Preparing for the open-book coding challenge

Applied Data Science 2, semester 2

Dmytro Shytikov (based on Duncan MacGregor’s slides)

2024-05-20

dmytroshytikov@intl.zju.edu.cn

Table of Contents

- GENERAL EXPLANATIONS
- DETAILS OF THE ASSESSMENT
- EXAMPLE – Q3 from ICA1
- GETTING READY FOR THE EXAM
- FINAL REMARKS

Learning outcomes

At the end of this lecture you should be able to:

- Understand how to prepare for the final open-book coding challenge;
- Understand what we expect you to do;
- Understand how the marking is done;
- Improve study skills;
- Practice open-book assessments.

GENERAL EXPLANATIONS

What is the coding challenge and why are we doing this?

The coding challenge includes:

- 3 problems (not just coding!); 25% each
- document everything in R Markdown 25%
- 3 hours in total
- semi-open-book. There is a vast list of websites about statistics and R that are *mostly* allowed (if they do not rely on some other external websites).
- The list includes (but is not limited to):
  - R Graphics Cookbook, 2nd edition
  - Cookbook for R
  - www.yihui.org
  - www.tidyverse.org
  - GitHub (but these websites are glitchy)
  - and many others

Why are we doing this?

Assessments are most effective if they are similar to what you would be doing “in real life”!

The semi-open-book format was chosen to make it difficult for you the use AI-based tools.

Packages that are available on the computers in the computer class

- `all basic packages`
- `tidyverse`
  - `dplyr`
  - `ggplot2`
  - `tidyr`
- `rmarkdown`
- `knitr`
- `tinytex`

All tasks can be solved using just these basic libraries.

Also, you should be able to install packages from https://cran.rstudio.com/.

What is allowed and what is not

Allowed:

- To install external packages from the Internet (https://cran.rstudio.com/ worked well);
- To access materials from BB Learn;
- To access websites from the white list (which will be published soon);
- Bring your printed or handwritten (the preferred way) materials.

Not allowed:

- To communicate with each other either in person, through chats, or any other means;
- To use your phone or other devices;
- To use external apps on the provided computer downloaded either from the Internet or from your USB drive;
- AI-based tools.

DETAILS OF THE ASSESSMENT

What we expect from you and why:

Our expectations	The reason for our expectations
You are able to handle data: import it, diagnose it briefly, and clean it if needed.	Data cleaning is an important part of your future work. We stressed its importance before. There was a very detailed tutorial about it.
You are able to make illustrative plots.	By year 2, you should know how to describe different types of data appropriately. We will talk about the details later.
You are familiar with the main analytical approaches *discussed during the study period*.	You were introduced to the respective approaches in all the details we could grant you including lectures, detailed solutions, and discussion with our course team.
You are able to produce nicely formatted and easy-to-read reports. We also expect a reasonable level of English	We do not expect your writing to be perfect, but it must be *at least* readable.

EXAMPLE – Q3 from ICA1

Spinal cord injury and novel biomaterials

Complete spinal cord injury (SCI) is a severe condition caused by the trauma of the spinal cord that leads to its breakage. Depending on the size and type of injury, it may lead to the loss of motor and sensory functions and disability. The common prognosis is bad, with less than 30% of patients being able to regain at least some functions over the affected body parts spontaneously within 1 year. The commonly accepted clinical scale to describe the patient condition is American Spinal Injury Association Impairment Scale (AIS scale), which has 5 grades: `A` (the complete motor and sensory functional impairment, complete disability), `B`, `C`, `D`, and `E` (the above-mentioned functions are normal, healthy subject).

Your team has developed a novel biomaterial that may help to regenerate injured nervous tissue after SCI. Studies on animals showed promising results, and now you perform a Phase I-II trial on patients with complete chronic SCI. Your team recruited patients with chronic SCI (at least 1 year since the trauma passed with no signs of recovery), recorded their condition upon admission, and installed implants with the novel material. 15-18 months later, you recorded patient AIS scores again.

