthly income *in the whole population* is between 995 and 1,275**

Name	Monthly Income
David	1,400
Michael	1,100
Anna	1,000
Monica	1,200
Emma	900
April	1,200
Frank	1,300
Daniel	1,500
Jennifer	800
Jodie	950

Statistical Testing in CIE
Confidence Intervals – Example

Center for Evaluation and Development

- Sample mean = 1,135; Standard Error = 71.5

Confidence Level	*Z*	Lower bound	Upper bound	Chance of being wrong
90%	1.645	1,017.4	1,252.6	10%
95%	1.96	994.9	1,275.1	5%
99%	2.575	950.9	1,319.1	1%

- The more confident we want to be, the more *imprecision* (i.e., the wider the interval) we have to accept
- Vice versa: if we want to increase precision (tighter interval), we must accept a higher chance of being wrong…

But… Maybe there is a way to get a tighter interval without increasing the chance of being wrong?

Statistical Testing in CIE
Sample size and Uncertainty

- Can we tighten the confidence interval without increasing the chance of being wrong?

$$\boldsymbol{Confidence\ interval\ (CI) = Sample\ mean\ \pm Z * Standard\ Error}$$

$$\boldsymbol{Confidence\ interval\ (CI) = Sample\ mean\ \pm Z * \frac{Standard\ Deviation}{Square\ root\ of\ sample\ size}}$$

Hmmm…
Do you notice something that you can affect here?

Statistical Testing in CIE
Sample size and Uncertainty

Center for Evaluation and Development

- Can we tighten the confidence interval without increasing the chance of being wrong?

$$\boldsymbol{Confidence\ interval\ (CI) = Sample\ mean\ \pm Z * \frac{Standard\ Deviation}{Square\ root\ of\ sample\ size}}$$

➢ Yes, by **increasing the sample size**!

- For a *given* level of confidence:

➢ ↑ sample size ⇒ ↓ standard error ⇒ ↑ CI lower bound and ↓ CI upper bound

   ➢ **Tighter interval for the *same* level of confidence**

Center for Evaluation and Development

Statistical Testing in CIE
Sample size and Uncertainty – Example

- Let's combine the data on income from Sample 1 and Sample 2 – i.e., now we have a sample size of 14

➢Sample mean = 1,135.7

➢Standard Deviation = 223.1

➢Sample size = 14

→**95% CI = [1,019 – 1, 252.6]**

Name	Monthly Income
David	1,400
Michael	1,100
Anna	1,000
Monica	1,200
Emma	900
April	1,200
Frank	1,300
Daniel	1,500
Jennifer	800
Jodie	950
John	900
Barbara	1,300
Frederic	950
Claire	1,400

Center for Evaluation and Development

Statistical Testing in CIE

Sample size and Uncertainty – Example

	Example 1	Example 2
Sample mean	1,135	1,135.7
Standard Deviation	226.1	223.1
Sample size	10	14
Standard Error	**71.5**	**59.6**
95% Confidence Interval	**[995 – 1,275]**	**[1,019 – 1, 252.6]**

- As expected, taking a larger sample from the population decreases the standard error and produces a tighter interval
- Thanks to the extra data, we are now 95% confident that the *true* average income in the population is between 1,019 and 1,253

Center for Evaluation and Development

Statistical Testing in CIE

Uncertainty in inferential statistics – Recap

- To recap, there are 2 sources of uncertainty:

1. Uncertainty because we want to draw conclusions about the *whole population of interest* based on *information from a sample*

2. Uncertainty due to *sampling* → e.g., calculating the mean in different samples will yield different answers

- The goal of inferential statistics is to account for the two types of uncertainty

Center for Evaluation and Development

Statistical Testing in CIE

Uncertainty in inferential statistics – Recap

- Confidence Intervals are an example of inferential statistics:

1. Uncertainty due to *using samples*
   - ➢This is captured by the confidence level

2. Uncertainty due to *sampling*
   - ➢This is captured by the standard error

- Basically, statistical testing in general uses the same ingredients – standard error, confidence level – and similar intuition as confidence intervals

Center for Evaluation and Development

Statistical testing in practice

Center for Evaluation and Development

Statistical Testing in CIE
Statistical testing in practice

- The general steps of statistical testing:

➢ Formulate a null hypothesis

➢ Use sample information to calculate a **test statistic (say t-stat)**
  - ➢ Captures the uncertainty due to sampling error

➢ For a given level of confidence, find the **critical value (Z previously)**
  - ➢ Captures the uncertainty due to using samples to infer about the population

➢ Make the decision as follows:
  - Test statistic < critical value → we cannot reject the null hypothesis
  - Test statistic > critical value → we can reject the null hypothesis

Statistical Testing in CIE
Statistical testing in practice – The p-value

- In practice, you don't have to find the critical value and compare it to the t-stat yourself

→Statistical software do that for you and calculate the test's **p-value**

- **p-value** = the probability of *being wrong* when you reject the null hypothesis

Statistical Testing in CIE
Remarks on the p-value

- *Remark 1*: The p-value is a probability, so it is always between 0 and 1
- *Remark 2*: The p-value is compared to the **significance level**
- *Remark 3*: Significance level and confidence level (that we discussed before) are closely linked
  - →Confidence level = 100 – significance level

Confidence Level	*Z*	Lower bound	Upper bound	Chance of being wrong
90%	1.645	1,017.4	1,252.6	10%
95%	1.96	994.9	1,275.1	5%
99%	2.575	950.9	1,319.1	1%

Center for Evaluation and Development

Statistical Testing in CIE
p-value and decision

- Conventional significance levels in social sciences are 10%, 5% and 1%, corresponding to 90%, 95% and 99% confidence, respectively

Significance level	p-value	Interpretation	Conclusion of the test
10%	0.1, 0.09, 0.08,…	There is (at most) a 10% chance of being wrong if we reject the null hypothesis	We reject the null hypothesis with 90% confidence
5%	0.05, 0.045, 0.04,…	There is (at most) a 5% chance of being wrong if we reject the null hypothesis	We reject the null hypothesis with 95% confidence
1%	0.01, 0.0099, 0.0098,…	There is (at most) a 1% chance of being wrong if we reject the null hypothesis	We reject the null hypothesis with 99% confidence

Statistical Testing in CIE

Statistical testing in practice – The t-test

- In our example, the null hypothesis is:

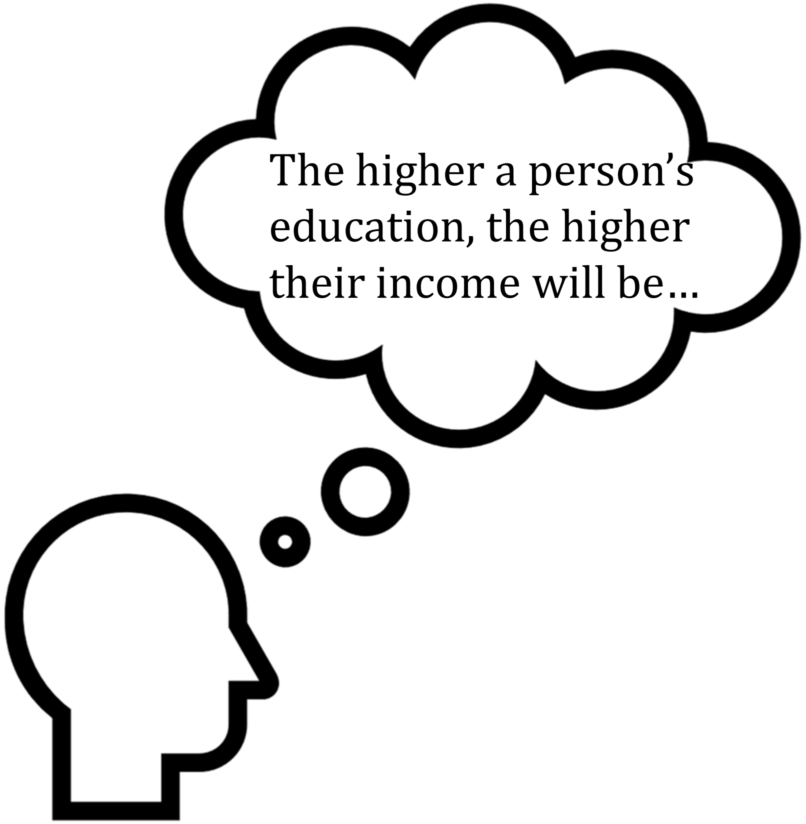
"The difference in income between those who completed primary education and those who did not is 0."

- One statistical test used to compare the mean between two groups is called a **Student's t-test** (often referred to simply as **t-test**)
- To carry out a t-test, we calculate the t-statistic (i.e., the t-test statistic) and use the p-value
- Let's see an example

Center for Evaluation and Development

Statistical Testing in CIE

p-value and decision – Example



Name	Completed Primary School	Monthly Income
David	Yes	1,400
Michael	Yes	1,100
Monica	Yes	1,200
April	Yes	1,200
Daniel	Yes	1,500
	Average	1,280

Name	Completed Primary School	Monthly Income
Anna	No	1,000
Emma	No	900
Frank	No	1,300
Jennifer	No	800
Jodie	No	950
	Average	990

Statistical Testing in CIE
p-value and decision – Example

Average Monthly Income – By group

Primary Education	**No primary education**	***Difference***	***p-value from t-test***
1,280	990	*290*	***0.032***

- We run a t-test on our sample data and find a p-value of **0.032**

→What do we conclude on the test?

Center for Evaluation and Development

Statistical Testing in CIE

p-value and decision – Example

Average Monthly Income – By group

Primary Education	No primary education	*Difference*	*p-value from t-test*
1,280	990	*290*	***0.032***

→What do we conclude on the t-test?

- In words, there is a 3.2% chance that the difference between the groups is due to random chance and does not exist in the population
- In other words, we can say with at least 95% confidence that people who complete primary school have a higher income in the *population*
- In a report: "The difference in average income between the two groups is *statistically significant* at the 5% (significance) level."

Center for Evaluation and Development

Statistical Testing in CIE
Recap

- Statistical testing exploits information from a *sample* to draw conclusions (i.e., infer) about the *population*

- Define your hypothesis and the associated null hypothesis

- Set the desired significance/confidence level

- Use the p-value to conclude → the smaller the p-value, the more confident we are to reject the null hypothesis

Center for Evaluation and Development

Statistical Testing in CIE
Remarks

- *Remark 1*: We focused on a test to compare means, but many other tests exist depending on the type of data (continuous, categorical) and what we want to compare (variance, median, etc.)

