Introductory Probability and Statistics: End-of-course summative assessment

INSTRUCTIONS (PLEASE READ CAREFULLY!)

- Your solutions should be handed in **individually** in the form of a report (in pdf format) on Gradescope. The report should **not be longer than 5 pages**. Any additional information must be included in the appendix.

- Please do not put your name on the report. Instead, use your student number. The file name for your report should be `[your exam number].pdf`.

- You are **not allowed** to discuss your solution with other students. You are **not allowed** to copy answers from other students, or from other sources. Any other sources that you use (e.g. books, webpages) **must be cited**. Your report will be checked for plagiarism, both using Turnitin and manually.

- Academic integrity is an underlying principle of research and academic practice. All submitted work is expected to be your own. AI tools (e.g., ELM) should not be used for this assessment. Using AI without authorization might constitute academic misconduct.

- Your written answers should be concise and written in such a way that it is perfectly clear what has been done and why. **Do not directly copy and paste software outputs.** Present results using nicely formatted tables and graphs.

- We strongly recommend that you use RStudio for the assignment as it is the best tool for this kind of work, but you are free to use any other software you are familiar with (e.g. Python, SPSS, ...).

- Regardless of what software you use, you must make your code/syntax available to the instructor and for peer review (see below).
  - Include your code in an Appendix, typeset using a `typewriter font`.
  - For code and data sharing, either (a) host the code and data in a GitHub repository and provide the link at the end of the report; or (b) submit the files via Gradescope. In both cases, include all data required to reproduce your results. Links may include your name; anonymisation is not required.

  - If using other software, ensure your analysis is reproducible and provide clear setup instructions (software versions, required packages, and any configuration).

- The **deadline** for handing in your report is on Tuesday 11 November at 13.00.

- You will then be assigned two other students' reports to review. Please submit your peer reviews by 13:00 on Tuesday 25 November. Guidance on how to write your review is provided on pages 5–8 of this document **(you are strongly advised to read this before submitting your report)**.

- Feedback on the report will be available on 2 December, and feedback on the peer review on 9 December. The assignment as a whole will be graded on the basis of the quality of your report (80%) and the quality of the reviews (20%) you write.

- Deadlines are strictly enforced, and late submissions will be subject to the University's standard late submission policy. Short extensions can be granted under certain specific conditions, as laid out in the University's Postgraduate Taught Regulations, but note that these specifically exclude the case of external work commitments. Anyone seeking an extension should contact me as soon as possible, and before the due date for the assignment.

Assignment: Biomarkers and pain (3 pages)

Your task is to analyse a dataset from a research project from two Scandinavian university hospitals. As a part of this process, you will have to manipulate the dataset in different ways (e.g. only using a subset of the observations for some of the tasks, and merging data from two files). You can do this by hand in Excel (time-consuming and not encouraged) or more efficiently using R (highly encouraged). There are various resources to help you learn how to manipulate data using R, such as Chapters 5 and 13.4 in the *R for Data Science* book at `https://r4ds.had.co.nz/`; however chapter 5 of *Modern Statistics with R* uses a version of this dataset as an exercise - so may be of most use. Please note, the data is changed each year (to prevent plagiarism), so make sure to use the version of the data found on the Learn page for your assignment and not the one associated with this textbook!

The study is concerned with patients that have a medical condition causing pain (see Moen et al., 2016). They were asked to rate their pain on a 0-10 scale (called VAS; 0 is no pain and 10 is the worst imaginable pain), both at time 0 (inclusion, during the acute phase of the condition) and at a 12-month follow-up. Blood samples from the patients, taken at inclusion, were analysed, and the levels of certain proteins ('biomarkers') believed to affect the condition were measured (available in the file `biomarkers.xlsx`). In addition to the biomarkers, other covariates were also measured for all patients (available in the file `covariates.xlsx`).

In the first column in `biomarkers.xlsx`, the patient ID and the timepoint for the blood sampling is shown (0 weeks, 6 weeks or 12 months after inclusion). As an example, `126-0weeks` denotes the measurements for patient 126 at 0 weeks. The rest of the columns represent the biomarkers, randomly selected from the full original dataset, which are as follows:

1. Interleukin-6 (IL-6)
2. Vascular endothelial growth factor A (VEGF-A)
3. Osteoprotegerin (OPG)
4. Latency-associated peptide transforming growth factor beta 1 (TGF-beta-1)
5. Interleukin-8 (IL-8)
6. C-X-C motif chemokine 9 (CXCL9)
7. C-X-C motif chemokine 1 (CXCL1)
8. Interleukin-18 (IL-18)
9. Macrophage colony-stimulating factor 1 (CSF-1)

These biomarkers were chosen in the original study because they are related to inflammation - some are pro-inflammatory and some anti-inflammatory. However, it is unknown whether increases or decreases in these biomarker values cause more pain.

