MATH5741M Coursework

Stuart Barber (s.barber@leeds.ac.uk)
Wednesday November 08, 2023

1 Important details

- The coursework will be a short report on a data analysis and will count for **20% of your final mark** for this module. It will only be assessed once (there will be no resit).
- Your completed coursework should be submitted as a single PDF file before the deadline of **2pm on Friday the 15th of December 2023**.
- There is a **limit of 1200 words** and also a limit of **5 pages**. The word limit excludes references, appendices, acknowledgements, tables, equations and figure/table captions. The page limit excludes references and appendices. Writing a report in significantly fewer words/pages is acceptable, and even encouraged (efficient writing is an important skill), so please don't force your report to be as long as possible.
- Please include your name and student identification number on the first page, and at the end attach a signed academic integrity form, which will be available in Minerva.
- Mathematical working/equations should **not** be included (e.g. there is no need to include the formula for a confidence interval or test statistic), and R code should be only be included in an appendix (appendices will not be marked but might be used to check your results).
- This coursework is intended not just as a method of assessment but also as part of your learning, by reinforcing the topics covered each week. You are expected to work on this report continuously **throughout the second half of the**

semester, applying new methods as you learn them. Do not leave everything until the final week.

- You are allowed (even encouraged) to discuss this coursework with other students, especially in the practical classes. However, your final report should be your own work. All reports will be checked for collusion and plagiarism. You should analyse only your own unique data set for your report.
- An example report from a previous student **will not** be provided. Too many students would just blindly copy that report, leading to many cases of plagiarism. Instead, detailed guidance on report writing is provided in this document.

2 Objectives

Dogs that are stray, unwanted or neglected are often sent to animal shelters to be rehomed. An important outcome for each dog is the time taken to rehome them. Previous research has suggested that, on average, it takes around 27 weeks to rehome a dog, with a variance of around 74. The author of that research believes that this average applies to all breeds of dog.

However, other researchers believe that the average time taken to rehome dogs depends strongly on their breed and will differ from 27 for some breeds. The goal of this coursework is to investigate this by analysing a sample consisting of three different breeds of dog.

Figure 2.1: Six dogs of various different breeds. How long should we expect each to spend at the shelter before being rehomed?

3 The data set

To obtain your data set, you need to follow three steps:

- Download the file `rehoming.Rdata` from Minerva and load it into R.
- Run the code `createsample(x)` in R with your student identification number in place of `x`.

This will create a data set called `mysample` in R, which you are advised to save in a separate file for future use, for example by typing `save(mysample, file = "mysample.RData")`. Each row consists of data for a single dog. You should analyse **only** the data in `mysample` for your report.

The `mysample` data set has the following columns:

- `Rehomed`, which is the dog's total rehoming time. This is the number of weeks from the dog's arrival at the shelter until being rehomed with a new owner.
- `Visited`, the number of weeks from the dog's arrival at the shelter until his/her first visit from a potential new owner.
- `Health`, a measure of the dog's physical health upon arriving at the shelter, on a scale of 0 (worst possible health) to 100 (perfect health).
- `Breed`, the breed of the dog, for example Staffordshire Bull Terrier, Whippet, etc.
- `Age`, categorised as "puppy" or "fully grown". Precise ages are not given.
- `Reason`, the reason the dog was taken to the shelter. For example this could be because the dog was stray, dangerous, neglected, unwanted by its owner, or had a health condition.
- `Returned`, a binary variable which tells us whether the dog has previously been rehomed but was returned to the shelter for any reason.

It would be sensible to check which breeds are included in your sample by running the R code `table(mysample$Breed)`. You should have only three different breeds of dog in your data set.

4 Analysis plan

Data cleaning: Use R to clean the `mysample` data set by removing any rows with missing observations for either rehoming time (missing values are recorded as `99999`) or breed (missing breeds are recorded as `NA`). In your report, you should note the number (and percentage) of observations that you removed for each reason, but you should not give a very detailed description of the data cleaning process. (See Practical 1.)

Data exploration: Split your data by breed, so that you have three samples of dogs from three populations (one population for each breed). Create a table or tables of numerical summaries to give a rough overview of how these three samples compare to each other (e.g. in terms of health, age, etc.). You don't need to include all of this information in your report, just any features that seem particularly interesting or relevant (if the breeds are found to differ in any meaningful way, this might help to explain the final conclusions of the report). Most importantly, you should summarise (numerically and graphically) rehoming time for each breed. Interpret any summaries you present as clearly and thoroughly as you can. (See Chapter 2 in the lecture notes.)

