Data Science & AI (TI) (CURPBATIN1810002024)
Data Science & AI (EN)

Inhoudsopgave

Data Science & AI (EN) (Leerpad)

This learning path contains an overview of the **Data Science & AI** course. This course will be taught for the first time in the academic year 2021-2022. The content will continue to be supplemented as needed over the course of the semester.

On the left, you will find a **table of contents** with the study guide and the 7 modules that make up the course. For each module, you will find the **learning objectives** (what do you need to know and be able to do?), references to **lesson materials** (slides, recordings, etc.), and **exercises**.

In this learning path, you will find a number of **reflection exercises** for each module. These are rather aimed at studying the theory and gaining a deeper insight into it, no extensive calculations are required. In addition, for each module, there are **lab assignments** where you will analyze data sets with Python. Those lab assignments are not included in this learning path, but made available through a public **Github repository**.

Feedback course content, learning path

If you suspect an **error** in the content of the learning path, the slides, or other learning material, or if something is **unclear**, do not hesitate to report this to the lecturer supervising your class group, or to the lecturer of the course, Bert Van Vreckem (mailto:Bert.VanVreckem@hogent.be?subject=Feedback%20course%20Data%20Science%20%26%20AI)

0. Study Guide (Sectie)

0.1 Purpose and place of the course in the curriculum (Pagina)

Purpose and place of the course in the curriculum

This course is an introduction to what is often called data science these days. The aim is to get you started in the correct collection, processing, and analysis of numerical data and to write a well-founded research report about it. You will learn to apply the right techniques to summarise and visualise numerical data, and you will learn how to check the relationship between different variables.

Successful companies make decisions not on the basis of gut feeling or intuition, but by collecting and analysing data. Using the techniques explained in this course, you will have sufficient background to answer questions such as:

- Is a (web) application fast enough for its users? Is the user experience consistent, or do response times vary widely?
- If you have to compare two systems, be it software or hardware, which one is the better performer? Is the difference between the two significant, or could differences in measurements be due to chance or other factors?
- When should purchases of new equipment (e.g. hard drives, servers, memory, etc.) be scheduled, based on historical usage data?

The competencies you acquire in this course are also useful outside of IT. After all, you will learn how to deal critically with data and information, and how to analyse and interpret them correctly. In the political and social debate, deliberate statements are made that are demonstrably false or that attempt to "bend the truth". The term that often comes up in this context is 'Fake News'. Charts that at first sight show objective data are manipulated in such a way that they nevertheless give a distorted (or completely wrong) picture of reality. One way of guarding against this is to take a critical look at the information that is distributed. In this way, the underlying reason for the disinformation can often be made clear.

Statistics and data science are therefore indispensable for (i) analysing data correctly and arriving at well-founded conclusions, and (ii) conducting your own research in order to be able to send well-founded conclusions into the world.

0.2 Learning objectives and competences (Pagina)

Learning objectives and competences

The aim of this course is to familiarize students with statistical methods for analysing and visualising data.

- Descriptive statistics
  - Knows some descriptive measures for data.
  - Can calculate some descriptive measures for data using statistical software (Python).
  - Knows different types of plots to represent data visually.
- Probability theory
  - Knows the basic rules for calculating with probabilities.
  - Knows the properties of some important probability distributions.
- Analysis of 2 variables
  - Can quantify and appropriately test the relationship between two variables.
  - Can construct a simple linear model to show the relationship between two or more variables.
- Time-series analysis
  - Can discuss some common models to predict time series and/or detect anomalies.
  - Can explain the importance of testing the accuracy of a model in a methodologically correct way.

0.3 Course Content (Pagina)

Course Content

In **module 1**, we explain some *basic concepts*, such as the scientific method, types of research, variables, *levels of measurement*, etc. We also discuss how *samples* are taken from a population and what mistakes can be made.

In **module 2**, we start with an introduction to descriptive statistics, i.e. the *analysis of a single variable*. This includes calculating centrality and dispersion measures (depending on the measurement level of a variable) and correct forms to visualise qualitative and quantitative data.

The central topic in **module 3** is the *central limit theorem*, an essential concept in statistics because it determines the conditions under which you can extrapolate measurement results from a sample to the population as a whole. The mathematical proof of the central limit theorem is not shown here, but to understand its meaning and applications you need some knowledge about *probability* and *probability distributions*. One of the most important applications of the central limit theorem is the procedure of a *statistical test*. In this module, we will explain the procedure of such a test and apply it to the so-called z-test and the t-test.

In the following modules, we will try to look for relationships between two variables, i.e. whether the value of one variable (the dependent) changes in a systematic way with respect to the value of another (the independent). This is a first (but certainly not sufficient!) indication that a causal relationship may exist between the two variables.

In Module 4, we focus on the situation where both variables are qualitative. In that case, you can use the *chi-square test* (χ^2 test) and *Cramér's V*. We also see ways to *visualise* this kind of data.

If the independent variable is qualitative, and the dependent is quantitative, you can use *the two-sample t-test*, as well as the concept of effect size. These techniques are explained in **module 5,** along with suitable visualisation techniques for this situation.