- *Remark 2*: The logic presented here is valid for other statistical tests. In other words, if you know the null hypothesis and the p-value, you can conclude!

Statistical Testing in CIE
Remarks

- *Remark 3*: We saw that increasing sample size could reduce uncertainty and improve confidence intervals. The same intuition holds for the t-test:
  - Larger sample size ⇒ information on a larger part of the population ⇒ reduce uncertainty due to using samples
  - Larger sample size ⇒ reduce uncertainty due to sampling
  - Overall, the larger the sample size ⇒ the more confident we are about our conclusion

Center for Evaluation and Development

Beyond statistical testing

Center for Evaluation and Development

Statistical Testing in CIE

A word of caution

Caution – Part 1

→A statistical test is a formula, it does not know the **context** of the analysis, where or how you got your data

- Researchers need to do the hard work to provide the formula with strong data and contextualize the findings

→The first step in bringing credibility/confidence to your tests is to carefully select the sample!

  - Please refer to see Year 2 training material on sampling

Statistical Testing in CIE
A word of caution

Caution – Part 2

- Simply testing the difference in the average income tells you there *exists* a difference between two groups

→It does ***not*** tell you that primary education leads to higher incomes, or that the difference in incomes is solely due to education

→A statistical test alone does ***not*** establish <u>causality</u>

Center for Evaluation and Development

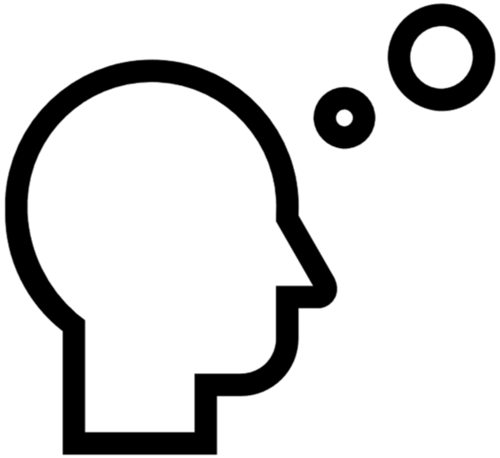
Statistical Testing in CIE

Statistical testing and causality

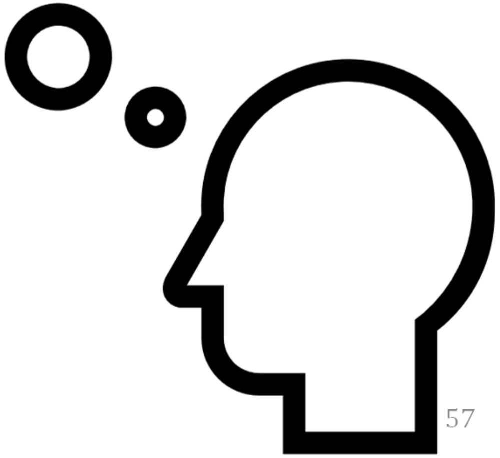
People that have completed primary school will have a higher average income than those that have not

→

Finishing primary school will increase your income



- Can we prove the second hypothesis with our current analysis? → NO
- Relevant for programme evaluation because we are interested in causality! → CIE



Statistical Testing in CIE

Statistical testing and causality

- In our sample, the group that have completed primary school counts 3 men and 2 women
- The group without primary education counts 4 women and only 1 man
- Should this be considered when testing our hypothesis?

Name	Completed Primary School	Monthly Income
David	Yes	1,400
Michael	Yes	1,100
Monica	Yes	1,200
April	Yes	1,200
Daniel	Yes	1,500
	Average	1,280

Name	Completed Primary School	Monthly Income
Anna	No	1,000
Emma	No	900
Frank	No	1,300
Jennifer	No	800
Jodie	No	950
	Average	990

Center for Evaluation and Development

Statistical Testing in CIE
Statistical testing and causality

- Armed only with our t-test, we cannot confirm that the observed difference in incomes is not caused by factors other than education, such as e.g., gender-driven phenomena (e.g., gender pay gap)

- How can we deal with this?

➢ Careful CIE design (Year 1) and careful sampling (Year 2)
➢ Regression analysis → our next topic

Center for Evaluation and Development

European Commission

END OF SESSION 3a

Center for Evaluation
and Development

Statistical Testing in CIE

Appendix

APPENDIX
*Critical Value **Z** in Confidence Intervals*

Confidence interval = Sample mean +/- *Z* x Standard Error

- ***Z*** = critical value
- It comes from a known (theoretical) statistical distribution and changes with the desired significance level/level of confidence
- For t-tests, in small samples (N < 30), the critical values are taken from **Student's *t* distribution** which accounts for the small sample size
- For samples with N > 30, critical values are based on the (standardized and centered) Normal distribution
  - We did not use Student's *t* critical values in our examples because in practice we rarely have a sample with fewer than 30 observations, and the critical values based on Normal distribution are popular numbers that we wanted to show

APPENDIX
One-sample t-test formula

- One-sample t-test → used to test whether the population mean is equal to a specific value μ

$$t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$\bar{x}$ = sample mean

μ = hypothesized value for the *population* mean (can be 0)

σ = standard deviation (of the sample mean)

n = sample size

$\sigma / \sqrt{n}$ = standard error

Center for Evaluation and Development

APPENDIX

Two-sample t-test formula

- Two-sample t-test → used to test whether the mean is statistically significantly different between two samples
  - Note: can be two different populations or two different groups of the same population

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{x}_i$ = mean in sample *i* = 1,2

σ_i^2 = sample variance (i.e., square of standard deviation) for sample *i* = 1,2

n_i = size of sample *i* = 1,2

$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ = an estimate of the standard error of the two combined samples

Center for Evaluation and Development

APPENDIX – General Steps of Statistical Testing

Center for Evaluation and Development

Statistical Testing in CIE
Statistical testing in practice

- The general steps of statistical testing:

➢ Formulate a null hypothesis

➢ Use sample information to calculate a **test statistic**
   - ➢ Captures the uncertainty due to sampling

➢ For a given level of confidence, find the **critical value**
   - ➢ Captures the uncertainty due to using samples to infer about the population

➢ Make the decision as follows:
   - ❑ Test statistic < critical value → we cannot reject the null hypothesis
   - ❑ Test statistic > critical value → we can reject the null hypothesis

Center for Evaluation and Development

Statistical Testing in CIE

Statistical testing in practice – The t-test

- In our example, the null hypothesis is:

"The difference in income between those who completed primary education and those who did not is 0."

- The statistical test used to compare the mean between two groups is called a **t-test**

- The associated test statistic is called **t-statistic** (or **t-stat** for short), or **t-test value** (see appendix for the technical details)

Center for Evaluation and Development

Statistical Testing in CIE

Statistical testing in practice – The t-test

- The statistical test used to compare the mean between two groups is called a **t-test**, for which we calculate the t-statistic (**t-stat**)
- Example:
  - Null hypothesis: “The difference in income between those who completed primary education and those who did not is 0.”

  - t-stat < critical value → Cannot reject the null hypothesis

  →<u>Intuition</u>: not enough evidence to say with confidence that primary school graduates and non-graduates earn different incomes on average

  - t-stat < critical value → Can reject the null hypothesis

  →<u>Intuition</u>: we have enough evidence to say with confidence that primary school graduates and non-graduates earn different incomes on average

Center for Evaluation and Development

APPENDIX – Standard Error

Statistical Testing in CIE
Sampling and uncertainty

- Recall our core problem here:

→We don't know if the *average calculated in the sample* is an accurate measure of the *true average in the population*

- We can get a sense of this accuracy by calculating a **confidence interval** – i.e., using sample data, how confident we are that the true mean in the population falls within a certain range of values
- The key building block of the confidence interval is the **standard error**, so let's start there

Center for Evaluation and Development

Statistical Testing in CIE
Sampling and uncertainty – Standard error

- **Standard error** = a measure of how much discrepancy we expect between the mean calculated in a sample and the "true" mean in the whole population
- Intuition:
  - Imagine we calculate the mean income in many different samples of same size from our population
  - We expect the mean will be different from sample to sample – as it was in our example
  - The standard error tells us how much variability we can expect in the values of the means calculated across the different samples

Center for Evaluation and Development

Statistical Testing in CIE
Sampling and uncertainty – Standard error

- **Standard error** = a measure of how much discrepancy we expect between the mean calculated in a sample and the "true" mean in the whole population
- Formally, the standard error of the mean is calculated as:

$$\boldsymbol{Standard\ error = \frac{Standard\ Deviation}{Square\ root\ of\ sample\ size}}$$

- Let's try an example to get a better sense of it

Statistical Testing in CIE
Example – Standard error

- Let's use the sample data from Sample 1 and focus on average monthly income (irrespective of education)

Name	Monthly Income
David	1,400
Michael	1,100
Anna	1,000
Monica	1,200
Emma	900
April	1,200
Frank	1,300
Daniel	1,500
Jennifer	800
Jodie	950

Sample mean = 1,135

Standard Deviation = 226.1

Sample size = 10

→**Standard Error = 71.5**

Center for Evaluation and Development

Statistical Testing in CIE

Example – Standard error

- Sample mean = 1,135
- Standard Error = 71.5

- The standard error tells us the following:
  - ➢ If we take many different (random) samples of 10 people from our population, and calculate the mean in each sample, it will be equal to 1,135 +/- 71.5
  - ➢ In other words, the mean calculated in any (random) sample of 10 people from this population will be between 1,063.5 and 1,206.5
- Standard error is useful in inferential statistics to calculate confidence intervals or test statistics

Center for Evaluation and Development

European Commission

Session 3b: Guided walkthrough of a t-test in Excel

C4ED – EUTF

October 2023

Center for Evaluation and Development

T-test – Recap

- The **t-test** is a method of inferential statistics that allows to carry out tests on *means*

- **One-sample t-test:** Use sample data to test whether the *population* mean is equal to a specific value – e.g., equal to 0

- **Two-sample t-test:** Use sample data to test whether the mean is equal/different between *two groups of the population*

→ The following examples focus on the ***two-sample t-test***

Center for Evaluation and Development

Setup

- Please open the Excel file called "3b_Example_Excel"
- The document includes several worksheets

➢ "Example 1" = Small sample (30 observations in total), extract of survey data on income by education level – Completed Primary Education vs. Completed Secondary Education

➢ "Example 2" = Full extract of real-world survey data with the following variables:

- The data used in these examples come from the EUTF Impact Evaluation study in Uganda

Center for Evaluation and Development

Example 1

Table 1 - Extract of survey data on income, by education level

Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Table 2 - Descriptive Statistics

	Average Income	Standard Deviation
Primary Education		
Secondary Education		

Table 3 - Comparison of Means

Difference in means	
p-value (t-test)	

Let's fill **Table 2**.