The data is in the `SCI_before.csv` and `SCI_after.csv` files. Merge both tables appropriately, analyze the data, draw conclusions, and give further suggestions.

Questions

1. Import, and arrange the data (merge both pieces of data and make the data possible to analyse), and make it suitable for analysis, e.g. the values. You should perform all the manipulations in R and provide the code.
2. Check your data carefully. Identify features of the data and discuss your conclusions. Make illustrative plots.
3. Formulate the correct statistical hypothesis, choose the appropriate statistical test, and check assumptions for this test. Explain your choice briefly. Then, perform this test and identify whether the difference between the experimental groups is statistically significant.
4. Discuss the data you got. What did you obtain? Are there any flaws in the experimental design and what would you suggest to your colleagues? Support your statements with the appropriate statistics and/or effect size estimates.

S1 study program and Q3

What you studied during S1:

1. R basics
2. Probability basics
3. The idea behind sampling and sampling distribution
4. The idea behind hypothesis testing
5. Data cleaning
6. Data visualization
7. t-test and alternatives (including the MW U-test)
8. R Markdown
9. Data analysis pipeline

- Q3, part 1 – data cleaning. **Worth 6 points.**

  `merge both pieces of data and make the data possible to analyse` → the data requires cleaning

  `make it suitable for analysis` → the data must be treated somehow

- Q3, part 2 – data cleaning and visualization. **Worth 9 points.**

  `Check your data carefully` → the data requires cleaning

  `Make illustrative plots`

- Q3, part 3 – hypothesis testing, data analysis pipeline, t-test, and alternatives. **Worth 6 points.**
- Q3, part 4 – data analysis pipeline, general understanding of biology. **Worth 4 points.**

  `What did you obtain? Are there any flaws? Support your statements...`

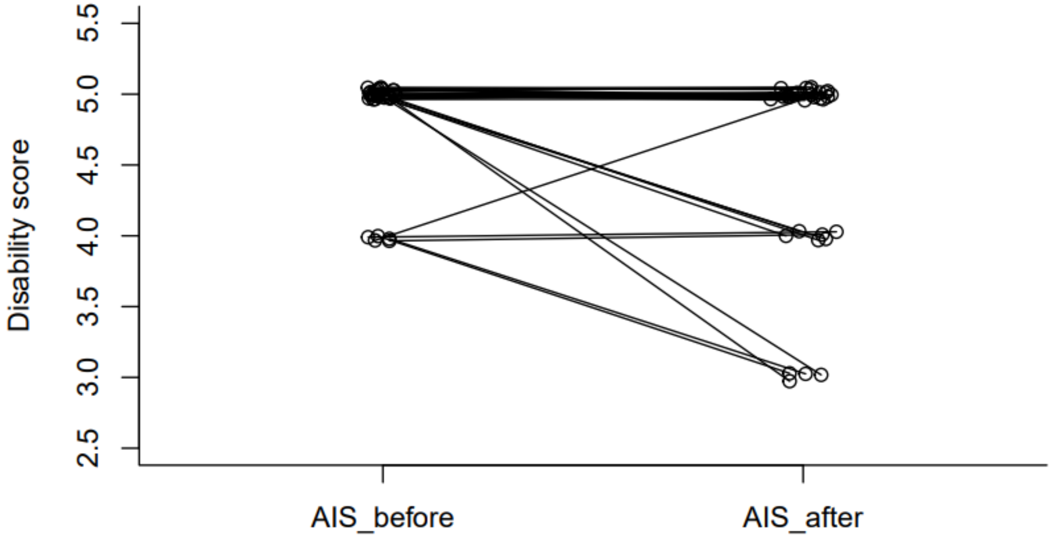
Data cleaning step

1. According to the task, you have reading *before* and *after* the surgery; the data are provided as two tables → likely, need to match patients.
2. The data scale is `A > ... > E` → the data are paired and quantitative, but discrete. So, we may transform it into numbers or ordered factors (basically, integers). We also need to merge tables.
3. As the data are paired, it is reasonable to calculate the *difference* between the AIS scores *before* and *after*.
4. A brief data cleaning can show the presence of some duplications.