Module 6 deals with regression analysis, which is typically used to investigate a relationship between two quantitative variables. Here we also look at suitable visualisation techniques.

The **last module** is a little bit separate from the rest of the course in terms of its subject, but this doesn't mean it is less important. Here, we discuss the principles of *time series analysis*, which can enable us to detect trends and, to a certain extent, make predictions.

0.4 Learning Material (Pagina)

Learning Material

Each module of this learning path refers to the corresponding learning material. This consists of:

- Slides to the lessons
- Book (see below)
- Code examples in Python
- Exercises

Book

We use the following book

Rajagopalan, G. (2021) *A Python Data Analyst's Toolkit: Learn Python and Python-based Libraries with Applications in Data Analysis and Statistics*. Springer. https://link.springer.com/book/10.1007/978-1-4842-6399-0 (https://link.springer.com/book/10.1007/978-1-4842-6399-0)

You can download this book for free from the HOGENT campus (or via VPN) as an e-book (pdf or epub) via the link above.

Please note that we do not follow the book from beginning to end. We have given our own structure to the subject content. In each module, we will refer to the relevant sections in the book.

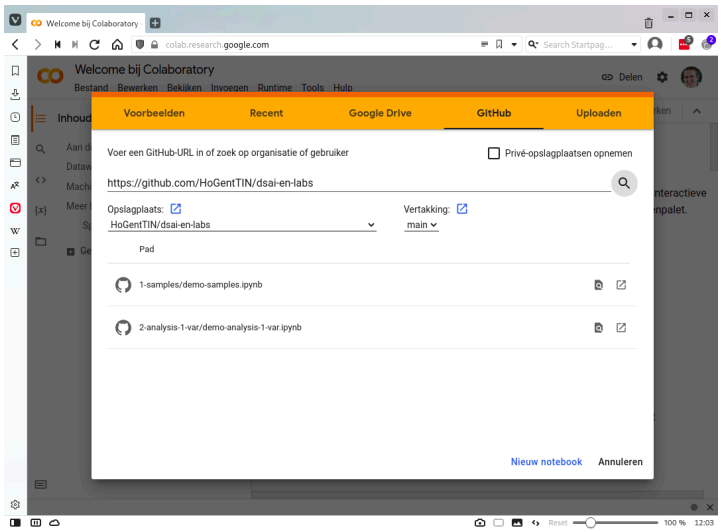
Software

The lab exercises are provided in the form of Jupyter Notebooks. This is a web-based technology whereby formatted text (Markdown), Python code and output, data, images, etc. can be combined in one file.

In the cloud

In principle, you do not need to install any software to get started! You can work online with Jupyter Notebooks via the Google Colab (http://colab.research.google.com) platform. A Google account is required for this.

1. Make sure you are logged into your Google account and visit to https://colab.research.google.com. (https://colab.research.google.com.)
2. You will see a pop-up to open a notebook. Click on the "Github" tab:



3. In the "Enter a Github URL" input box, you can paste this URL: https://github.com/HoGentTIN/dsai-en-labs (https://github.com/HoGentTIN/dsai-en-labs)
4. You will see an overview of all Jupyter Notebooks within this repository. Click on the one you want to open it.
5. Finally, make a personal copy (e.g. in Google Drive) where you can save your own changes. This can be done via the "File" menu at the top of the page.

You can also download the entire Github repository as a ZIP file and then upload the contents to your Google Drive (under the Colab Notebooks folder). That way you have everything. Please note that there will probably be changes and additions to the Github repository of lab assignments during the course of the semester. Therefore, keep an eye on the **commit history** and copy the latest version of the assignments if necessary.

On your own laptop (optional!)

If you find it useful to work on your own laptop, you may do so (but it is not required for the successful completion of this course). In this case, you will need the following software:

- Python
- Visual Studio Code, with extensions:
  - Python, Pylance (Microsoft)
  - Jupyter, Jupyter Keymap, Jupyter Notebook Rendering (Microsoft)
  - Optionally: GitLens (GitKraken), Markdown All in One (Yu Zhang)
- Git & a Github-account

We have prepared a Github repository (https://github.com/HoGentTIN/dsai-en-labs) for the Python lab commands that contains the assignments and associated datasets, as well as sample code. Please make a local clone of this repository (or download as ZIP) and open the directory in Visual Studio Code. By the way, VS Code itself will offer to install the most important extensions as soon as you open a Python script or Jupyter Notebook.

Background Information

The content of this course is largely based on the course "Research Techniques" in the previous curriculum. It used lecture notes that were written by the lecturers. Although we no longer use this course, it may still be useful. The course is publicly available via Github, but you need LaTeX to compile it into PDF. Therefore, we will make a PDF available through Chamilo, under Documents.

0.5 Lecture types (Pagina)

Students enrolled in the daytime curriculum receive **three hours of lectures** per week. These lessons alternate between **classroom instruction** with exercises and lab assignments.