Center for Evaluation and Development

Example 1 – Descriptive Statistics

Mean

G4

Insert Function button here

Table 1 - Extract of survey data on income, by education level

Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Table 2 - Descriptive Statistics

	Average Income	Standard Deviation
Primary Education		
Secondary Education		

Table 3 - Comparison of Means

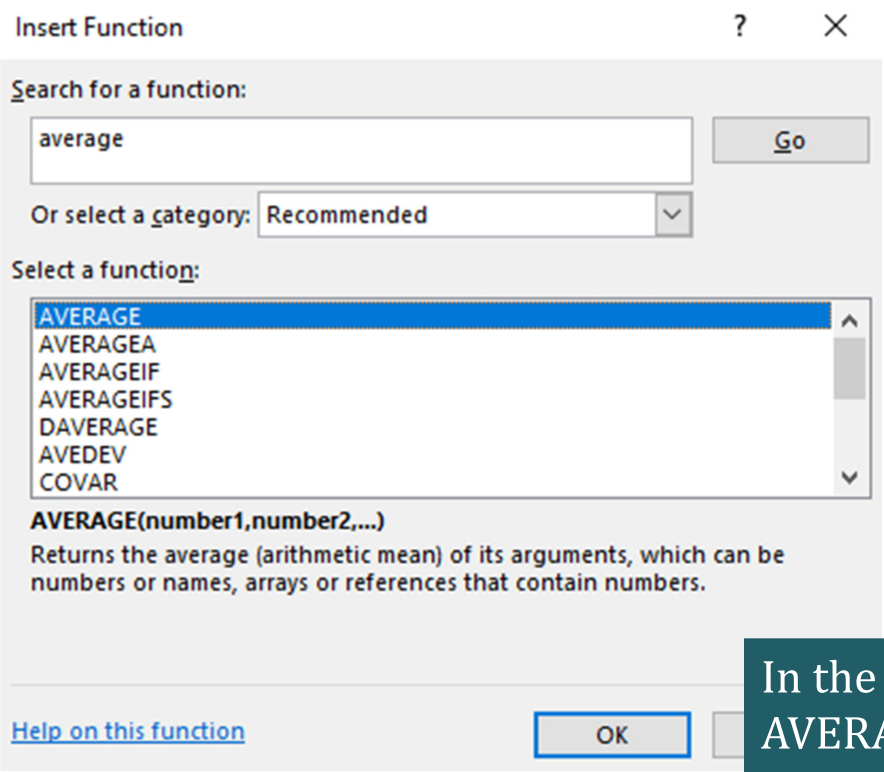
Difference in means	
p-value (t-test)	

In Table 2, select the cell corresponding to the **Average Income of people with Primary Education**, and click on “Insert Function” (LHS of the formula bar).

Center for Evaluation and Development

Example 1 – Descriptive Statistics

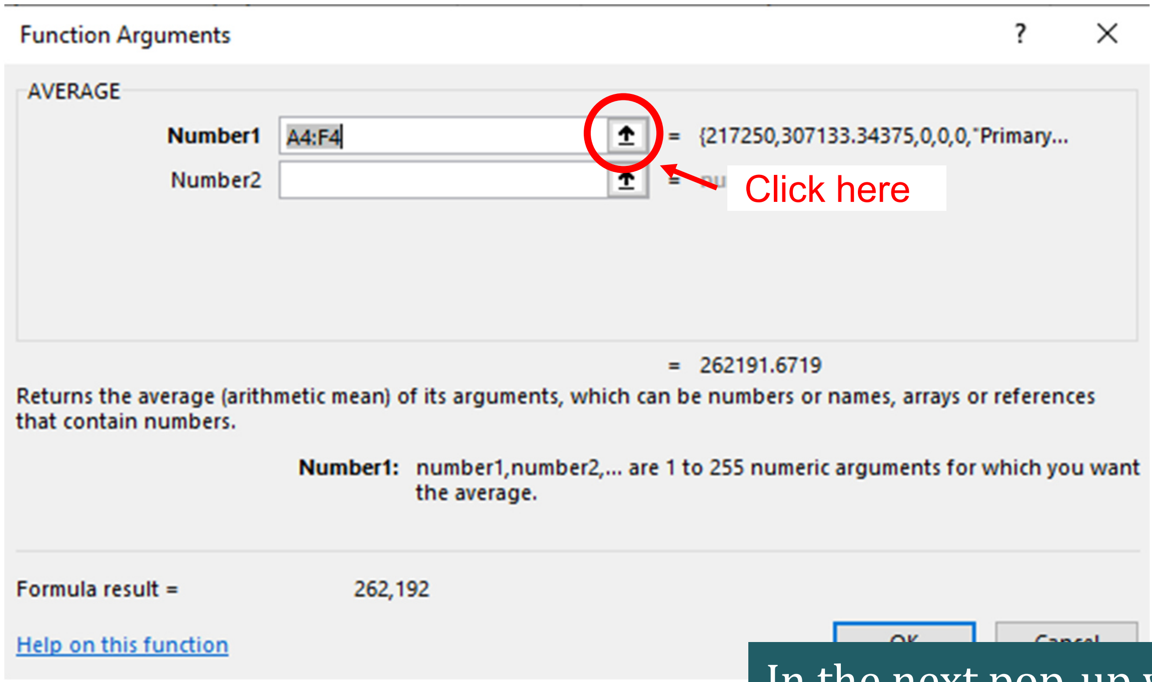
Mean



In the pop-up window that appears, select the function AVERAGE and click on "OK"
[HINT: in case the function does not appear in the list, type "average" in the search bar and click on "Go"]

Center for Evaluation and Development

Example 1 – Descriptive Statistics
Mean

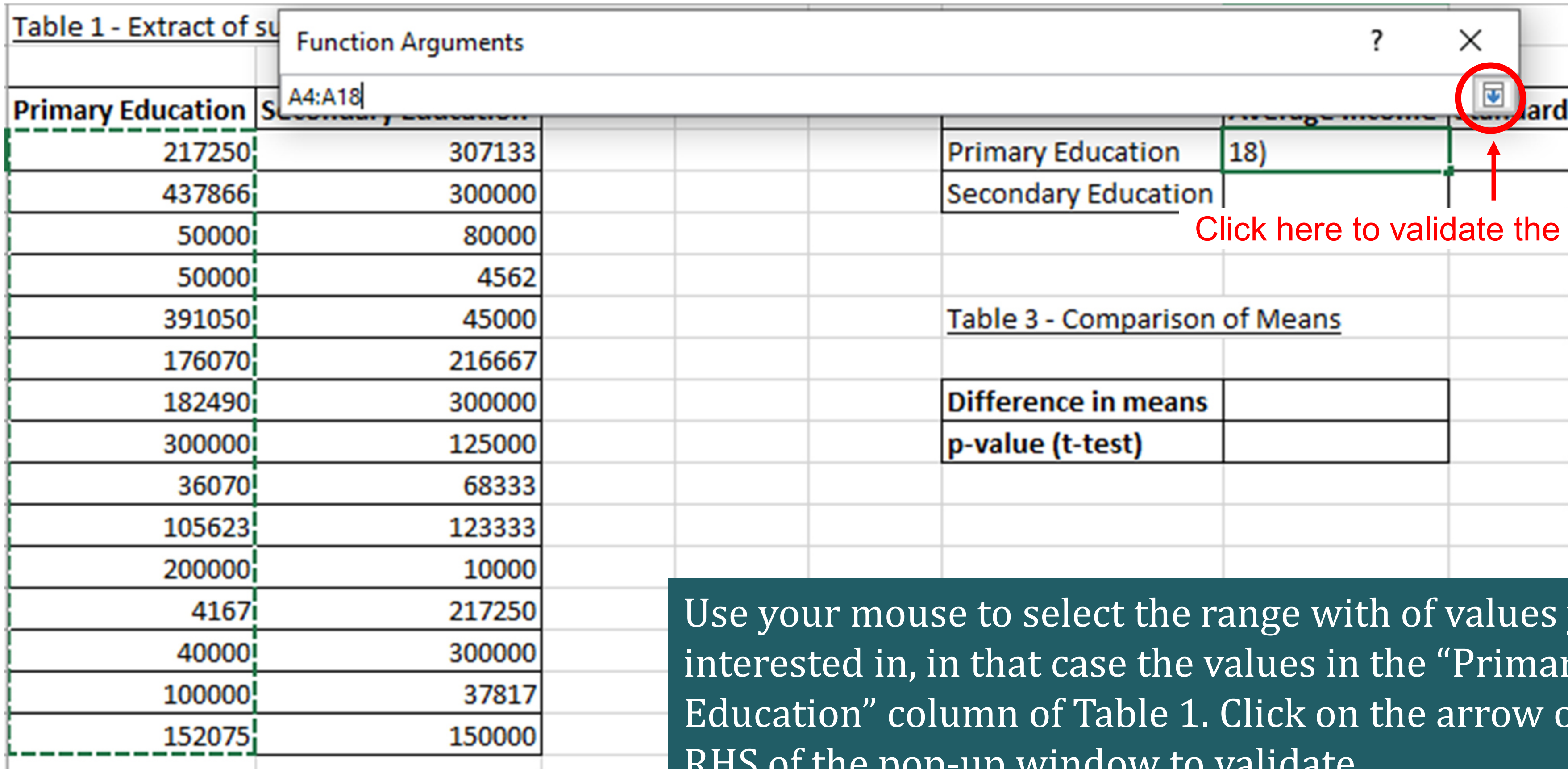


In the next pop-up window that appears, click on the arrow on the RHS of the "Number1" field