Good plots

1. The dependent variable is quantitative → we need a quantitative scale;
2. The independent variable is qualitative → we need a discrete scale;
3. The data are paired → we need to emphasize it in your figure;
4. It is good to show the primary points;
5. The data are discrete → we need to avoid overplotting.

The plot clearly requires a lot of work, but we did not expect you to make it exactly like this.



Other viable options

A boxplot (with the primary points);

A histogram of the differences between both time points;

Any plot that can confirm the paired nature of data;

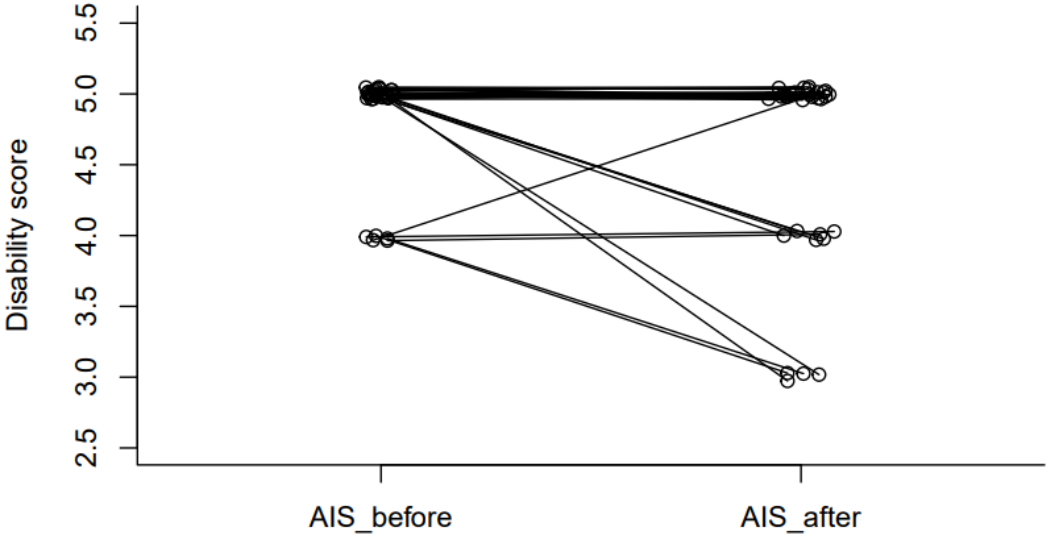
A table:

Scores before		**Scores after**	
	A	**B**	**C**
A	19	4	3
B	1	2	2

Choosing the statistical test

1. The dependent variable is quantitative → we need a test that can process a quantitative dependent variable;
2. The independent variable is qualitative → we need a test that can make use of two groups;
3. The data are paired → we need a paired test;

So, we need a test that can compare two sets of paired measurements. A paired t-test may work. What about its assumptions?