The **exercises** for which you do not need a computer are included in this learning path. There are also **lab assignments** where you will analyse and visualise data sets using Python.

0.6 Work and study hints (Pagina)

Work and study hints

Many students find the content of the *Data Science & AI* course difficult. This is understandable because the subject is outside the comfort zone of the average computer science student and we all know that mathematical subjects are not the most popular ones in our study programme.

There are two ways of dealing with this. You can take the path of least resistance: concentrate on the subjects you like and go through this course the day before the exam in the hope of getting enough points to get a credit (i.e. 10/20). Experience shows that this strategy is not successful. Experience shows that this strategy is not successful. In June, typically only a third of the students passed. In September, we often see a much higher pass rate, which in our opinion suggests that if you put enough effort into this subject, it is certainly achievable.

Some tips to succeed in this subject straight away:

- Attend the lectures and actively take notes (Lundin, 2020);
- Also work for this subject outside of the lectures. Revise the theory you have seen and complete the exercises you have not yet finished. Make a note of things you do not understand or of things that you get stuck on, and ask your question at the next lecture.
- Use *good learning techniques*. You will find a good overview of learning techniques whose effect has been scientifically proven on the website of The Learning Scientists (http://learningscientists.org):
    - *Spaced practice*: Study in several small sessions (at least once a week) and not in large blocks. Book a fixed moment in your weekly agenda/lesson planning!
    - *Retrieval practice*: Take a blank piece of paper and try to write down as many things as possible about a certain topic from memory (i.e. without looking in the lecture notes). Then check this against your written notes and in the lectures notes.
    - *Elaboration*: Ask yourself questions about how things (e.g. formulas, test procedures, . . .) fit together and why that is the case. Consult with fellow students. Ask your lecturer for more explanation if necessary. Make connections between different topics in the course
      course (e.g. compare testing procedures).
    - *Interleaving*: Alternate subjects while studying.
    - Use *concrete examples* to understand abstract ideas. Some examples are already given in the course, try to think of others yourself. Consult with fellow students and ask your lecturer for feedback, if necessary.
    - *Dual coding*: Combine word and image, try to visually represent the material you are studying.

> **!** Common Pitfall
>
> In the end, it comes down to investing enough time and effort to study for this subject. It is normal for learning to be difficult and laborious. If everything goes automatically, you have not learned a thing. You can only grow outside your comfort zone!

0.7 Study guidance and planning (Pagina)

Study guidance and planning

Planning

The table below gives an overview of the lesson planning for daytime education. Please note that this can differ from reality, e.g. when a contact moment has to be omitted due to a public holiday or other circumstances.

Week	Subject
1	Intro, study guide, sampling
2	
3	Analysis of 1 variable
4	
5	Probability and the central limit theorem
6	Hypothesis testing: the *z*-test
7	Hypothesis testing: the *t*-test
8	Analysis of two qualitative variables: chi-square test and Cramér's V
9	Analysis of qualitative vs quantitative variables: *t*-test for 2 samples, effect size
10	Analysis of 2 quantitative variables: regression analysis
11	Time series
12	Revision and backup session

Study Guidance

- The **first point of contact** in case of difficulties or questions about this course is the lecturer with whom you follow the lessons or if need be, the lecturer for distance learning (TIAO). Preferably ask your questions during the **contact moments**, either the weekly lessons for daytime education or the contact evenings for distance learning.
- If you don't get the chance to ask your questions during a contact moment, you can ask them via **MS Teams**. A general Team has been created to which every student following this course is subscribed. You can find the link from the main page of the Chamilo course, the icon/link on the right side.
- **Please do not ask questions via email or Teams chat** unless it is for personal matters. We would like to bundle all questions about the subject matter via Teams, so that the answer is immediately available to everyone and we do not have to answer the same questions over and over again.
- Spotted an error in the slides or other course material? Please let the subject leader, Bert Van Vreckem (mailto:bert.vanvreckem@hogent.be?subject=%5BDSAI%5D%20Feedback%20learning%20material), know. Errors or suggestions for the lab assignments that are offered via Github can be reported via a Github issue (https://github.com/HoGentTIN/dsai-en-labs/issues). Thank you for taking the time to report them!

Contact Details of Lecturers

These are the lecturers for the subject, the groups they supervise, and when the scheduled contact moment takes place.

We repeat that you should only contact the lecturers by e-mail if it concerns a personal matter. Questions about the subject can be asked via Teams (see above).