Example 1 – Descriptive Statistics

Mean

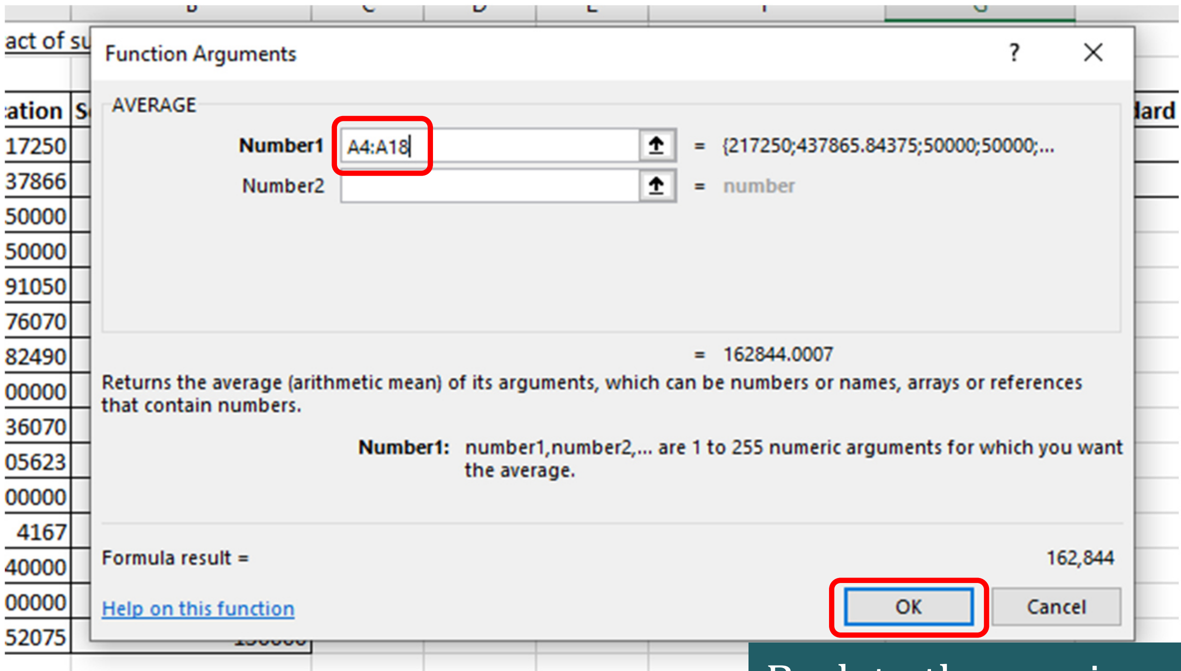
Center for Evaluation and Development



Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Use your mouse to select the range with of values you're interested in, in that case the values in the "Primary Education" column of Table 1. Click on the arrow on the RHS of the pop-up window to validate.

Example 1 – Descriptive Statistics
Mean



Back to the previous pop-up window, click on “OK” to validate the selected range of values.

Example 1 – Descriptive Statistics

Mean

=AVERAGE(A4:A18)

Table 1 - Extract of survey data on income, by education level

Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Table 2 - Descriptive Statistics

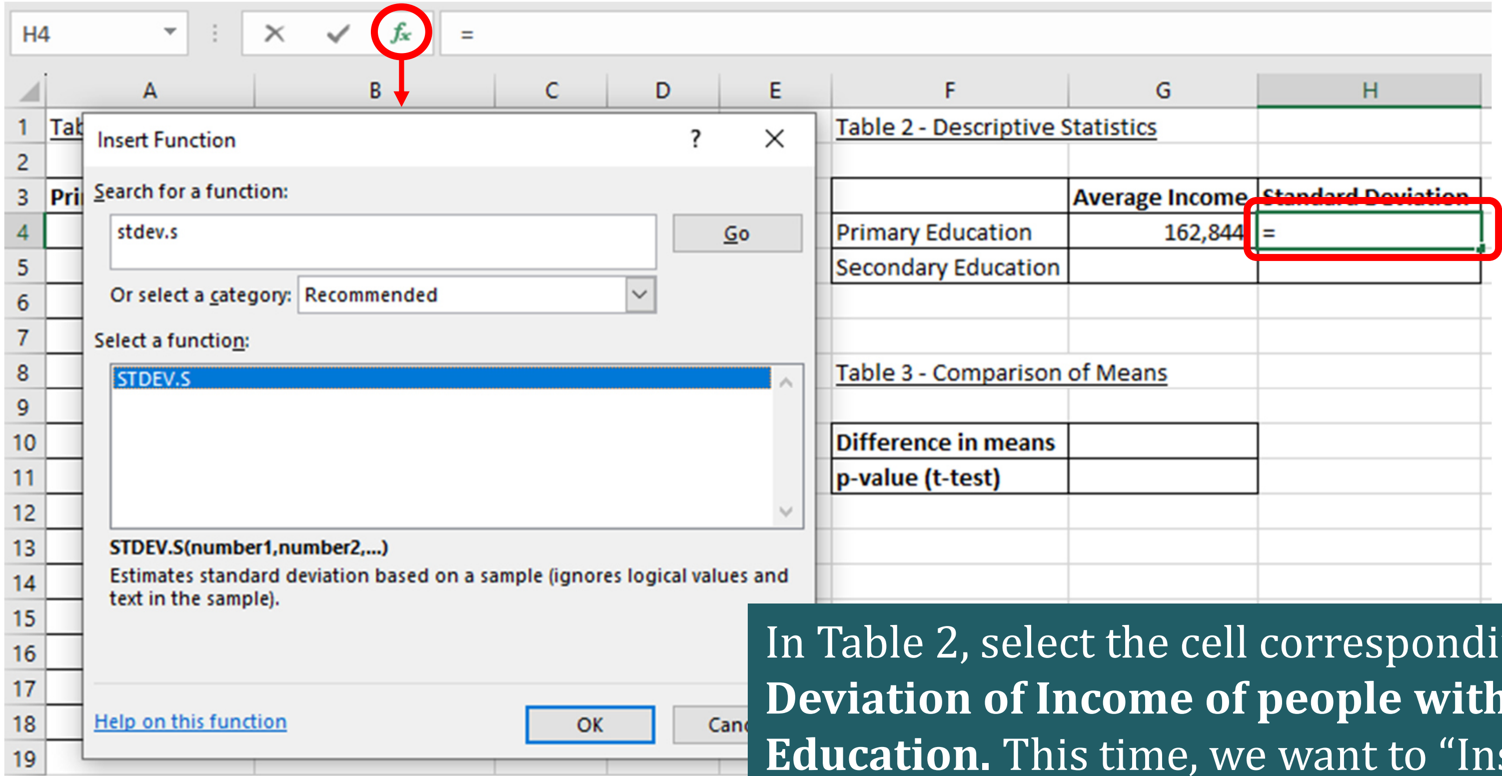
	Average Income	Standard Deviation
Primary Education	162,844	
Secondary Education		

Table 3 - Comparison of Means

Difference in means	
p-value (t-test)	

Now, we see the calculated value in Table 2. Note that the formula appears in the formula bar – in Excel you can type formulas directly into cells.
Let's turn to Standard Deviation.

Example 1 – Descriptive Statistics
Standard Deviation

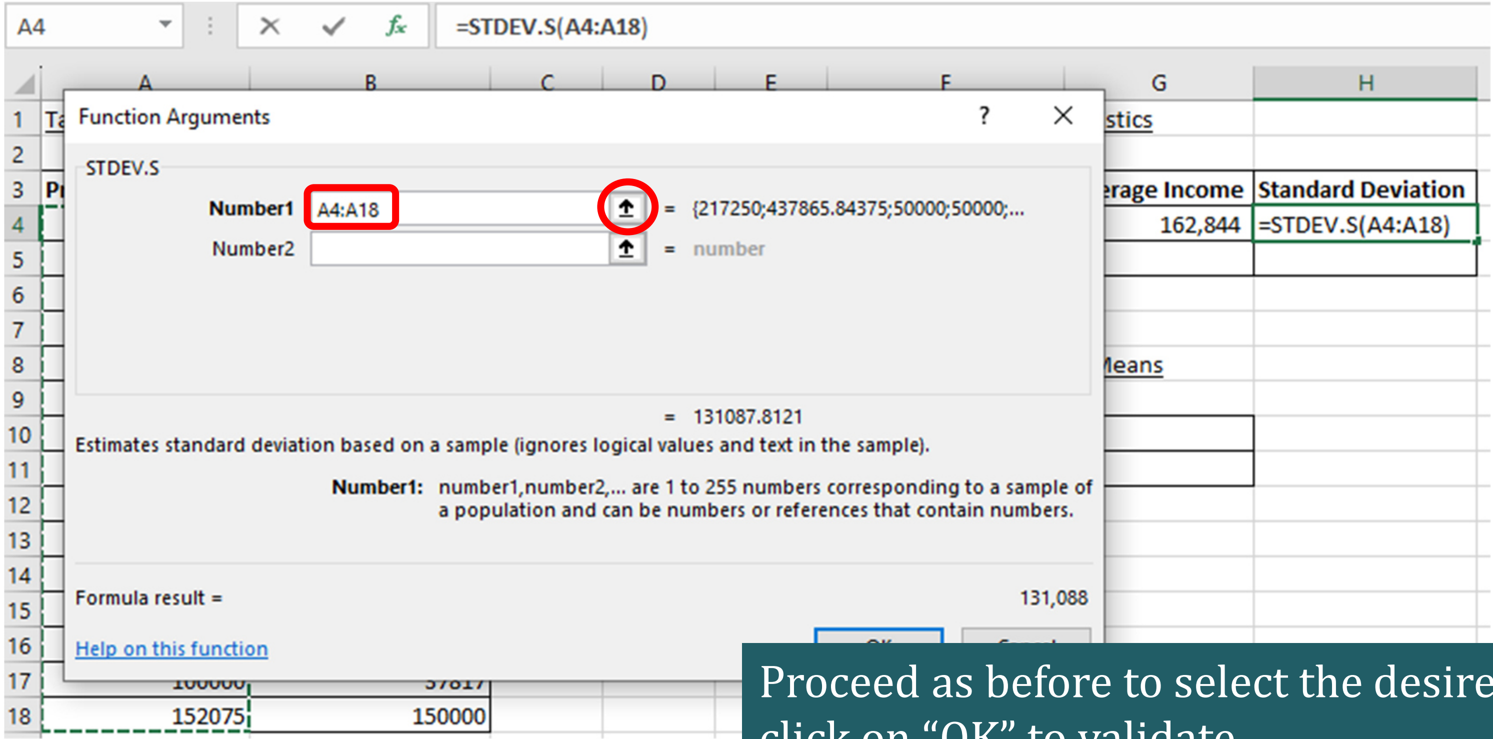


In Table 2, select the cell corresponding to the **Standard Deviation of Income of people with Primary Education.** This time, we want to "Insert Function" called STDEV.S.