Modelling and estimation: Based on your graphical summaries, try to propose distributions that might be reasonably used to model rehoming time for each breed (see Chapter 3 in the lecture notes). Estimate the parameters of each of your proposed models (Chapter 4). It is acceptable to say that **none** of the distributions you know seem suitable. If you don't think there is a suitable model among the distributions used in this module, explain why (i.e. say which features of your data make these distributions unsuitable). Check the suitability of any proposed models (see Practical 7).

Inference: For each breed, calculate a confidence interval to test whether mean rehoming time is 27 weeks. State which types of confidence interval (i.e. based on a z test or based on a t test) you used for each breed and **state very clearly why** you made these choices. You should only calculate one interval for each breed, but you can use different types for different breeds. Summarise your results in a table or graph, and state your conclusions/interpretations clearly. (See Chapter 5 in the lecture notes.)

Discussion: Interpret your findings. Do your results have **practical** significance (this is not the same as statistical significance)? Discuss any real-life implications of your results and any directions for future research. Discuss any **limitations** of your analyses, such as any assumptions that might be in doubt. Remember here that your goal is not a perfect analysis, so sometimes it will be necessary to perform an analysis under assumptions that are not ideal. The goal is the best analysis that you can manage with the tools and data available to you, but you must therefore be honest and transparent when discussing the flaws of that analysis. You should also discuss any particular **strengths** of your analysis (think about which features of the data were advantageous to you).

Comparison: For each pair of breeds, calculate a confidence interval to test whether mean rehoming time is the same for both. Briefly state which assumptions you made when calculating the confidence interval for each pair and why you made these assumptions (you should only calculate one interval for each comparison, but you can use different assumptions for different pairs of). Summarise your results in a table or graph, and state your conclusions/interpretations clearly. (See Chapter 6 in the lecture notes.)

5 Report writing guidance

The most important advice is that **you should interpret** your summaries, graphs, tables, confidence intervals, and the results of any hypothesis testing. It is unreasonable to expect the reader to interpret your results and draw conclusions for you, and this will be punished harshly in the marking.

For any report you write (or any poster, oral presentation, etc.), you should always consider the **target audience**. This will help you to decide how much you need to explain and how to explain it. For this report, imagine that your audience is familiar with basic statistics (including hypothesis testing and confidence intervals) but that they do not have access to your data set and are not familiar with (and are not interested in seeing) R code/output.

For general advice on presenting statistical reports, see:

- Making Data Meaningful: a guide to writing stories about numbers
- Making Data Meaningful: a guide to presenting statistics

You can find many good examples of statistical reporting in medical journals. For example, the ROLARR study.[1] In particular, pay attention to the way that each estimate is typically accompanied by a measure uncertainty (e.g. a standard deviation or a confidence interval), whether in the text or in the tables. The graphs/tables are thoroughly explained and self-contained in such a way that they would be quite clear even if presented separately to the paper (of course, your graphs/tables will be different, but the style and clarity are the main point here).

It is important that you split the report into meaningful sections (e.g. Introduction, Results, Conclusions/Discussion). A sensible structure would be as follows.

Title

Give your report a sensible title that immediately tells the reader what it is about. “Report” and “Coursework” are not acceptable titles.

Introduction

The introduction section should explain, in plain English (that could be understood by a non-statistician), the background and objectives of your analysis. For example, what questions do you hope to answer using your data? What are the variables of interest and why do they matter? You should also provide some context. For example, what breeds of dog do you have in your sample and what are their main characteristics? But you should avoid numerical or graphical summaries here unless absolutely necessary. All pages should be clearly numbered, starting from the introduction.

Results

Numerical summaries should be presented here and tabulated sensibly. If the summaries are split into categories, make it clear which category each summary corresponds to (e.g. we see summaries from `galaxies.Rdata` split by galaxy type in Table 5.1 here). All tables and graphs should be numbered and should be referred to by number in the text. Summaries from the tables should not be repeated in the text unless absolutely necessary.

State from how many observations (n) each summary was calculated. Summaries of location (e.g. mean or median) should be accompanied by summaries of spread (e.g. standard deviation or IQR/both quartiles). Try to avoid using tables and graphs that just communicate exactly the same information (for example, if you have a forest plot of confidence intervals then you don't need a table of the same confidence intervals).

Table 5.1: Median (with interquartile range in parentheses) of effective radius and number of globular clusters for galaxies of three visual types (spiral, lenticular and elliptical).