- Sabine De Vreese (mailto:sabine.devreese@hogent.be?subject=%5BDSAI%5D&body=Beste%20mevrouw%20De%20Vreese%0A%0A):
  - Aalst: A2.2A-2B-2C-2D-2E (Monday afternoon)
  - Gent: G2.A2-B2-C2 (Wednesday morning)
  - Virtual campus: G2.VC (Wednesday afternoon)
- Stijn Lievens (mailto:stijn.lievens@hogent.be?subject=%5BDSAI%5D%20&body=Beste%20meneer%20Lievens%0A%0A):
  - English taught, IC (Wednesday morning)
- Lieven Smits (mailto:lieven.smits@hogent.be?subject=%5BDSAI%5D&body=Beste%20meneer%20Smits%0A%0A):
  - Distance learning/TIAO
  - Gent G2.A3-B3-D1 (Wednesday morning)
- Bert Van Vreckem (mailto:bert.vanvreckem@hogent.be?subject=%5BDSAI%5D%20&body=Beste%20meneer%20Van%20Vreckem%0A%0A):
  - course lead
  - G2.C3-C4-D2-E2-ATN (Tuesday morning);
  - G2.A1-B1-C1 (Tuesday afternoon);
  - G2.B4-E1-ATP (Thursday morning)

0.8 Evaluation (Pagina)

Evaluation

- First examination opportunity
  - 100% Written exam with use of PC (personal laptop) and aids (see below)
- Second examination opportunity
  - Same as above

Sample exam

You can find a sample exam in the Github repo for the lab assignments, under the RepeatExercises (https://github.com/HoGentTIN/dsai-labs/tree/main/RepeatExercises) directory.

Aids

On the exam, the following aids may be used (electronically or printed/on paper):

- Lecture slides and other teaching materials published on Chamilo
- The necessary software for statistical analysis: Google Colab or Python and VS Code
- The Python sample code (demo-*.ipynb) in the Github repository for the lab assignments
- Your own notes and solutions from labs and exercises
- Code snippets (that you prepared yourself) in your IDE
- Internet use is also allowed, except what is listed below

Make sure your laptop is configured correctly so you can start the exam right away. Technical problems with the software to be used are your responsibility. For technical problems, follow the instructions in the document 'First aid for digital panic'. "First aid for digital panic (https://www.hogent.be/sites/hogent/assets/File/Panic%20Card_studenten.pdf)".

Any form of communication with fellow students or third parties is of course forbidden. That includes, but is not limited to:

- Cell phone, smartphone or smartwatch
- Headphones, wired or bluetooth earphones (earplugs or earmuffs without electronics are allowed)
- Communication apps like Discord, Messenger, WhatsApp, Teams, email client, chat applications, etc. Remember to turn all desktop notifications OFF, as incoming notifications are also considered communication.
- Websites that allow communication, e.g. Facebook, Q&A sites like StackOverflow, Forums, etc. (if during the exam you do accidentally land on such a website via a search, immediately close the tab/window).
- **Large Language Models** such as ChatGPT, LLama, Copilot Chat, Bing Chat, Google Bard/Gemini, etc. The use of **Github Copilot is also prohibited.**
- Other applications, websites, etc. by which you outsource answering the exam questions to an external party (e.g. Mechanical Turk).
- *Any other form of communication or method by which you do not answer exam questions yourself that is not listed above.*

Guidelines during the exam

- The total score you obtain for this examination will be recalculated to an examination mark out of 20.
- All students (VC and on-campus) will make a **video recording of the exam** which will be shared immediately after the exam via **Panopto** with the lecturer of the class group in which you are enrolled.
  - For the on-campus exam, a recording of your screen and the built-in webcam of your laptop is sufficient. VC students must also film their workspace via an external webcam.
  - All guidelines for Panopto recordings of exams apply. Read these carefully and strictly follow the instructions. Prepare in advance by creating a folder with the right name and sharing it with the right lecturer.
- Download the assignment from exam.hogent.be as a Jupyter Notebook (.ipynb).
- Immediately rename the Jupyter Notebook file dsai-2324-ReeksX-FAMILYNAME-GIVENNAME.ipynb **before** opening it, replacing FAMILYNAME and GIVENNAME with your name.
- In the first Markdown cell, enter your name, student number, exam date and time and class group. If you are an IOEM student that is entitled to extra time, indicate that too.
- Add code cells where necessary for working out the questions.
  - **TIP!** Avoid reusing generic variable names (e.g. data, df, ...) between different questions, but instead use descriptive names where possible (e.g. penguins, agri_businesses, aus_athletes, etc.). After all, if you switch between working on different questions during the exam, you might overwrite variables and get wrong results or runtime errors.
- **ATTENTION:** Only what you write down in the Markdown cells provided for this purpose counts as an answer! The content of the code blocks only serves to substantiate your answer, so we can see how you arrived at your answer if it did not match the expected outcome.
- If the answer to a question is a real number, always **round it to exactly four decimal places**, unless explicitly asked otherwise. Very small numbers can be written in scientific notation (e.g. 1.2345e-6 or 1.2345x10^-6)
- **Submission:**
  1. You submit the **Notebook file** (.ipynb, with your name in the file name!) via exam.hogent.be. Do not pack in a .zip/.rar, or submit in any other file format!
  2. Immediately after the exam, make sure your **Panopto recording** is loaded, has the correct name and has been shared with the correct lecturer.
  3. Students taking the exam on-campus must **sign for attendance.**

> **!** Pay Attention!
>
> If your submission is incomplete (Notebook, Panopto recording and/or signature on attendance list is/are missing), you cannot prove that you took the exam in a correct and fair manner. In that case, it is also not possible for us to grade you and you will receive 0/20 as your exam grade.

