Research methods in AI

Assignment 1

Assignment 1 consists of three parts (A-C). In three weeks, we go from formulating a research question, to estimating the effect of questionable research practices on type I error rate.

Part A

Description

In this first part, we make a testable research question. The question is based on one of the six case studies. These case studies are fictional, but mostly realistic, descriptions of data that was collected in a scientific study. You can pick a case study that appeals to you.

For now, there is no actual data, but there are descriptions of the variables in the dataset. Use these variables to come up with a research question, while taking into account:

- Researchers usually want to find a statistically significant difference or association between things, so think of something that makes sense to you but is also not obvious (otherwise there would be no reason to test it).
- The question should be testable, so it should be possible to answer it if we had the data.
- Next week you will make (simulate) the variables that are needed to answer the question, so don't make it too complicated for yourself.

Write a short text, like a mini-version of an article introduction, that describes the context of the study, the (fictional) gap in our knowledge, and the research question your study wants to answer. You can of course have a short look at existing studies in this field and cite those, but it's OK if things remain fictional for now (we will spend more time on realistic research questions in Assignment 2).

Make sure that all group members agree on the text you hand in.

Hand in

Your introduction text (max. 500 words), as a .pdf file.

Homework for the next meeting

As mentioned before, next week you will need to do simulations in R, so have a look at the R materials on Brightspace and use the exercises of week 1 to practice.

Part B

Description

In this second part of Assignment 1 we continue with the research question. To answer the question, we need some data, but this study is fictitious so no data exist.

We will "solve" this by doing a simulation study. Simulating data means that we get to decide what the real truth is. In this assignment, we simulate the data in such a way that no real effect exists! In other words, we will simulate data in line with the null hypothesis. For example, if you think that *length* and *height* are related, you actually simulate them to be independent of each other.

This is easier than simulating data where the alternative hypothesis (our expectation) is true, but we have another reason for doing this. It means that now we can use our fake data to simulate the type I error rate: the proportion of times we would find a significant result even though nothing is going one (H_0 is true).

Follow these steps, using R, while describing what you are doing:

Step 0. Translate your research question into an alternative hypothesis (expectation) and a null hypothesis ("nothing going on" or "no effect") and write these down.

Step 1. Simulate the variables needed to answer your research question (and, if you want, some other variables as well). Take into account:

- The variables should take on somewhat realistic values, so think about the measurement types and distributions. The ranges provided in the case descriptions are just suggestions, so you don't need to stick exactly to those.
- Remember, the simulation should be in line with the null hypothesis. This means that *in your simulation, the null hypothesis has to be true.*

Step 2. Perform the correct statistical test[1] on these variables to test your null hypothesis and extract the p-value of interest. Note that you have simulated the data in such a way that you know the null hypothesis is actually true! However, you could get a significant result just by chance (which would be a type I error).

Step 3. Write a loop to repeat the data simulation and testing (steps 1 and 2) a lot of times, and save all the p-values. You should now get a vector containing lots of possible p-values for a study that does a test where no effect exists (H_0 is true).

[1] Choose one of the tests from lecture 2.

Step 4. Describe these p-values, in text and graphs. What can you tell about the type I error rate, is it as expected?

Hand in

Hand in two documents:

1. A short description (max. 300 words) of what you have done and of the findings (including at least one graph) as a .pdf file.
2. Your R code, with comments (`# like this`) as a .R file. Note that all group members should agree on, and understand, each line of code.

Homework for the next meeting

If you did everything correctly, you should get some predictable numbers for the type I errors. Next week we will start messing with the analyses (by committing *Questionable Research Practices*) to change this!

In the meantime, improve your R skills with the exercises and ebooks. Also have a look at control flow (like `if`-statements), for example using the video *Conditional statements and loops* which is shared on Brightspace.

Part C

Description

Now we can finish Assignment 1 by adding questionable research practices (QRPs) to the simulation. What happens to the type I error rate if researchers base their data-analytic choices on the p-values they get? You have learned about this in the lectures and you can read a bit of a recap on page 6.