Depending on your educational background you may already be familiar with these biomarkers, but do not worry if you are not - you task is to be the statistician on a hypothetical research team explaining your solution, results, and conclusions to colleagues who are writing a research paper with you. Your introduction to the report only needs to have enough information to justify the hypotheses you make. The majority of your report should therefore address the following two tasks:

1. **Statistical hypothesis testing.** The researchers are interested in several questions about the levels of the different biomarkers. For instance: do the levels at inclusion vary between males and females? Do the levels at inclusion vary from those 12 months later? From those 6 weeks later? Do the biomarker levels at inclusion for patients with high VAS (≥ 5) differ from those for patients with low VAS (< 5)? Answering these questions will help the researchers understand how the biomarkers are related to the pathophysiology of the condition. Your task is to:
   (a) Choose one of these questions. Note that the researchers want you to answer the question **for each biomarker in the material** (i.e. "Is there a difference for IL-8? Is there a difference for VEGF-A?" and so on)! Describe the question and why you think it may be of interest.

(b) Formulate the question as hypotheses about parameters of distributions (make sure to describe what your random variables and your distributions are!).

(c) Perform suitable hypothesis tests to test your hypotheses. Describe which test you used and why. Draw conclusions based on the results of your tests.

(d) You will now have performed multiple hypothesis tests.

i. Describe what the potential problems with multiple testing are. Calculate the probability of making at least one type I error assuming that your tests are independent and that all null hypotheses are true.

ii. One common remedy for problems associated with multiple testing is called *Bonferroni correction.* Search for information about Bonferroni correction online or in books. Describe what it is used for and explain how to use it. Then redo your tests using Bonferroni correction. Draw conclusions based on the results of your Bonferroni-corrected tests.

2. **Regression modelling.** Up until now, it has been difficult to make predictions of how well patients with this medical condition will recover. Your task is to construct a regression model using the 12-month VAS as the response variable and biomarker levels (at inclusion) and covariates as explanatory variables.

(a) Describe your model. Fit the model, but only use data from 80 % of the patients (see part (c) below), and present the fitted parameter values in a table.

(b) Discuss how well the model fits the data.

(c) Use your model to make predictions for the remaining 20 % of the patients. Compare their predicted 12-month VAS to their actual 12-month VAS. Discuss your findings. (This is called an *out-of-sample evaluation* of the model, as the data used for evaluating the model is different from the data used for fitting the model.)

(d) In conclusion, do you think that the model can be useful for predicting the 12-month VAS of patients (and therefore their pain status one year after onset)?

INSTRUCTIONS FOR PEER REVIEW (FOUR PAGES – PLEASE READ CAREFULLY!)

- You will be assigned reports from two other students, which you will be asked to review.
- Your review will consist of scores and comments. Your scores will **not** directly affect the score we give to the report (which counts towards the student's grade).
- Your review should follow the template provided on pages 6-8 of this document.
- Your review should be handed in as pure text in the review form on Gradescope.
- Before you start writing your review, read the report all the way through.
- While I will ask you to give the report a score, bear in mind that your role is that of a fellow statistician checking someone else's work, not that of a grader or examiner. The goal is (a) to help your fellow student improve their analysis and their statistical skills and (b) to improve your own skills in reading and commenting upon statistical analyses. Keep comments **constructive**!
  - Be constructive and explain your comments. **Avoid:** "This is a poor solution. You should use method Y instead.". **Use:** "Because condition X is not fulfilled, it would be better to use method Y, which..."
  - Offer suggestions, not commands. **Avoid:** "You need to...". **Use:** "You could...", "I'd advise you to...".
- The deadline for your reviews is Tuesday 25 November at 13.00.
- The assignment as a whole will be graded on basis of the quality of your report (80%) and the quality of the reviews you write (20%). Your reviews will be graded based on quality rather than quantity.
- You are **not allowed** to discuss your reviews with other students. Any other sources that you use (e.g. books, webpages) **must be cited**. Your reviews will be checked for plagiarism, both using Turnitin and manually.
- Feedback will be available on 9 December
- Deadlines are strictly enforced, and late submissions will be subject to the University's standard late submission policy. Short extensions can be granted under certain specific conditions, as laid out in the University's Postgraduate Taught Regulations, but note that these specifically exclude the case of external work commitments. If your extension has been granted, please send me an email as soon as possible, and before the due date for the assignment.