1. Fundamental concepts, sampling (Sectie)

1.1 Learning Goals (Pagina)

Learning Goals

By the end of this chapter you must be able to:

- Define the following basic concepts:
  - scientific method and empirical validation
  - variable, value
  - measurement levels: nominal, ordinal, interval, ratio
  - population, sample, sampling frame
  - random vs.~ select sample
  - representativeness of a sample
- Identify and explain scientific research objectives;
- Identify and explain the different steps in a research process;
- List the four measurement levels, formulate the characteristics of each and give an example;
- Determine the measurement level for a given variable;
- List the types of sampling errors, formulate the characteristics of each and give an example;
- Based on the description of a sampling method:
  - make a distinction between a random and a selective sampling method;
  - identify and explain the type of sampling error(s) made;
  - make proposals to improve the sampling method;

1.2 Learning Materials (Pagina)

Learning Materials

Textbook

Rajagopalan (2021, chapters. 1-6)

Lecture Slides

- Course introduction, study guide (https://chamilo-downloads.hogent.be/Chamilo/Libraries/Resources/Javascript/Plugin/PDFJS/web/viewer.html?file=https%3A%2F%2Fchamilo-downloads.hogent.be%3Fapplication%3DChamilo%255CCore%255CRepository%26go%3DDocumentDownloader%26object%3D6186625%26DownloadHost%3D1%26security_en-0-intro.pdf)
- Basic concepts, sampling (https://chamilo-downloads.hogent.be/Chamilo/Libraries/Resources/Javascript/Plugin/PDFJS/web/viewer.html?file=https%3A%2F%2Fchamilo-downloads.hogent.be%3Fapplication%3DChamilo%255CCore%255CRepository%26go%3DDocumentDownloader%26object%3D6186629%26DownloadHost%3D1%26security_en-1-sampling.pdf)

Lecture Recordings (in Dutch)

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

1.3 Exercises (Pagina)

Reflection Exercises

Basic Concepts in Research

Exercise 1 - Retrieval Practice: measurement levels

Measurement levels are an important concept in descriptive statistics because most visualization and analysis techniques depend on this. It is therefore important to know and understand the different measurement levels.
Retrieval practice is a study technique that has been scientifically proven to be effective and which leads to improved learning outcomes (Roediger & Karpicke, 2006).

1. Take a blank sheet of paper and try to reproduce an overview of all measurement levels without consulting the syllabus or other sources. Describe the specific properties for each measurement level and give some examples.
   Take enough time for this (e.g. at least 5 to 10 minutes). Do not immediately look into the syllabus, but try to remember as much as possible.
   When you are done with this, indicate everything that you have noted so far with an (e.g. green) marker.
2. If possible, consult with a fellow student and try to complete the overview that you have each made. Mark all additions with a marker in a different color (e.g. yellow).
3. Finally, verify your notes using the syllabus and correct/complete incorrect or missing information if necessary. Indicate this in a third color (e.g. orange or red).

Thanks to the highlighted colors in your notes, you now have an overview of what you already know and what you need to study. Repeat this exercise a number of times during the course of the semester. Never start by looking at the result of a previous attempt, but immediately try to retrieve as much information as possible from your memory on a blank sheet. Compare the results afterwards. You should notice that over time, you will mark more and more in green and you will use less red or even none at all.

Sampling Methods and Sampling Errors

Exercise 2 - sampling methods

A researcher wants to investigate as accurately as possible the consumption habits of residents aged 18 years and up in a certain municipality with 3 residential areas. She distinguishes 4 age groups so she eventually has a total of 12 subgroups. From the municipality, she requests the percentual composition of the population in the city, and based on the numbers she calculated the number of respondents required for each age group. We call this type of sample a *quota sample*.

- Is this a random sample? Why (not)?
- Is this sample representative of the population?
- Which types of sampling errors can occur?
- What are the advantages and disadvantages?

Exercise 3 - sampling methods

A research agency wants to investigate the purchasing behavior of washing products. They interrogate a number of women between 25 and 55 because they assume that these are representative of most customers.

- Is this a random sample? Why (not)?
- Is this sample representative of the population?
- Which sampling error(s) was (were) made?
- What improvements would you suggest for the way this sample is constructed?

Exercise 4 - sampling methods

The trade union wants to investigate the working conditions of the employees of an IT company. This company has a total of 3200 employees spread over 12 branch offices. Because the total number of employees is quite large, 40 employees are selected at random from each branch office. The size of the sample is, therefore, $n = 480$.

- Is this a random sample? Why (not)?
- Is the sample representative of the population? If not, when could it be representative?
- What kind of error(s) is/are being made here?
- What improvements would you suggest for the way this sample is constructed?

Exercise 5 - sampling methods

We want to conduct a survey into the quality of our education at Hogeschool Gent. For this purpose, the students present in a certain lecture will be questioned.