Choose one or a few questionable research practice(s) that the researchers could use in this study. You can choose one or more from this list:

- Sequential testing with optional stopping
- Removing outliers with some criteria, but choosing whether to use this (or which criteria to use) based on the test results
- Trying multiple dependent variables (so hypotheses) and reporting only those giving desirable results
- Reporting on specific levels (groups) of a nominal independent variable, depending on the results
- Adding covariates (additional independent variables) to your model, or removing them, to get a lower p-value for the main independent variable
- Rounding p-values down (e.g., $p = .056$ is treated as $p \leq .05$)

Copy your script from Part B (after making changes based on the feedback) and implement the QRP(s) of your choice. Don't change the way you simulate datasets, only the way you analyse them! You should see the effect of the QRP(s) on the distribution of the p-values. What is this effect, and what does that mean for studies where researchers commit a QRP?

Hand in

Hand in two documents:

1. The full text of the assignment (Parts A-C) as a .pdf file, which includes (updated versions) of:
   (a) The introduction section leading up to the scientific research question (Part A) and hypotheses (Part B).
   (b) A description of the QRP(s) you implemented (new)
   (c) An explanation of the methods/code (Part B and new).
   (d) A description of the results of the simulations without (part B) and with (new) the QRP.
   (e) A short discussion of what this means (new).

(f) A description of the author contributions: which group member did what?

2. The R code you used (part B and new), with comments (`# like this`).

In a group assignment, some division of tasks makes sense. However, each member of the group is responsible for the full end product. This also means that each member should be able to explain every line of R code.

You will be graded based on the files you hand in and the oral discussion of your assignment in the meeting of week 5.

Homework for the next meeting

Next week we will make start on Assignment 2. There is no specific homework, so you can just study the materials of week 1-3.

Background

Frequentist significance testing is based on the p-value: the probability of finding a sample (and therefore statistic) as extreme as or more extreme than the one we have found, *if in reality the null-hypothesis were true*. If this probability is low, we reason that our null-hypothesis is probably not true (since we know that our data is "true", because we have observed it). In other words: when we do a test, we collect data, test if that data seems plausible if "nothing is going on" (H_0 is true) and if the data is not plausible under H_0, then this null-hypothesis is probably wrong. Our threshold for "not plausible" is finding a p-value lower than α (often 0.05), so in those cases we reject the null hypothesis. If you struggle with the concept of p-values, it's a good idea to watch the Crash Course videos that are linked to on Brightspace.

This assignment is based on data simulation. By simulating our own data, we can now "know the truth". For example, if we simulate IQ scores for two groups of people, both with a mean IQ score of 100, we know there is no real difference in mean IQ between the two groups in our *theoretical population*. However, a statistical test performed on *a sample* can still give a significant result, because our data (sample) is random. In the simulation, we know that rejecting the null-hypothesis has to be a type I error: we say there is an effect, but in reality no such effect exists. The data was constructed in such a way that the null-hypothesis is true, but we reject it by coincidence when we do our significance test.

In the lectures and exercises, you have seen what the distribution of the p-value looks like if in reality the null-hypothesis is true. If all is right, for some proportion (α) of cases, the p-value will be below our significance threshold (that same α). In those cases, we commit a type I error. Researchers can make choices that inflate the proportion of type I errors, so it is no longer equal to α. This is problematic, because it leads to more published false positive findings.

In this assignment, you simulate how this happens. You simulate a (large) number of datasets, for which we know that the null-hypothesis is true. On each of these datasets you perform the appropriate statistical test (for example an independent samples t-test), but you do so while committing questionable research practices (QRPs). The lecture of week 3 covers several of these QRPs and you can choose which ones you want to simulate.

The goal of the assignment is to show how committing a QRP impacts the type I error rate. Doing the assignment should help you understand how hypothesis testing works, and why we have a "replication crisis" in science.

Case descriptions

Case 1

Study

One of the main selling points of teaching statistics in study programs is that students become critical consumers of science and science communication. The current study sought to examine whether statistics proficiency indeed predicts someone's ability to spot mistakes in news articles describing scientific discoveries. Participants read 14 news articles, seven of which contained a mistake, the other seven were accurate. For each article, they had to determine whether it was flawed ("fake news") or not.