Center for Evaluation and Development

Example 1 – Descriptive Statistics

Standard Deviation



Proceed as before to select the desired range of values and click on “OK” to validate.

Center for Evaluation and Development

Example 1 – Descriptive Statistics

Table 1 - Extract of survey data on income, by education level

Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Table 2 - Descriptive Statistics

	Average Income	Standard Deviation
Primary Education	162,844	131,088
Secondary Education	152,340	112,866

Table 3 - Comparison of Means

Difference in means	
p-value (t-test)	

You can follow the same steps to fill the values for average and standard deviation of income for people with Secondary Education.

Center for Evaluation and Development

Example 1

Table 1 - Extract of survey data on income, by education level

Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Table 2 - Descriptive Statistics

	Average Income	Standard Deviation
Primary Education		
Secondary Education		

Table 3 - Comparison of Means

Difference in means	
p-value (t-test)	

Let's fill **Table 3**.

Example 1 – Descriptive Statistics

Difference in means

Center for Evaluation and Development

Table 1 - Extract of survey data on income, by education level

Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Table 2 - Descriptive Statistics

	Average Income	Standard Deviation
Primary Education	162,844	131,088
Secondary Education	152,340	112,866

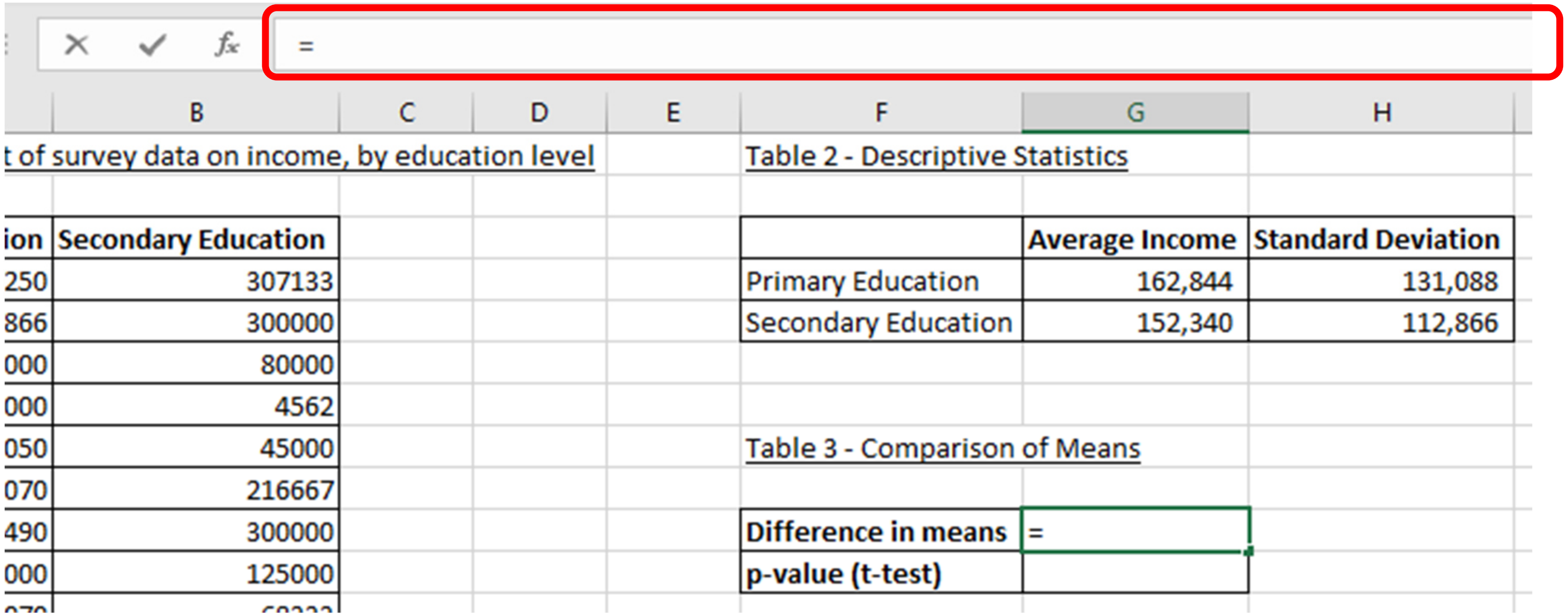
Table 3 - Comparison of Means

Difference in means	
p-value (t-test)	

We will use a **formula** to calculate the **Difference in means.**

In Table 3, select the cell corresponding to the **Difference in means**, and click in the formula bar.

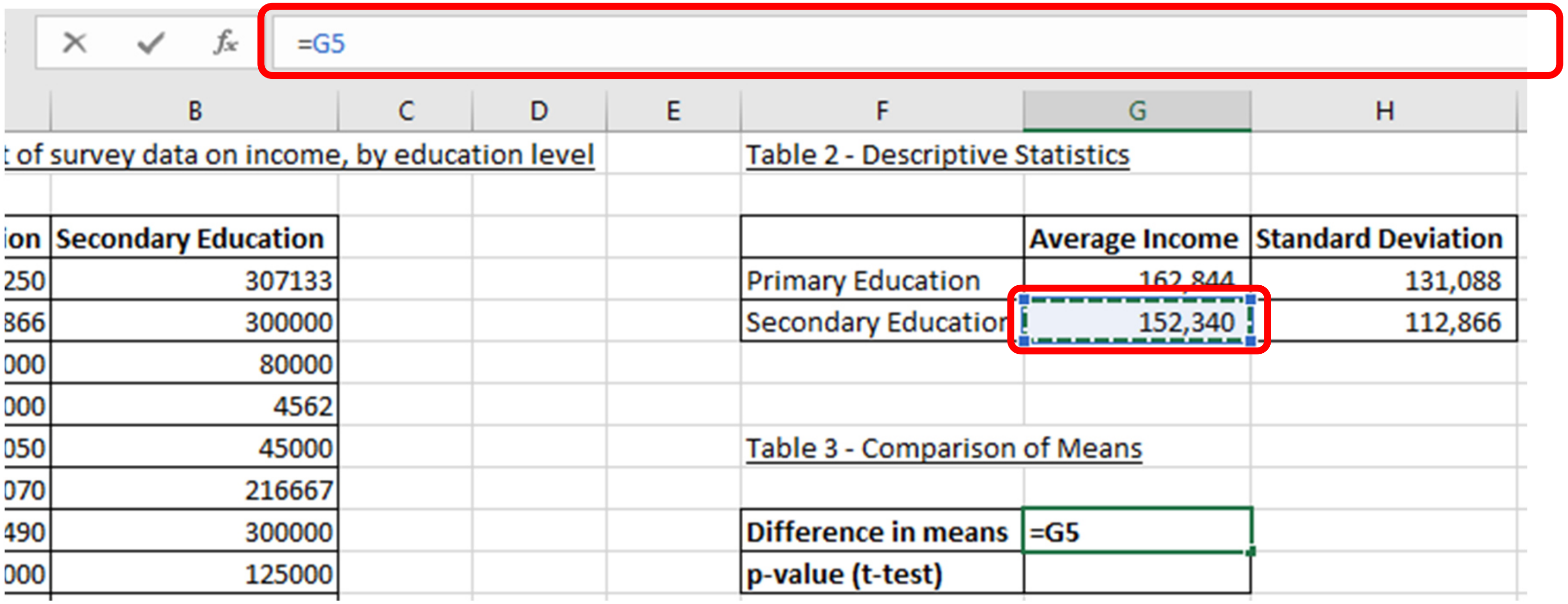
Example 1 – Descriptive Statistics
Difference in means



1. In the formula bar, type "=".

Example 1 – Descriptive Statistics
Difference in means

Center for Evaluation and Development



=G5

t of survey data on income, by education level

ion	Secondary Education
250	307133
866	300000
000	80000
000	4562
050	45000
070	216667
490	300000
000	125000

Table 2 - Descriptive Statistics

	Average Income	Standard Deviation
Primary Education	162,844	131,088
Secondary Education	152,340	112,866

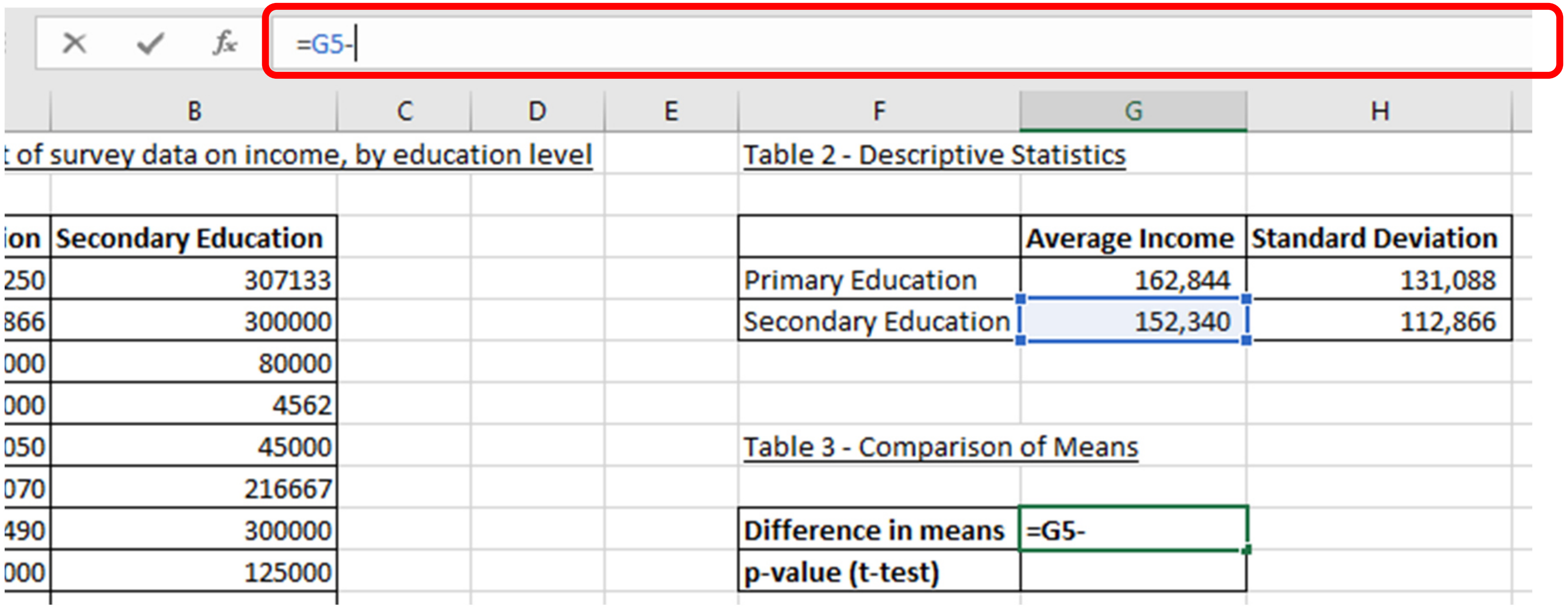
Table 3 - Comparison of Means

Difference in means	=G5
p-value (t-test)	