- Is this a random sample? Why (not)?
- Is the sample representative of the population?
- What kind of error(s) is/are being made here?
- Suppose the lecturer who teaches the subject is present during the questioning. What influence can this have and what mistake is made in this case?
- Suppose the questioning is not held during a lecture but after an exam. What mistake can be made then?
- What improvements would you suggest for the way this sample is constructed?

Exercise 6 - Retrieval practice: sampling errors

Use the procedure for retrieval practice from Exercise 1 to study the types of sampling errors. List the different sampling errors, describe each and give an example.

2. Analysis of 1 variable (Sectie)

2.1 Learning Goals (Pagina)

Learning Goals

By the end of this chapter you must be able to:

- Specify the appropriate measures for central tendency and dispersion for each measurement level;
- Reproduce and understand the formulas for calculating the mean, variance, and standard deviation of a sample;
- Calculate the central tendency and dispersion of a given variable;
- Identify suitable visualization techniques for each measurement level;
- Apply suitable visualization techniques for a given variable;
- Interpret a given graph, this includes naming the graph type and deriving the central tendency and dispersion.

Overview central tendency and dispersion measures

Measurement Level	Central Tendency Measures	Dispersion Measures
Qualitative	Modus	/
Quantitative	Mean Median	Variance, standard deviation Range, interquartile range

Suitable measures for central tendency and dispersion for each measurement level. The combination of central tendency and dispersion is often used when summarizing a series of measurements.

Overview visualisation techniques

Measurement Level	Chart type
Qualitative	Bar chart of frequencies
Quantitative	Boxplot Histogram Density plot

Suitable chart types per measurement level for visualizing a single variable.

2.2 Learning Materials (Pagina)

Learning Materials

Text book

- Visualisation techniques: chapter 7
- Central tendency measures: p 347 "Measures of central tendency"
- Dispersion measures: p 348 "Measures of dispersion"
- Shape: p 349: "Measures of dispersion"

Lecture Slides

- Analysis of 1 variable (https://chamilo.hogent.be/index.php?go=CourseViewer&application=Chamilo%5CApplication%5CWeblcms&course=47890&tool=Document&publication_category=282180&browser=Table&tool_action=Viewer&public

Lecture Recording (in Dutch)

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

2.3 Exercises (Pagina)

Reflection Exercises

Exercise 1 - retrieval practice: analysis of 1 variable

Use the procedure for retrieval practice from the similar exercises in Module 1
to study the *techniques for analysis and visualization of a single variable*.

For each measurement level, provide:

- The appropriate measures for central tendency and dispersion (name, definitions and formulas)
- The appropriate graph types

Central tendency and dispersion measures

Exercise 2 - sample mean and sample variance of a frequency table

Consider the formulas for the sample mean \(\overline{x}\), the sample variance \(s^2\) and the standard deviation \(s\). How should these formulas be adapted to calculate these values when we are dealing with a frequency table? A frequency table gives an overview of how often each different value (of a qualitative variable) occurs in the sample.

Apply your formula to the data in the table below:

Pins *x*	**Frequencies \(f_x\)**
0	2
1	1
2	2
3	0
4	2
5	4
6	9
7	11
8	13
9	8
10	8

While playing a skittles game, the number of pins that were knocked over with each throw is recorded. For each possible score x, the number of times this score was obtained during a throw was recorded.

Results (for your convenience): n = 60, mean = 7, variance ≈ 5.83, standard deviation ≈ 2.41

Exercise 3 - formula for sample variance

In the formula for the sample variance, the difference between the measurement values and the mean is squared. Why? Couldn't we devise a simpler formula that is an equally good measure of the dispersion of a dataset? Here are three proposals (the third one is the "real" formula):

\(\begin{align} s_{1}^{2} &= \frac{1}{n-1} \sum_{i=1}^{n} (\overline{x} - x_i) \\ s_{2}^{2} &= \frac{1}{n-1} \sum_{i=1}^{n} \left| \overline{x} - x_i\right| \\ s_{3}^{2} &= \frac{1}{n-1} \sum_{i=1}^{n} (\overline{x} - x_i)^{2} \end{align}\)

Apply each formula to the two data sets below. By comparing the results, you should be able to decide whether the formulas are suitable as a dispersion measure.

\(\begin{align*} X &= \left\{ 4,4,-4,-4 \right\} \\ Y &= \left\{ 7,1,-6,-2 \right\} \end{align*}\)

Exercise 4 - coefficient of variation

On your own, look up what the coefficient of variation for a sample is. How is it defined for a full population and what could you do with it?

Visualisation Techniques

Exercise 5 - data visualisation hall of shame

Look for examples of bad graphs in news reports, articles, interest group publications, etc.

Why is the chosen graph "bad"? What mistakes are being made? What changes should be made to correct the graph? Who will find the most ridiculous example of a bad graph within the class group?