	Spiral (n = 7)	**Lenticular (n = 9)**	**Elliptical (n = 29)**
Effective radius (light-years)	11100 (7800)	8300 (3400)	17800 (19300)
Number of globular clusters	300 (370)	400 (404)	1242 (4370)

Tables and graphs should have descriptive captions. They should clearly state which summaries are presented (means/medians/quantiles/etc.) and any units of measurement. The numbers in them should be presented to a sensible number of decimal places (relative to how accurately the original data were recorded). Numbers in tables should usually be right aligned.

Abbreviations and mathematical notation should be unambiguously defined the first time they are used. Variables should be properly named in clear English, not just referred to by their column names in R (e.g. "Effective radius", not just "Eradius" or `galaxies$Eradius`). Axes on graphs should be labelled, as in Figure 5.1. The caption of a graphical summary should state the type of graphical summary (scatter plot/box plot/histogram/etc.).

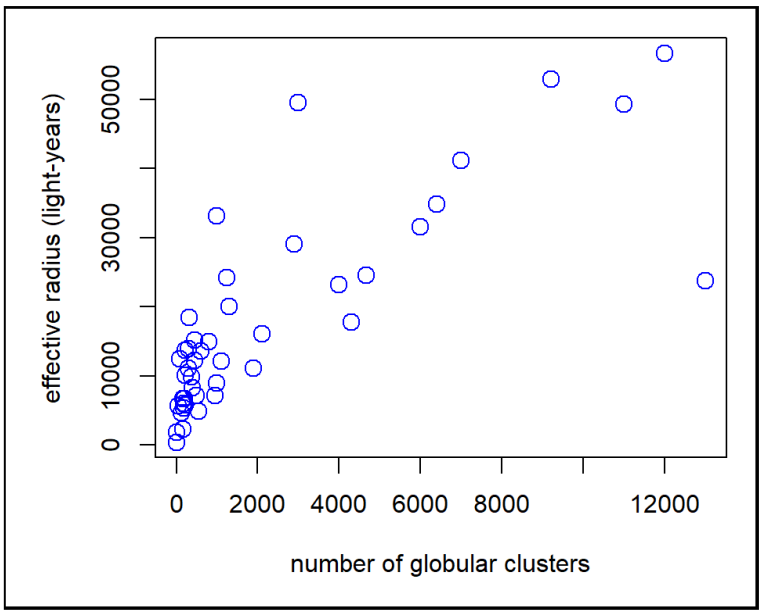


Figure 5.1: A scatter plot of effective radius against number of globular clusters for 45 galaxies with central black holes of known mass.

Do not summarise variables that are not relevant. Do not include graphs or tables that you will not interpret in the text. Do not copy and paste chunks of R output directly into your report (put only the relevant numbers into a properly formatted table instead). Do not expect the reader to interpret the graphs or summaries for you. You also do not need to present an example of every numerical/graphical summary that you have learned. It is better to just present the summaries that you consider most appropriate for describing your data.

Only perform one test of each hypothesis. You need to choose the type of confidence interval/test that is most appropriate for each hypothesis (i.e. the test for which the necessary assumptions are most justified). For example, you should not perform both a z test and a t test of the same hypothesis, you need to choose between them and justify that choice. However, you are free to use different types of tests/intervals for different hypotheses, so you aren't restricted to only using one type of test/interval throughout the report. You **must** state any assumptions that you make for each of these (e.g. about the underlying distribution or variance) and you **must** state the evidence that supports these assumptions. **Do not** state $n \leq 30$ or $n > 30$ as your only reason for your choice of test/interval (I don't care if the internet tells you that this is enough).

When presenting a confidence interval, you should always present both the interval and the point estimate (of the mean or difference in means). It is often helpful to put them into a caterpillar or forest plot so that they can be compared. Code will be provided that

shows how to produce the plot below in R, and you can adapt this code for your own purposes. Please try to understand the code rather than applying it blindly.