1. In the formula bar, type "=".
2. With your mouse, select cell G5 – Average income, secondary educ.

Center for Evaluation and Development

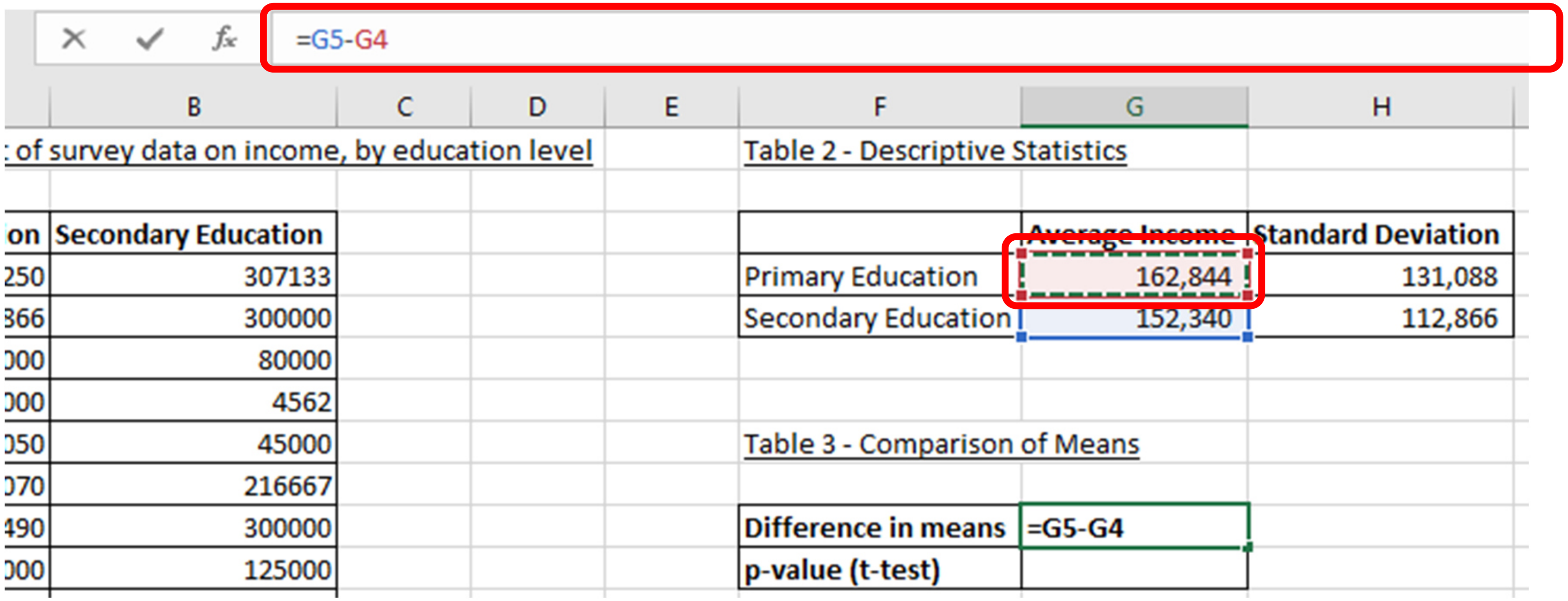
Example 1 – Descriptive Statistics
Difference in means



1. In the formula bar, type "=".
2. With your mouse, select cell G5 – Average income, secondary educ.
3. Then, in the formula bar, type "-".

Center for Evaluation and Development

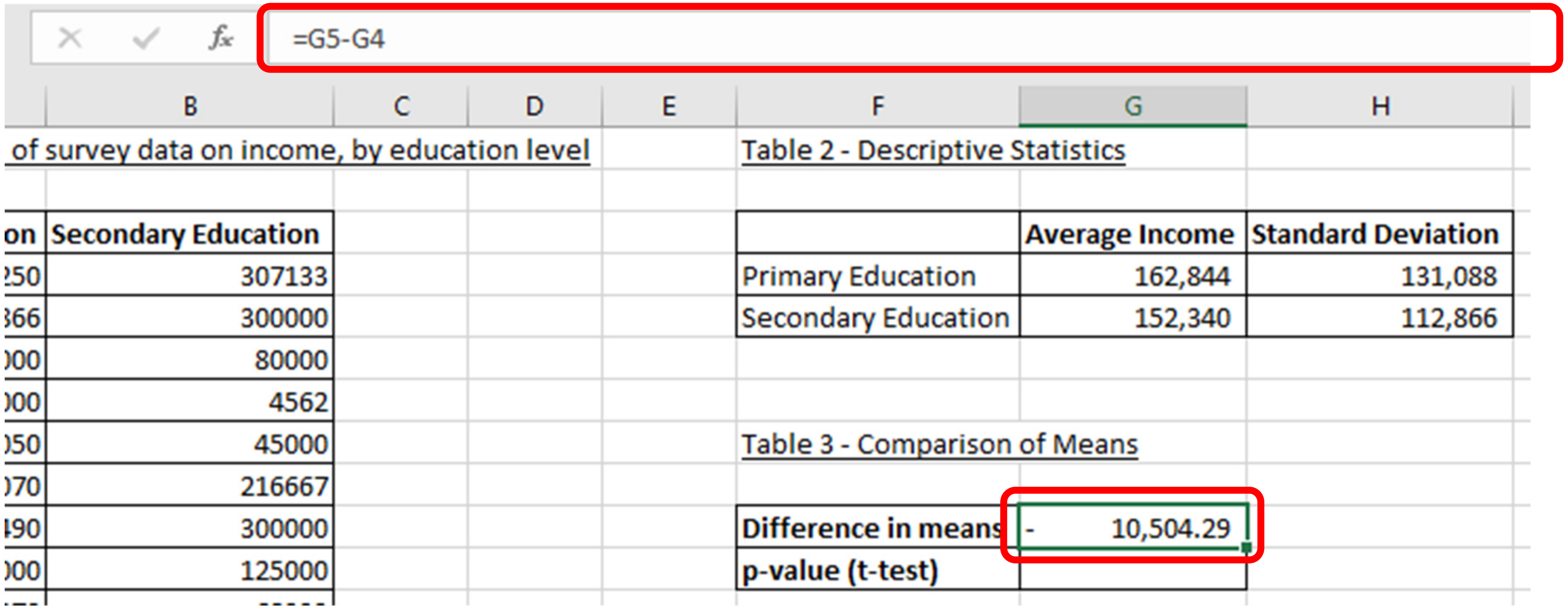
Example 1 – Descriptive Statistics
Difference in means



1. In the formula bar, type “=“.
2. With your mouse, select cell G5 – Average income, secondary educ.
3. Then, in the formula bar, type “-”.
4. With your mouse, select cell G4 – Average income, primary educ.

Example 1 – Descriptive Statistics
Difference in means

Center for Evaluation and Development


=G5-G4

of survey data on income, by education level

on	Secondary Education
250	307133
366	300000
000	80000
000	4562
050	45000
070	216667
490	300000
000	125000

Table 2 - Descriptive Statistics

	Average Income	Standard Deviation
Primary Education	162,844	131,088
Secondary Education	152,340	112,866

Table 3 - Comparison of Means

Difference in means	- 10,504.29
p-value (t-test)	

1. In the formula bar, type "=".
2. With your mouse, select cell G5 – Average income, secondary educ.
3. Then, in the formula bar, type "-".
4. With your mouse, select cell G4 – Average income, primary educ.
5. Press "Enter".

Center for Evaluation and Development

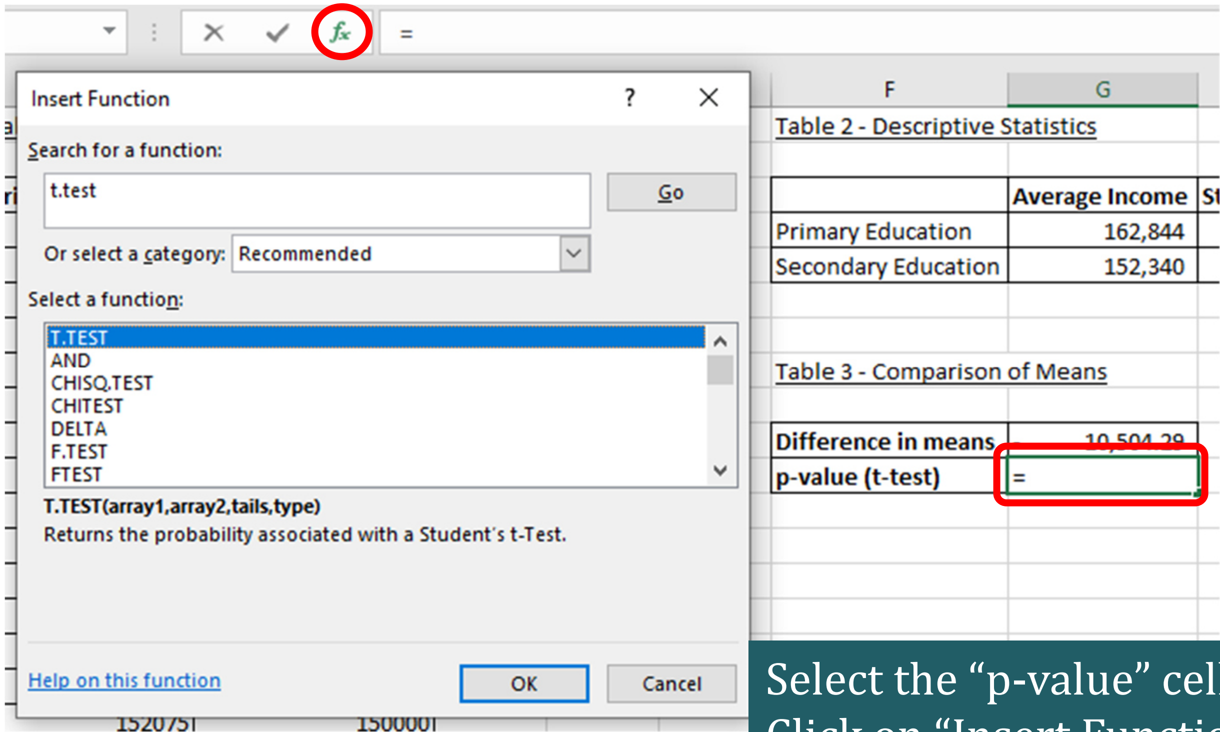
Example 1 – Statistical Testing
t-test

- Now, let's do a t-test to compare the means between the two groups and see whether they are statistically significantly different from each other in the *population*.