3. Probability theory, central limit theorem, statistical tests (Sectie)

3.1 Learning Goals (Pagina)

Learning Goals

After studying this chapter, you will be able to

- Probability:
  - explain the basic concepts of probability (probability space, event, outcome)
  - sketch the (standard) normal distribution and explain its properties
  - calculate a z-score
  - calculate probabilities in a (standard) normal distribution (left tail probability, right tail probability)
- The central limit theorem
  - formulate the central limit and explain its importance for statistical analysis
  - calculate confidence intervals for the mean of a large or small sample
- Statistical tests:
  - explain the basic concepts of statistical testing (test, null hypothesis, alternative hypothesis, sample size, level of significance, p-value, acceptance and critical range, ...)
  - list and explain the procedure of a statistical test
  - perform a z-test and t-test on a dataset.
  - be able to choose the correct variant of the z- or t-test for a given situation
  - explain the possible errors (type I, type II) in hypothesis testing

3.2 Learning Materials (Pagina)

Learning Materials

Text Book

- Probability: chapter 9, more in particular
  - p. 327 "Probability"
  - p. 335 "Probability distributions", (standard) normal distribution
- The central limit theorem: chapter 9, more in particular
  - p. 355 "Central limit theorem"
- Statistical tests: chapter 9, more in particular
  - p. 358 "Hypothesis testing"
  - p. 362 "One sample z-test"
  - p. 370 "T-distribution"

Slides

- Central limit theorem (https://chamilo.hogent.be/index.php?go=CourseViewer&application=Chamilo%5CApplication%5CWeblcms&course=47891&tool=Document&publication_category=283173&browser=Table&tool_action=Viewer&public
- Hypothesis testing (https://chamilo.hogent.be/index.php?go=CourseViewer&application=Chamilo%5CApplication%5CWeblcms&course=47891&tool=Document&publication_category=283173&browser=Table&tool_action=Viewer&public

Addional links:

- Really good visual explanation about the central limit theorem: here (https://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm)

Lecutre Recordings (in Dutch)

Central Limit Theorem Part 1 + 2

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Demo Central Limit Theorem

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Labo 3.01

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Introductie + Demo Hypothesis Testing

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Lab 3.02 Hypothesis Testing Exercise 1

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Labo 3.02 Hypothesis Testing Exercise 2 + 3

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

3.3 Exercises (Pagina)

Exercises on Probability

Exercise 1

Consider the two games of chance mentioned in the slides:

- Bet on: at least one six when throwing a fair die 4 consecutive times.
- Bet on: at least one "double six" when throwing two fair dice 24 times.

Question: calculate the exact probability of winning each of these two games. Compare your answer with the long-term frequency of winning shown in the slides.

Exercise 2 (Reliability of backups)

A hard drive has a 1% probability to crash. Therefore, we take two backups, each having a 2% probability to crash. The three components work independently of each other. Data is lost only when all three components crash.

Question: what is the probability that the data is not lost?

Exercise 3 (Expectation of $X - a$)

Consider the probability distribution of the random variable X giving the sum of the eyes when throwing two dice.

- This probability distribution is symmetric around 7. Verify (by using the definition of expectation) that the expectation of X is also 7.
- Suppose that you have to pay 8 EUR to play a game where your earnings are the sum of the eyes when throwing two dice. Denote your profit (or loss) by $Y.$ What is the relationship between Y and X? Give the expectation of Y. What is the relationship with the expectation of X?
- Can you see a general relationship between the expectation of $X - a$and that of X, where a is a real number.

Exercise 4 (Variance of X/a)

Consider the probability distribution of the random variable X giving the sum of the eyes when throwing two dice.

- Calculate the variance of X.
- Suppose the profit you make is only half of the sum of the eyes when throwing two dice. Denote your profit (or loss) by $Y.$ What is the relationship between Y and X? Give the variance of Y. What is the relationship with the variance of X?
- Can you see a general relationship between the variance of X/a and that of X, where a is a real non-zero number.

Exercise 5 (Expectation and variance of $(X-\mu)/\sigma$)

Using the information from the two previous exercises, determine the expectation and variance for a random variable Z given by

$Z = \frac{X - \mu_X}{\sigma_X}$.

Reflection Exercises

Central limit theorem

Exercise 1

When students write in their bachelor thesis proposal that they want to conduct a survey, we usually try to dissuade them from doing so. With what you have learned about sampling and the central limit theorem, can you think of some reasons why we

do this?

Confidence intervals

Exercise 2

1. How do you calculate the upper and lower limits of a 95% and 99% confidence interval?
2. A 99% confidence interval is [wider/narrower/the same width] as a 95% confidence interval. Why?
3. What would a 100% confidence interval look like?