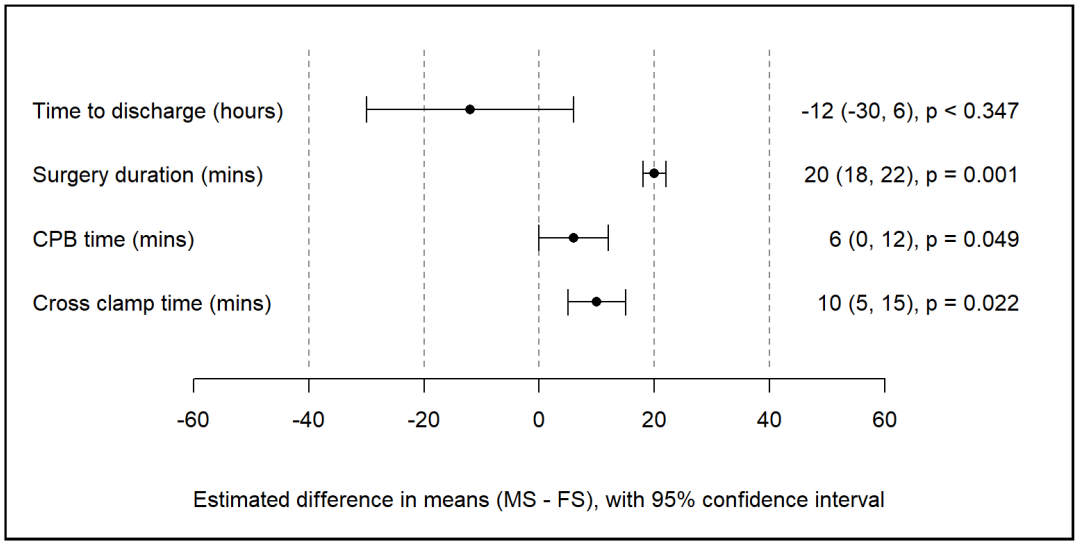


Figure 5.2: A caterpillar/forest plot of four 95%-confidence intervals for four different parameters in a heart surgery trial comparing mini-sternotomy (MS) to full sternotomy (FS) via a two-sample t test.

Whenever you calculate a confidence interval, **interpret it thoroughly** rather than just checking if it contains 27. Is your interval wide or narrow, and why does this matter? Is your interval to the left or to the right of 27, and what does this suggest? Are the values in your interval large or small? Are they close to 27 or far from 27? A confidence interval provides much more information than the result of a single hypothesis test, so don't waste that information.

Discussion

Your discussion/conclusions section should state your main findings in plain English, as clearly and simply as possible. Interpret your findings and their potential real-life implications. Do not just repeat your numerical results. Do not make conclusions that are not supported by your analysis. Do not state your conclusions too strongly (e.g. "our data **prove** that…"). State any limitations of your analysis (e.g. poor quality data, outliers, potentially misleading summaries, or questionable assumptions) and discuss how these might affect your conclusions. Do your findings raise any questions that might be of interest for future research?

Remember, the goal is not to perform a perfect analysis. The goal is to accurately and transparently present the analysis that you have chosen, and to discuss any limitations or potentially flawed assumptions. But also remember to discuss any strengths of your analysis.

Acknowledgements

You can acknowledge anyone who helped.

References

You don't need to include any references in this report, but you might want to. The Leeds Harvard style is often used in dissertations, but here you are free to use any referencing style that works for you. Please try to avoid referring to websites such as Wikipedia because these often change over time and their accuracy is questionable (where possible, peer-reviewed publications such as books and journal articles are preferred). You should include in your references any R packages that you have used (for example, typing `citation("survival")` into R will tell me which paper to cite for the `survival` package).

6 Exercise

This exercise is not part of your coursework. It is here simply so that you can test your own understanding of some of the advice given above. Far too many students either misunderstand that advice or ignore it.

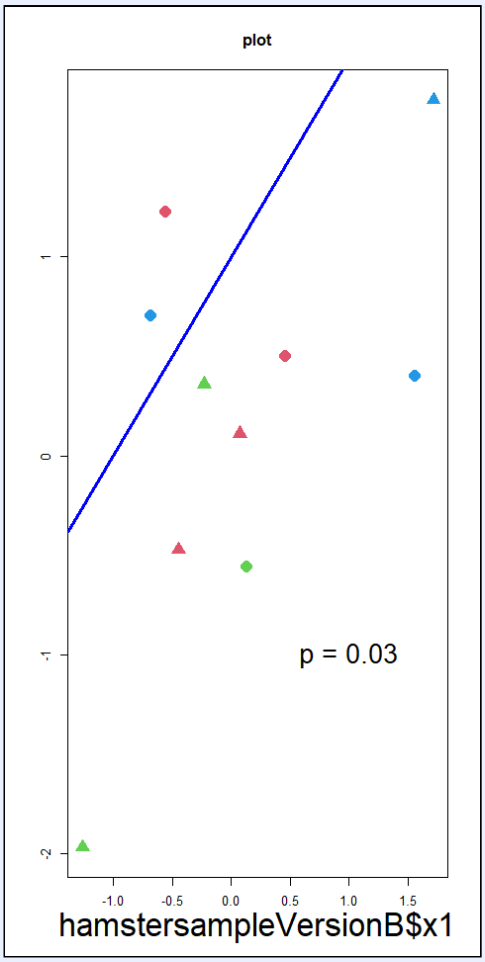
Look at the table, figure and references given below. Try to list all of the flaws that you spot. How could they be improved?