Participants

237 people who have recently obtained a degree in computer science

Variables (name: variable measurement, approximate range or levels)

ID (participant identification number): categorical, $1 - N$

Gender: categorical, female/male/other

Age (in years): numerical, 20 (youngest) – 36 (oldest)

StatsCourses (number of stats courses passed, expressed in credits): numeric, either 5, 10, 15, or 20

GradeStats (average grade on the stats courses participants took): numeric, 5.5 (lowest) – 9.5 (highest)

ReadNews (how often participants read the newspaper): numeric, ratings on a discrete 7-point scale from 1 (never) to 7 (daily)

Performance (after reading each article, participants had to indicate whether it contained a mistake or not; this variable indicates the number of correct decisions): numerical, 0 (lowest) – 14 (highest)

Comprehension (reading comprehension score, after each article participants answered a two-alternative multiple choice question about the article): numerical, 5 (lowest) – 14 (highest)

GoalStudy (during the debriefing, participants were asked what they thought the goal of the study was): categorical, 1 (guessed correctly)/ 0 (guessed incorrectly)

Case 2

Study

Machine learning practitioners often seek to optimize model performance while minimizing computational resources. This study investigated factors affecting an AI engineer's ability to develop efficient machine learning models. Participants were given identical datasets and asked to develop models with the best performance-to-computation ratio. Their working habits, development environment preferences, and prior training were measured to identify which factors most strongly predict a practitioner's ability to create efficient AI systems.

Participants

184 machine learning engineers with at least 2 years of professional experience

Variables (name: variable measurement, approximate range or levels)

ID (participant identification number): categorical, $1 - N$

Gender: categorical, female/male/other

Age (in years): numerical, 23 (youngest) – 45 (oldest)

Education (highest degree obtained): categorical, BSc/MSc/PhD

AlgoTraining (formal training in algorithmic complexity, in course hours): numerical, 0 – 120

YearsExperience (years working in machine learning): numerical, 2 – 15

CaffeineDaily (average daily caffeine consumption in mg): numerical, 0 – 650

SleepHours (average hours of sleep per night): numerical, 4.5 – 9

ProgrammingLanguage (primary language used): categorical, Python/R/Julia/Other

DebugTime (percentage of development time spent debugging): numerical, 10 – 75

ModelEfficiency (composite score of model performance divided by computational resources used): numerical, 0.12 (lowest) – 0.89 (highest)

DocumentationHabit (quality of code documentation assessed by reviewers): numerical, ratings on a discrete 5-point scale from 1 (poor) to 5 (excellent)

DeadlinePressure (participant reported stress level during task): numerical, ratings on a discrete 7-point scale from 1 (none) to 7 (extreme)

Case 3

Study

The field of natural language processing relies heavily on pre-trained language models. This study examines various language models' performance on a range of linguistic tasks after controlled fine-tuning procedures. The research aims to determine which architectural features and fine-tuning approaches yield the best results across different languages and task types.

Participants

128 language models with varying architectures and training regimes

Variables (name: variable measurement, approximate range or levels)

ModelID (model identification number): categorical, $1 - N$

Architecture: categorical, Transformer/LSTM/GRU/Hybrid

Parameters (number of parameters in billions): numerical, 0.1 – 175

PreTrainingTokens (tokens seen during pre-training in trillions): numerical, 0.05 – 4.2

LanguagesCovered (number of languages in pre-training corpus): numerical, 1 – 104

AttentionLayers (number of attention layers): numerical, 0 – 96

SemanticScore (performance on semantic understanding benchmarks): numerical, 35.2 – 94.8

SyntaxScore (performance on syntactic tasks): numerical, 42.7 – 97.3

LowResourceScore (performance on languages with limited data): numerical, 22.5 – 88.6

OpenSource (availability of model weights and architecture): categorical, 0 (closed) / 1 (open)

Case 4

Study

University cafeterias are investigating how to reduce food waste using AI-powered prediction systems. This study monitored 32 university cafeterias in the US that implemented food waste prediction systems over a 6-month period. The research tracked various factors that might influence the effectiveness of these systems, including the type of prediction algorithm used, staff compliance with AI recommendations, and contextual factors like menu diversity and student demographics.