4. Analysis of 2 qualitative variables (Sectie)

4.1 Learning goals (Pagina)

Learning goals

After this chapter you should be able to:

- explain the basic concepts of analysis of 2 variables: a statistical relationship between 2 variables, dependent/independent variable
- explain why a correlation does not necessarily imply a causal relationship
- correctly visualise two qualitative variables
- set up a crosstab for 2 qualitative variables
- calculate the statistics χ^2 and Cramér's V
- perform the χ^2-dependency test
- perform the χ^2 goodness-of-fit test
- calculate standardised residuals
- explain Cochran's rules and check whether they are fulfilled in a given case

Summary analysis of two variables

Independent	Dependent	Statistic/test
Qualitative	Qualitative	χ^2-test Cramér's V
Qualitative	Quantitative	two sample t-test Cohen's d (effect size)
Quantitative	Quantitative	Linear regression, correlation

Independent	Dependent	Chart type
Qualitative	Qualitative	Clustered bar chart, mosaic diagram
Qualitative	Quantitative	Box plot, bar chart using error bars
Quantitative	Quantitative	Scatter plot, regression line

4.2 Learning Materials (Pagina)

Learning Materials

Text book

- p 379. Chi-square test of association"

Lecture slides

- Slides for module 4. (https://chamilo.hogent.be/index.php?go=CourseViewer&application=Chamilo%5CApplication%5CWeblcms&course=47891&tool=Document&publication_category=283173&browser=Table&tool_action=Viewer&public

Lecture Recordings (in Dutch)

Slides + demo

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Labo

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

4.3 Exercises (Pagina)

Reflection Exercises

Exercise 1

Look for an example in the news of scientific research where the press claims that there is a causal connection between two phenomena. Try to find the original scientific publication. Is the claim made in the article correct? Do the researchers also claim that there is a causal connection, or do they express themselves more cautiously?

5. Analysis of 2 variables: qualitative vs quantitative (Sectie)

5.1 Learning Goals (Pagina)

Learning Goals

After this chapter you should be able to:

- Use appropriate techniques to visualise the relationship between a qualitative and quantitative variable
  - Grouped box plot, probability density plot, violin plot
  - Bar chart with error bars
- Apply the t-test for 2 paired or independent samples
- Calculate an effect size (Cohen's d)

5.2 Learning Materials (Pagina)

Learning Materials

Text Book (Rajagopalan, 2021)

- p. 372. Two-sample t-test
- p. 373. Two-sample t-test for paired samples

Slides

Analysis of qualitative vs. quantitative variables (https://chamilo.hogent.be/index.php?go=CourseViewer&application=Chamilo%5CApplication%5CWeblcms&course=47891&tool=Document&publication_category=283173&browser=Table&tool_action=Viewer&publication=

Lecture Recordings (in Dutch)

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

5.3 Exercises (Pagina)

Reflection Exercises

This module doesn't have any reflection exercises.

6. Analysis of 2 quantitative variables (Sectie)

6.1 Learning Goals (Pagina)

Learning Goals

After this chapter you should be able to:

- Use appropriate techniques to visualise the linear relationship between two quantitative variables:
  - Scatter diagram
  - Regression line
- Calculate the parameters of a regression line (slope and intersection with the y-axis)
- Calculate covariance, correlation and coefficient of determination and interpret their value

6.2 Learning Materials (Pagina)

Learning Materials

Textbook (Rajagopalan, 2021)

The subject of linear regression is not discussed in the text book. We therefore refer to the slides and to the Python code with accompanying explanations in the Jupyter Notebook file accompanying the lab assignments (demo-regression.ipynb).

Slides

Analysis of two quantitative variables (https://chamilo.hogent.be/index.php?go=CourseViewer&application=Chamilo%5CApplication%5CWeblcms&course=47891&tool=Document&publication_category=283173&browser=Table&tool_action=Viewer&publication=

Lecture Recordings (in Dutch)

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

6.3 Exercises (Pagina)

Reflection Exercises

This module doesn't have any reflection exercises.

7. Timeseries analysis (Sectie)

7.1 Learning Goals (Pagina)

Learning Goals

After this chapter you should be able to:

- Visualise a time series using Python
- Explain the concept of a time series model
- Calculate the simple moving average (SMA) of a time series
- Apply a suitable variant of exponential smoothing to a time series and make predictions (forecast) based on the constructed model:
  - Single Exponential Smoothing
  - Double Exponential Smoothing or Holt's method for time series with a rising or falling trend
  - Triple Exponential Smoothing or the Holt-Winters method for time series with a seasonal trend
- Assess the quality of a time series model on the basis of Mean Squared Error or Mean Absolute Error.

7.2 Learning Materials (Pagina)

Learning Materials

Textbook

The subject of time series analysis is not covered in the handbook. Please study the slides and the Python-code.

Slides

The slides can be found here. (https://chamilo.hogent.be/index.php?go=CourseViewer&application=Chamilo%5CApplication%5CWeblcms&course=47891&tool=Document&publication_category=283173&browser=Table&tool_action=Viewer&publication=

Lecture Recording (in Dutch)

Time series demo

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Time Series Lab Covid Part 1

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Time series Lab Covid Part 2

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

Time series Lab Golden Cross

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.

7.3 Exercises (Pagina)

Reflection Exercises

This module doesn't have any reflection exercises.

Example Exam (Pagina)

Lecture Recording (in Dutch)

Hier staat inhoud die niet geprint of offline gebruikt kan worden. Gelieve de webpagina te bezoeken met behulp van onderstaande QR-code om de volledige inhoud te bekijken.