Table

	Hamster	Mouse
mu	3.5555456	1.40

	Hamster	Mouse
m	3.056775767	1.43232333

Figure



Refs

- Gelman A. P Values and Statistical Practice.
- Gelman A. Abandon Statistical Significance.
- https://en.wikipedia.org/wiki/P-value
- https://www.simplypsychology.org/p-value.html [accessed December 21 2022]
- R package “survival” was used.

In general, you should try to look at any table or figure from the point of view of the reader. Would they be clear to readers who have not done the analysis themselves? If not, then they need to be improved.

Similarly, you should look at the references from the point of view of the reader. Based on the information provided in your references, would the reader easily be able to find exactly the resource you are referring to? If not, you need to either provide more information or present the information more clearly.

There is flexibility in how your present graphs or tables. However, ignoring the advice given her will most likely lead to a loss of marks. Please be careful and pay attention.

7 Solution to the exercise

Problems with the table:

- No caption to explain what the table is about. Are these estimates or true parameter values? What are they estimates of (i.e. which parameters of which variables)?
- The table is not numbered, so it cannot easily be referred to in the text (e.g as “Table 2”).
- No units of measurement.
- No explanation of what mu and m denote.
- Assuming that mu and m are the mean and median, we two have measures of location in the table but no measures of spread (usually the standard deviation should be provided with the mean and the IQR should be provided with the median).
- Just typing mu is lazy, and the Greek symbol μ would be better. Even better, just write the word “mean” to remove any ambiguity.
- No sample sizes.
- Too many decimal places.
- Inconsistent numbers of decimal places.

Problems with the figure:

- There is no explanation of what the plot is. Is it a scatter plot of one variable against another? Is it a Q-Q plot? Is it something else entirely? We can't tell from what is given here. A caption should be included to explain the type/purpose of the plot.
- The figure is not numbered, so it cannot easily be referred to in the text (e.g as "Figure 4").
- The horizontal axis label is in the form `data$variable`, which is meaningless to the reader.
- The title and numbers on the ases are too small to read easily.
- `x1` and `hamstersampleVersionB` are not likely to be helpful unless the reader has already worked through all of the analysis code.
- There is no label on the vertical axis.
- There are no units given anywhere.
- The plotted points are different shapes and colours but with no explanation. If colour/shape is communicating important information here then a legend could be used to explain. If colour/shape are not communicating important information, don't use them (they will be an unnecessary distraction to the reader).
- It is not centred.
- It is squashed vertically for no apparent reason.
- There is no explanation of what the line represents.
- There appears to be a p value in the plot, but with no statement that it is a p value, or what test it is from, or what is being tested. This information could go in a caption.

Problem with the references:

- Writing just "Refs" is pretty lazy.
- The first two references here are papers, but it is not stated when they were

written or in which journal they were published.

- There are two papers by Gelman, so the reader will not know which one you are referring to if you just write “(Gelman)”. Include the dates, so that you can refer to them as (Gelman, 2013) and (Gelman, 2019).
- The second paper has more then one author, so their names should be included.
- Stating just the links to websites is lazy. As in any reference, you should state the author (if available), the title, and the date written (if available).
- For websites, always include the date accessed.
- Wikipedia is almost always a bad choice because it can be changed at any time and by anyone. At least include the date it was accessed.
- Don’t just state the packages used. Cite the paper associated with the package. For example, for the `survival` package, this paper can easily be found by typing `citation("survival")` into R. Don’t just write `citation("survival")` in your list of references.
- References should be in a sensible order, e.g. ordered alphabetically by the first author’s surname.

---

1. Jayne D, et al. Effect of Robotic-Assisted vs Conventional Laparoscopic Surgery on Risk of Conversion to Open Laparotomy Among Patients Undergoing Resection for Rectal Cancer (The ROLARR Randomized Clinical Trial). JAMA. 2017;318(16):1569-1580. doi:10.1001/jama.2017.7219 ↩