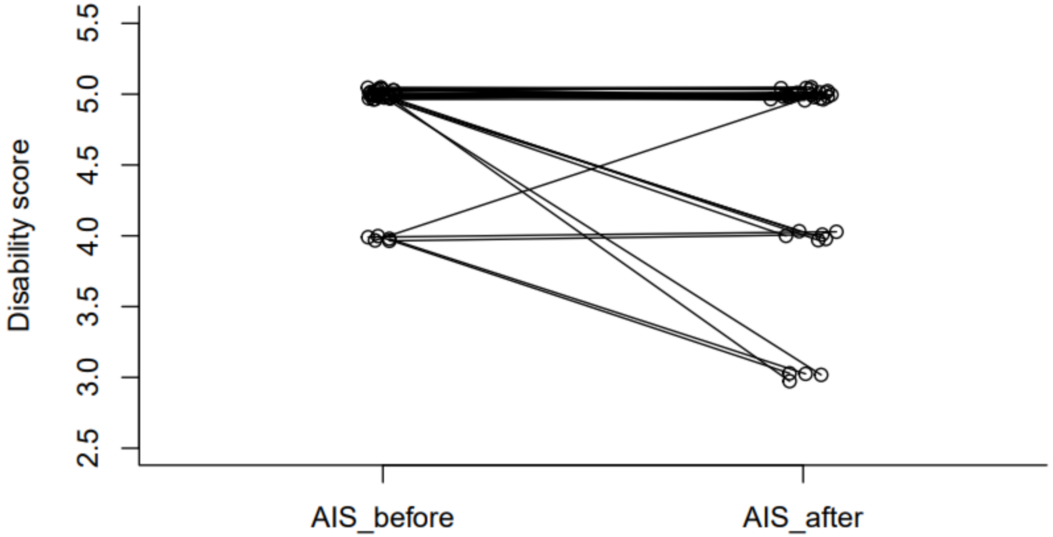
Choosing the statistical test

1. The dependent variable is quantitative → we need a test that can process a quantitative dependent variable;
2. The independent variable is qualitative → we need a test that can make use of two groups;
3. The data are paired → we need a paired test;

So, we need a test that can compare two sets of paired measurements. A paired t-test may work. What about its assumptions?

1. The data are not normally distributed → a parametric test is a bad choice here.
2. The data are discrete → a t-test is a bad choice here.

Something non-parametric should be used. The possible choice `wilcox.test()` was mentioned in the respective lecture.



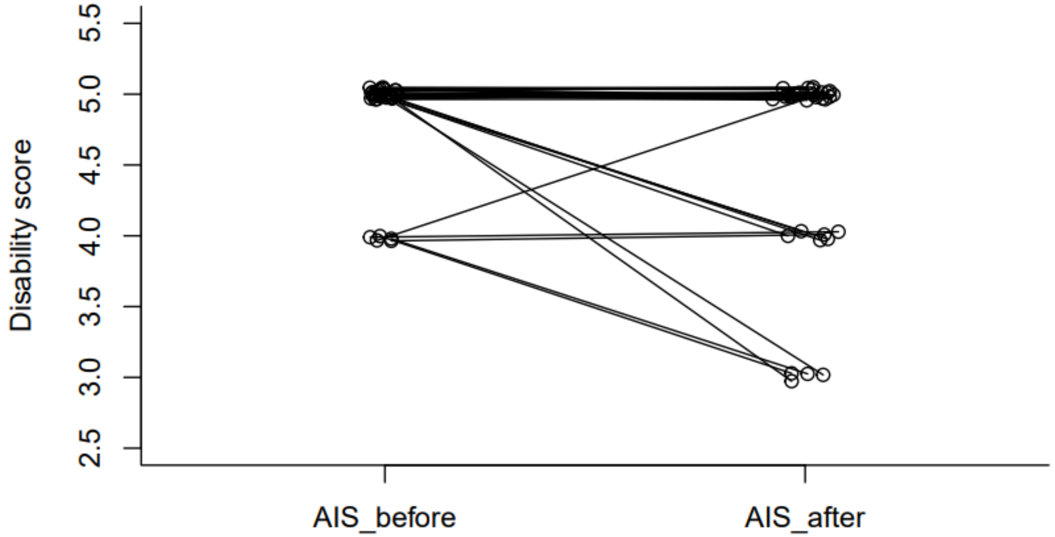
Statistical hypotheses

- H_0: the true median (or average, depending on your test) *score change* is 0;
- H_A: it *is not* 0.

It would give 2 points

A one-sided hypothesis is legal here, but you must *clearly* explain why you think so (because most of the patients are *already* completely disabled. Their condition cannot worsen further than the A score).

The one-sided hypotheses would give only 1 point **if not justified**.



Discussion

- Summarize your results and provide the mean difference (p-value): 0.3 points of the AIS score ($P_U < 0.05$). Adding Cohen’s d or similar would be valued.
- *Formally*, a placebo group would be required. But how do you imagine this?
- Why do some patients improve and some do not? Did we take into account all the data? Clinical condition? The location of the trauma?
- Is there at least *any* improvement in those patients whose A score did not change? Maybe their sensitivity got better?
- There is no data about toxicity.
- Can you imagine something more creative? What about incorporating stem cell-derived factors into the implants? How about coating them with autologous stem cells?
- After the study, $p_{(U)} \approx 0.025$. What will you do with the p-value reduced further? Will it change your opinion?