- The easiest way to do a t-test in Excel is to use the T.TEST function – you can do it with "Insert Function".
- T.TEST gives you the p-value of the test.

Example 1 – Statistical Testing

t-test



Select the “p-value” cell in Table 3.
Click on “Insert Function” and select T.TEST.

Center for Evaluation and Development

Example 1 – Statistical Testing

t-test

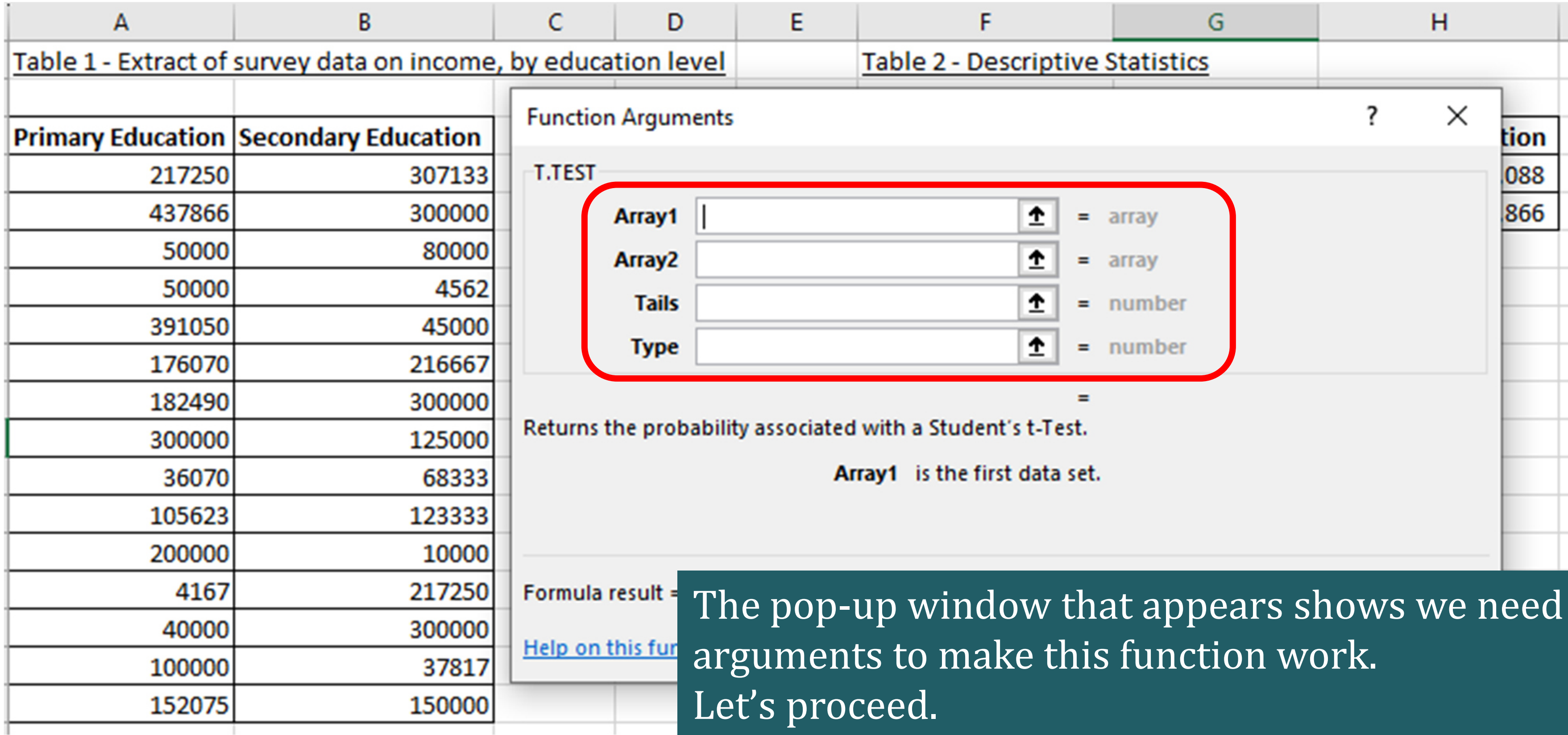


Table 1 - Extract of survey data on income, by education level

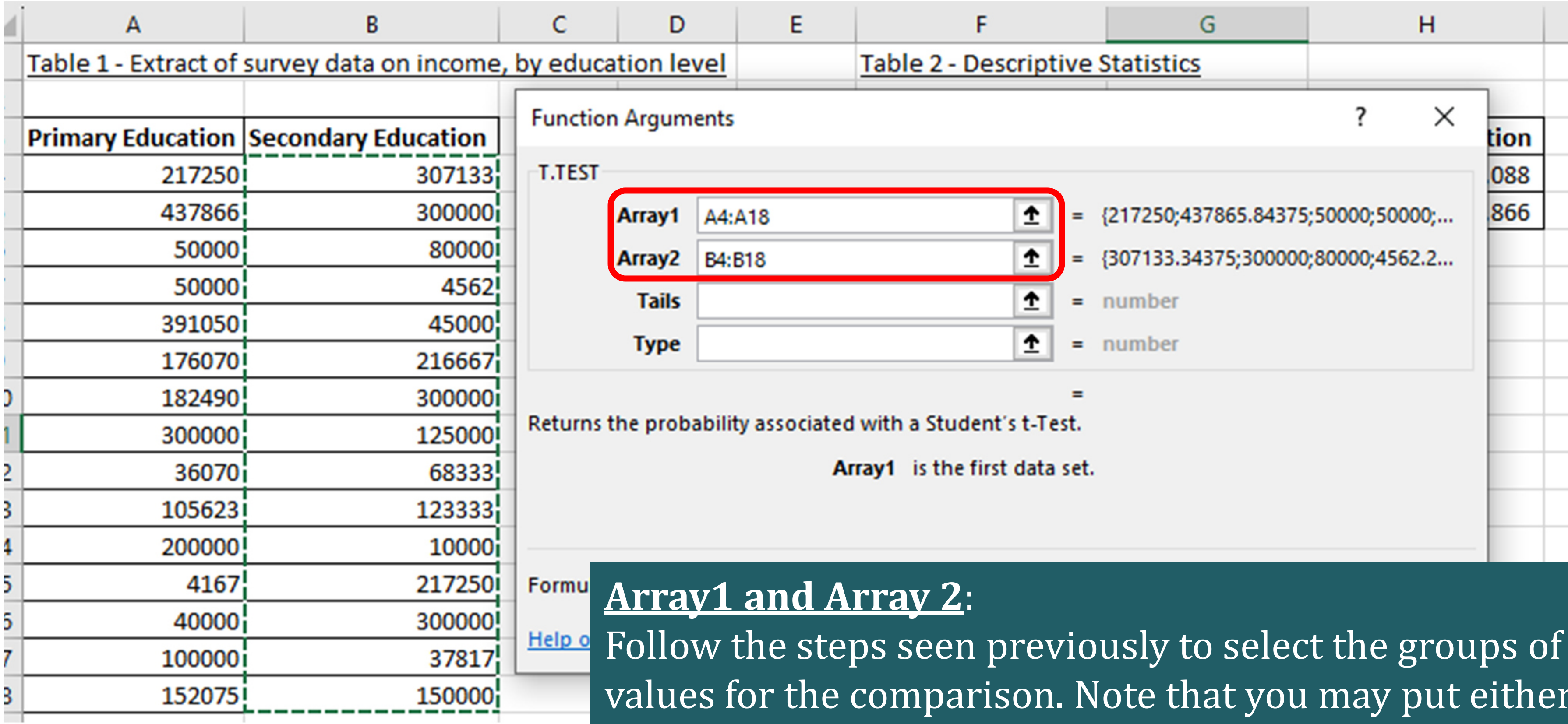
Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

The pop-up window that appears shows we need 4 arguments to make this function work.
Let's proceed.