Template for peer review

Your peer review should consist of scores and comments based on the following five points.

1. Statistical methodology. Score X/20

   - Scoring example 1: the author was asked to create a regression model but instead computed a correlation. The author has solved the wrong problem. Score 0/20.
   - Scoring example 2: the author attempts to use the correct methods and adequately describe the methods and underlying assumptions, but makes some mistakes with the software, leading to incorrect results. Score 15/20.
   - Scoring example 3: the author uses good methods but does not provide any motivation for using them and does not discuss any underlying assumptions. Score 10/20.
   - Scoring example 4: the author uses good methods and gives a clear description of the methodology, along with motivations and a discussion of the assumptions behind the methods. Score 20/20.
   - **Write a comment regarding the following:** Are the statistical methods and plots used applicable for this data? Are they used correctly? Are the underlying assumptions described and checked? Do the computations appear to be correct? Are there any methodological choices that you disagree with? If so, explain why and suggest alternative approaches (if available).

2. Replicability. Score X/5

   - Scoring example 1: it is not clear how the data was merged, if any patients were removed from the data and how the analysis was performed. Score 0/5.
   - Scoring example 2: it is perfectly clear how the analysis was performed (R code is included), but not how the data was manipulated prior to the analysis. Score 2/5.
   - Scoring example 3: it is clear from the text how the analysis and data manipulation was performed, but there is no R code included in the report. Score 3/5.
   - Scoring example 4: it is perfectly clear how the analysis and data manipulation was done and the corresponding R code is included in an appendix. Score 5/5.
   - **Write a comment regarding the following:** Is the data handling and the methods used described in such a way that you would be able to replicate the analysis? Is working R code included in an appendix? If not, describe what needs to be added to make the work replicable.

3. Conclusions. Score X/5
    - Scoring example 1: the statistical analysis shows that the levels of several biomarkers are higher for females than for males, but the author writes that the levels are higher for males for for females. Most of the conclusions in the report are incorrect. Score 0/5.
    - Scoring example 2: some of the conclusions in the report are incorrect. Score 2/5.
    - Scoring example 3: all conclusions drawn from the analysis appear to be correct. Score 5/5.
    - **Write a comment regarding the following:** Are the conclusions drawn from the analysis correct? If not, explain why.
4. Presentation. Score X/5
    - Scoring example 1: it is not possible to understand the report without reading the instructions for the assessment. Score 0/5.
    - Scoring example 2: it is not possible to understand parts of the report without reading the instructions for the assessment. Score 2/5.
    - Scoring example 3: it is possible to understand all of the report without reading the instructions for the assessment, but the text is difficult to follow (e.g. far too much text, bad grammar). Score 3/5.
    - Scoring example 4: the text is easy to follow and it is possible to understand all of the report without reading the instructions for the assessment. Score 5/5.
    - **Write a comment regarding the following:** Is it possible to understand the report without reading the instructions for the assessment? Are the problems solved in the report clearly described? Is it possible to understand what has been done and how conclusions were reached? Are there any parts of the report/analysis that are unclear and would benefit from clarification? If so, what needs to be clarified?
5. Citations. Score X/5
    - Scoring example 1: citations are needed but not provided. Score 0/5.
    - Scoring example 2: the author only uses methods covered in the courses and so no citations are needed. Score 5/5.
    - Scoring example 3: citations are needed, but only provided in some cases. Score 2/5.
    - Scoring example 4: citations are provided where needed. Score 5/5.

- **Write a comment regarding the following:** Are citations provided for methods not covered in the course? Are books and webpages that have been used cited?

Finally, you should provide a total score: X/40, obtained by summation of the five scores above.

Example

```
Statistical methodology: 14/20.
Comment: ...bla bla... The methods used in part 1 (c) are well suited to
the problem. However, in 1 (b) you use a one-sided alternative hypothesis,
while in 1 (c) you use two-sided tests instead of one-sided tests, meaning
that you in fact are testing against a two-sided alternative hypothesis. I
believe that a one-sided test would have been appropriate in this case.

Replicability: 5/5
Comment: The description of the data manipulation process and the analyis
are clear. The R code for both are included in an appendix. I found it interesting
that you used the subset-function to get a subset of the observations in part
1 (I used the filter-function instead!).

Conclusions: 2/5
Comment: ...bla bla...

Presentation: 3/5
Comment: ...bla bla... Figure 3 is a very good illustration of the data,
but would have been easier to understand if you had put labels on your axes.
...bla...

Citations: 5/5
Comment: ...bla bla...

Total score: 29/40
```