GETTING READY FOR THE EXAM

How to prepare for an open-book assignment?

You can look up anything, but you cannot look up everything!

- Have your code and lecture materials/notes where you can find them;
- Label code that did and didn’t work;
- Write good data-wrangling maneuvers;
- Make a hand-written “cheat sheet”, e.g. of frequently-used commands, statistical tests, and their assumptions, and plots;
- Bookmark what you can find in each of the allowed resources (remember to cite when using in assessment) ;
- Make sure you are familiar with the assessment instructions (check Blackboard Learn)

Simple data cleaning maneuvers

- Check for `NAs`;
- Check for duplications;
- Check which values are in each column (`table()`, `unique()`, any other ideas?);
- Convert to the wide format if applicable;
- Arrange and merge data if it looks reasonable;
- Make your data **suitable for your analysis**.

If you see `Check your data carefully`, you should be careful about the data. It may require data cleaning.

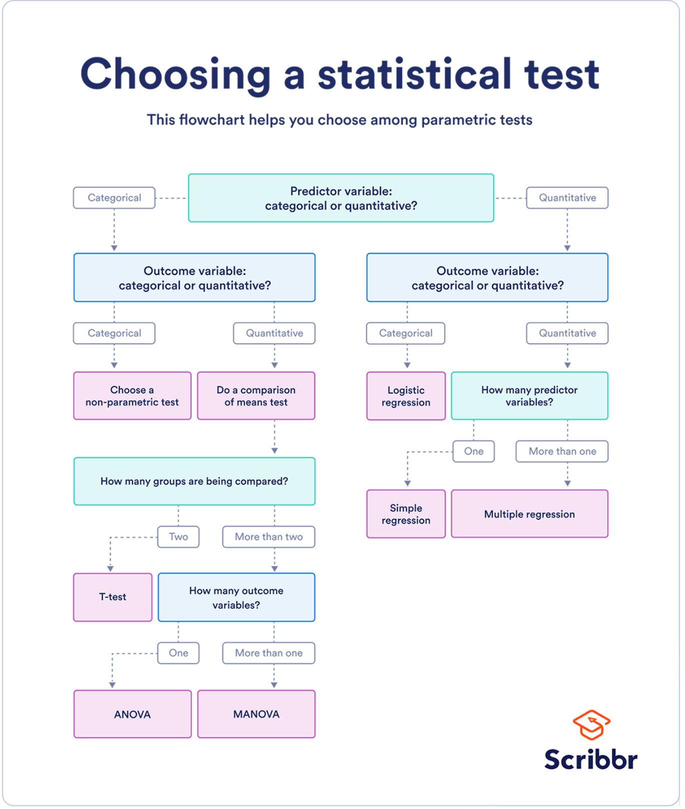
How to describe continuous data

Variables	**Plot**
Independent: categorical	Boxplot
Data: unpaired	Whisker plot (+SD/SE/CI/IQR)
	Strip chart
Independent: categorical	All above with the primary points that are linked
Data: paired	Histogram of the *difference*
Dependent: continuous	2D scatter plot or similar
Independent: continuous	A trend line would be helpful

How to describe categorical data

Variables	Plot
Independent: categorical	Barchart
	Table (with percentages)
	Pie chart

How to deal with data



Choosing a statistical test

This flowchart helps you choose among parametric tests

Predictor variable: categorical or quantitative?