Participants

32 university cafeterias across different regions

Variables (name: variable measurement, approximate range or levels)

CafeteriaID (cafeteria identification number): categorical, $1 - N$

RegionUS: categorical, North/South/East/West/Central

StudentPopulation (number of students served): numerical, 800 – 12500

StaffSize (number of kitchen staff): numerical, 8 – 45

AITrainingHours (hours of staff training on the AI system): numerical, 2 – 40

AlgorithmType (complexity of prediction algorithm): categorical, Simple Regression/Random Forest/Neural Network

MenuDiversity (number of unique dishes offered per week): numerical, 28 – 124

WasteBeforeAI (average kg of food waste per day before implementation): numerical, 45 – 320

WasteAfterAI (average kg of food waste per day after implementation): numerical, 15 – 290

PredictionAccuracy (percentage of accurate demand predictions): numerical, 62 – 91

StaffCompliance (percentage of AI recommendations followed): numerical, 35 – 98

CostSavings (estimated monthly savings in GBP): numerical, 200 – 4500

StudentSatisfaction (student rating of food quality): numerical, ratings on a discrete 5-point scale from 1 (poor) to 5 (excellent)

Case 5

Study

This research examined the impact of AI algorithms on online dating success. The study collected data from a dating platform that implemented various matching algorithms, where sets of users were randomly assigned to an algorithm. Success was measured through match rates, conversation length, and self-reported satisfaction. The research investigated which algorithmic approaches led to the most successful matches and whether different groups benefited equally from algorithmic matching.

Participants

5,280 dating platform users who actively used the service for at least 3 months

Variables (name: variable measurement, approximate range or levels)

UserID (user identification number): categorical, $1 - N$

Age (in years): numerical, 18 – 75

Gender: categorical, female/male/non-binary/other

RelationshipGoal: categorical, Casual/Serious/Friendship/Undecided

PreviousRelationships (number of past significant relationships): numerical, 0 – 12

ProfileCompleteness (percentage of optional fields completed): numerical, 15 – 100

AlgorithmType (matching algorithm assigned): categorical, Behaviour-Based/Stated-Preference/Hybrid/Random

MatchRate (percentage of swipes resulting in matches): numerical, 0.5 – 65

AvgConversationLength (average message count per conversation): numerical, 1 – 250

DateCount (number of in-person meetings arranged): numerical, 0 – 28

UserSatisfaction (self-reported satisfaction with matches): numerical, ratings on a discrete 10-point scale from 1 (very dissatisfied) to 10 (very satisfied)

TimeOnApp (average daily minutes on platform): numerical, 2 – 120

RelationshipSuccess (whether user formed relationship lasting 3 > months): categorical, 0 (no) / 1 (yes)

Case 6

Study

Climate researchers collected data on a set of failed and successful machine learning models designed to predict extreme weather events. The study analyzed various model characteristics, training approaches, and data sources to determine factors associated with accurate predictions of hurricanes, floods, and heatwaves. Particular attention was paid to the models' false positive and false negative rates, as both have significant real-world implications for emergency planning.

Participants

94 weather prediction models from meteorological agencies and research institutions

Variables (name: variable measurement, approximate range or levels)

ModelID (model identification number): categorical, $1 - N$

DisasterType: categorical, Hurricane/Flood/Heatwave/Wildfire/Drought

ModelAge (months since model development): numerical, 2 – 84

DataSources (number of distinct data stream types): numerical, 1 – 12

SatelliteData (whether model uses satellite imagery): categorical, 0 (no) / 1 (yes)

HistoricalRange (years of historical data used): numerical, 5 – 150

UpdateFrequency (how often model is retrained, in days): numerical, 1 – 365

ModelType (type of model): categorical, Physical/Statistical/NeuralNetwork/Hybrid

LeadTime (days ahead of event prediction is made): numerical, 1 – 21

Accuracy (overall prediction accuracy): numerical, 0.54 – 0.91

FalsePositiveRate (rate of predicting events that don't occur): numerical, 0.02 – 0.45

FalseNegativeRate (rate of missing events that do occur): numerical, 0.01 – 0.38

ComputeCost (annual computing cost in thousands of GBP): numerical, 5 – 1250