Center for Evaluation and Development

Example 1 – Statistical Testing

t-test

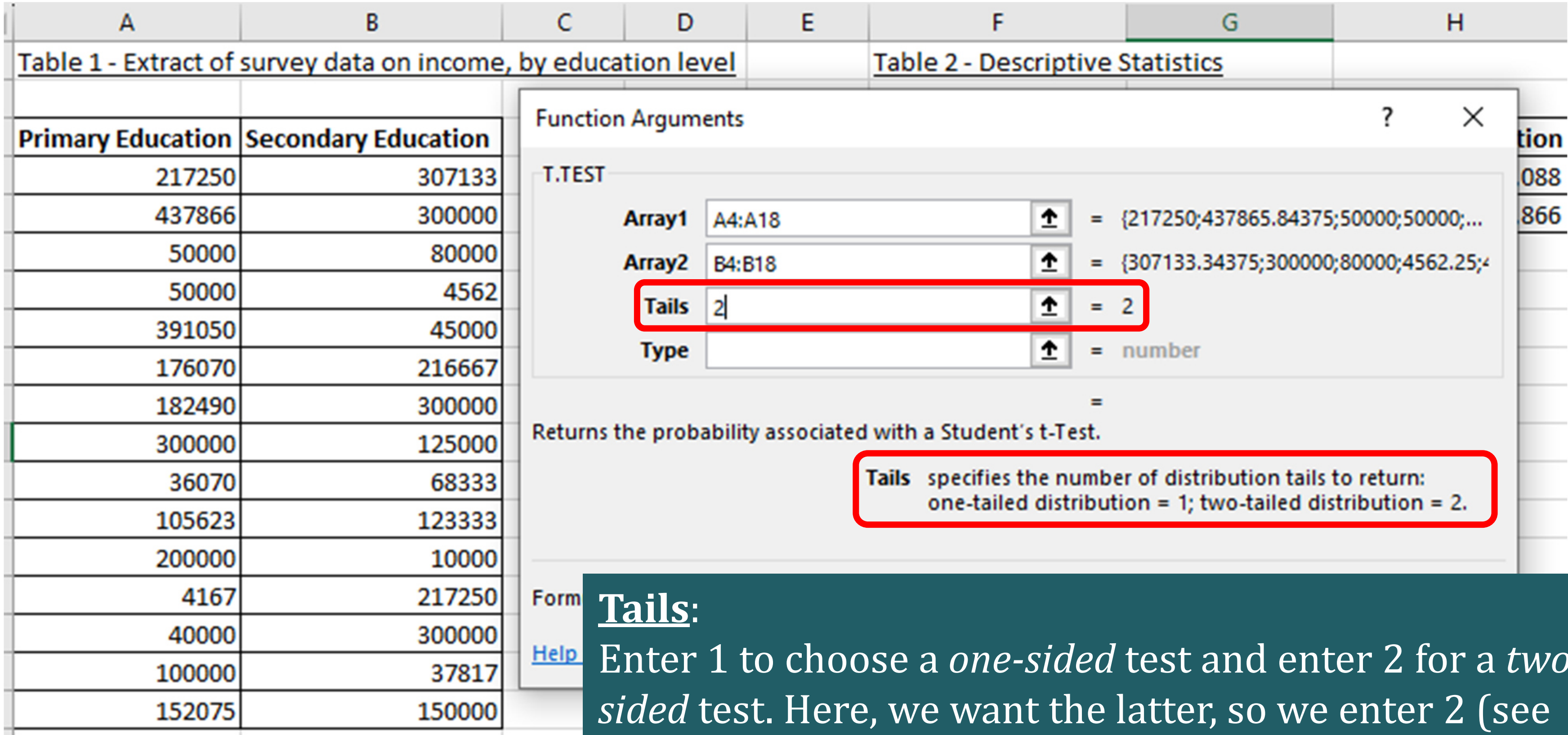


Array1 and Array 2:
Follow the steps seen previously to select the groups of values for the comparison. Note that you may put either group in Array1 or Array2, it will not change the result.

Center for Evaluation and Development

Example 1 – Statistical Testing

t-test



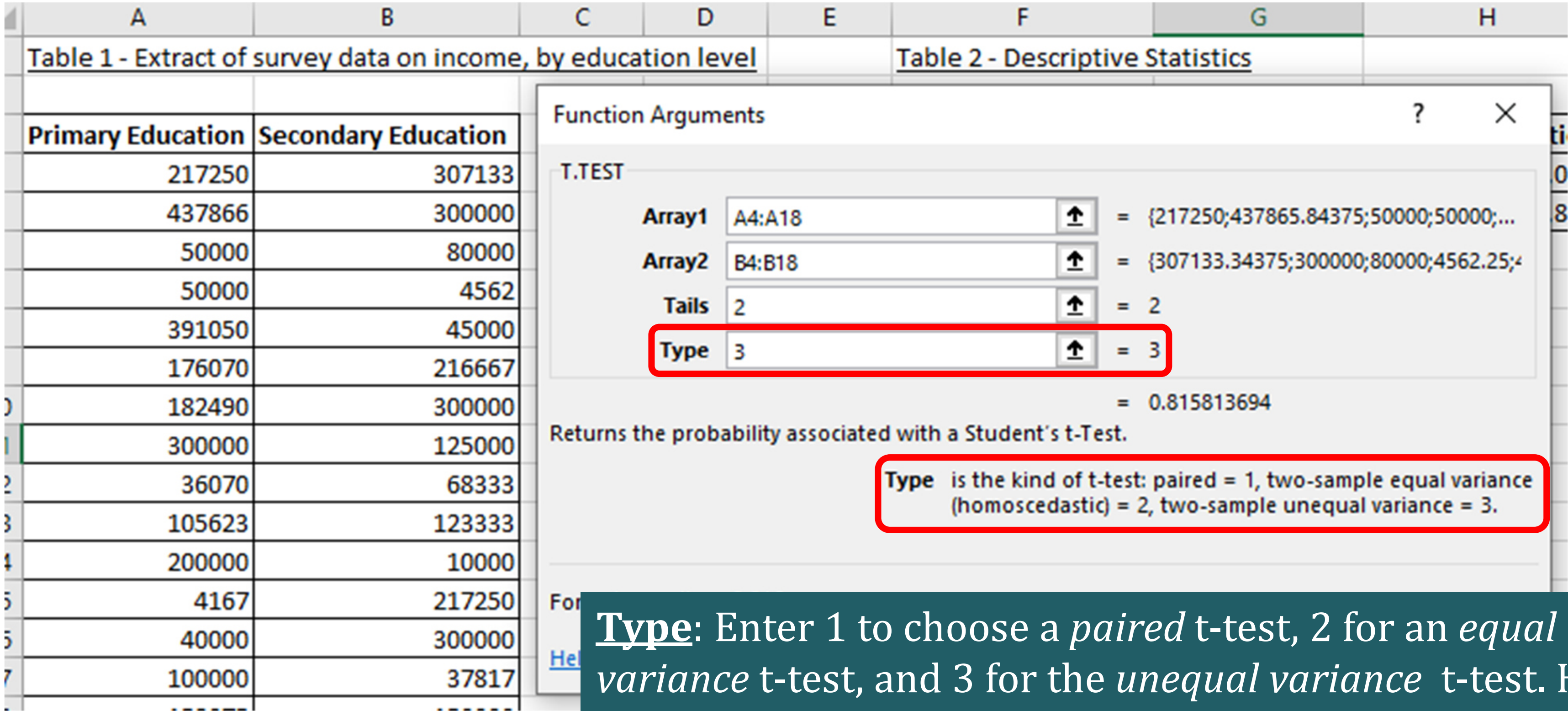
Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Tails:
Enter 1 to choose a *one-sided* test and enter 2 for a *two-sided* test. Here, we want the latter, so we enter 2 (see appendix slides on one-sided vs. two-sided t-tests).

Center for Evaluation and Development

Example 1 – Statistical Testing

t-test



Type: Enter 1 to choose a *paired* t-test, 2 for an *equal variance* t-test, and 3 for the *unequal variance* t-test. Here, we want the latter, so we enter 3 (see appendix slides on the different types of t-test). Then click on "OK".

Center for Evaluation and Development

Example 1 – Statistical Testing

t-test

=T.TEST(A4:A18,B4:B18,2,3)

Table 1 - Extract of survey data on income, by education level

Primary Education	Secondary Education
217250	307133
437866	300000
50000	80000
50000	4562
391050	45000
176070	216667
182490	300000
300000	125000
36070	68333
105623	123333
200000	10000
4167	217250
40000	300000
100000	37817
152075	150000

Table 2 - Descriptive Statistics

	Average Income	Standard Deviation
Primary Education	162,844	131,088
Secondary Education	152,340	112,866

Table 3 - Comparison of Means

Difference in means	- 10,504.29
p-value (t-test)	0.816

The function T.TEST gives the p-value of the test.
Here, the p-value is 0.816, so we cannot reject the null hypothesis that the difference between group means is 0.

Center for Evaluation and Development

European Commission

END OF SESSION 3b

Center for Evaluation and Development

3b – Example t-test in Excel

Appendix

T-test – One-sided and Two-sided test

- Consider a **t-test** to compare means between two groups, and let's call the group means X1 and X2.
- We saw that the typical null hypothesis for such a t-test is: "X1 – X2 = 0".
  - If we reject the null, we say the difference is *statistically significant*.
  - Such a test ignores the *direction* of the difference – i.e., whether it is positive or negative.
  - Hence it is called a **two-sided test** – i.e., it doesn't matter "on which side of 0" the difference in means falls.
- The t-test also allows a null hypothesis that specifies a direction, e.g.:

"X1 – X2 > 0" or "X1 – X2 < 0"

  - This is a **one-sided test** – i.e., it does matter "on which side of 0" the difference in means falls.

T-test – One-sided and Two-sided test

- The choice of one-sided vs. two-sided test influences the critical values used for calculating the p-value, and hence the decision.
- That is, the p-value for a one-sided test cannot be used to decide on the two-sided test, and vice versa.
- In practice, most CIE studies in social sciences/economics use two-sided tests.

T-test – Paired, Equal, Unequal

- **Paired t-test**: should be used when the data are *paired*, i.e., measurements come from the same unit.
  - E.g., Comparing the income of the same individuals *before* and *after* an event.
- **t-test with equal variance**: should be used when you assume the variance (or standard deviation) is equal/similar in both groups.
- **t-test with unequal variance**: should be used when you assume the variance (or standard deviation) is different in both groups.

- Equal vs. Unequal variance t-test → which one to use?

T-test – Paired, Equal, Unequal

- **Equal vs. Unequal**:
  - In practice, one can use a test to gauge whether the variances in both groups are equal (with an F-test).
  - However, many researchers warn about decisional chains that rely on a sequence of tests
  - Statistical tests are not infallible, so if the first test says, e.g., that variances are equal when they are not in reality, then the decision to use a t-test with equal variance is also wrong, and hence its conclusion is irrelevant
  - Practical advice:
    - ➢ If in doubt, use the **t-test for unequal variance** – it is more "robust" than the alternative because, in case variances are actually equal, the *unequal variance t-test* is conservative, i.e., it tends to be harder to reject the null hypothesis
    - ➢ In most real-world datasets, both groups in the sample are large enough so that the difference between the two types of t-test is marginal