- Categorical
  - Outcome variable: categorical or quantitative?
    - Categorical → Choose a non-parametric test
    - Quantitative → Do a comparison of means test
      - How many groups are being compared?
        - Two → T-test
        - More than two → How many outcome variables?
          - One → ANOVA
          - More than one → MANOVA
- Quantitative
  - Outcome variable: categorical or quantitative?
    - Categorical → Logistic regression
    - Quantitative → How many predictor variables?
      - One → Simple regression
      - More than one → Multiple regression

Scribbr

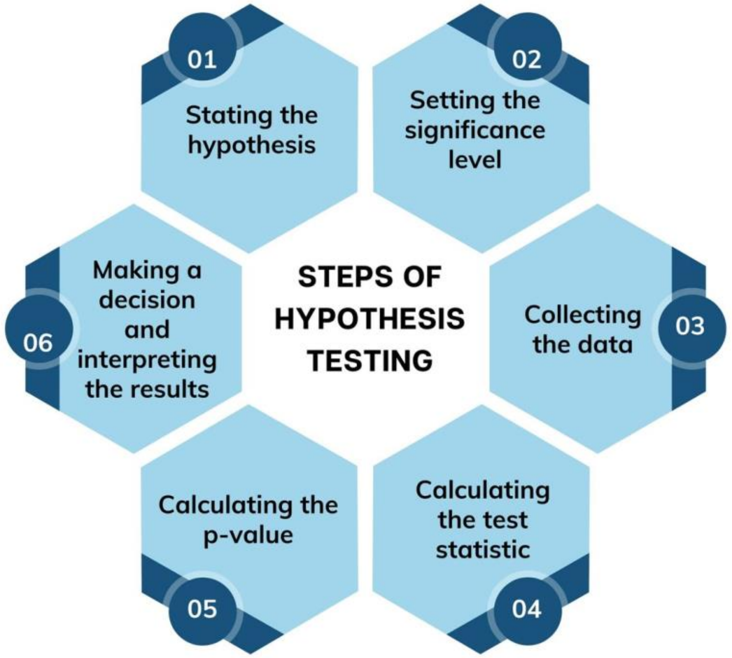
[www.scibbr.com, Rebecca Bevans, 2020](www.scibbr.com)

Why do we want you to write statistical hypotheses?

We want you to demonstrate that you understand *how* the respective test works, but not do it mechanically just because you heard it is used under these circumstances.

Some approaches (bootstrapping) *require* you to understand what you are doing.

In some cases, better results can be obtained if you use the one-sided hypothesis vs the two-sided one *correctly*.


STEPS OF HYPOTHESIS TESTING
01
Stating the hypothesis
02
Setting the significance level
03
Collecting the data
04
Calculating the test statistic
05
Calculating the p-value
06
Making a decision and interpreting the results

www.pickl.ai

How to discuss results?

- Formal summary of your results with some statistics, p-value (but **not** only the p-value), possibly, effect size estimate, and your conclusion.
- What do these data mean? Why do you think it means something good/bad?
- Are there any shortcomings in the study (apart from the sample size)? Was something not done? Is the mechanism of the effect clear? Are all the clinical covariates included in the explanation?
- Is the power of the study good enough? No? Why do you think so? Can you say how many replicates to add? *Power analysis would make your words more convincing*!

www.uk.assignmentgeek.com

Learn to knit!

- Prepare a list of useful commands for `knitr` code chunks and YAML headers.

Got troubles?

- Check whether the `.MD` and `.TEX` files are generated at all and whether there are some weird symbols:

```
---
...
output:
  pdf_document:
    keep_md: true
    keep_tex: true
---
```

- Reinstall TinyTex:

```
install.packages('tinytex')
tinytex::reinstall_tinytex()
```

- Change the TeX engine
- Let me know.

www.bookdown.org

FINAL REMARKS

Conclusions

Now, you should understand:

- how marking is done;
- what we *likely* look at in your submission;
- what *likely* to expect during the exam;
- what *likely* to use in some situations;
- how to prepare for the exam;
- there will be 2 test sessions.

Further suggestions

```r
data() # check in-built R datasets

data("Indometh") # paired data ANOVA
data("ToothGrowth") # 2-way factorial ANOVA
data("diamonds") # many-many things...
data("cars") # regression and correlation
data("iris") # many-many approaches...
```

Or you can check for some data online.

THANK YOU FOR ATTENTION!