BCB744 (BioStatistics): Final Integrative Assessment

Smit, A. J.
University of the Western Cape

2025-07-19

Table of contents

Honesty Pledge

This assignment requires that you work as an individual and not share your code, results, or discussion with your peers. Penalties and disciplinary action will apply if you are found cheating.

ℹ Acknowledgement of the Pledge

Copy the statement, below, into your document and replace the underscores with your name acknowledging adherence to the UWC's Honesty Pledge.

I, _____, hereby state that I have not communicated with or gained information in any way from my peers and that all work is my own.

Instructions

Please carefully adhere to the following guidelines. Non-compliance may result in deductions.

- **Convert Quarto to HTML:** Submit your assignment as an HTML file, derived from a Quarto document. Ensure your submission is a *thoroughly annotated report*, complete with meta-information (name, date, purpose, etc.) at the beginning. Each section/test should be accompanied by detailed explanations of its purpose.
- **Testing Assumptions:** For all questions necessitating formal inferential statistics, conduct and document the appropriate preliminary tests to check statistical assumptions. This includes stating the assumptions, detailing the procedures for testing these assumptions, and specifying the null hypotheses (H_0). If assumptions are tested graphically, elucidate the rationale behind the graphical method. Discuss the outcomes of these assumption tests and provide a rationale for the chosen inferential statistical tests (e.g., *t*-test, ANOVA).
- **State Hypotheses:** When inferential statistics are employed, clearly articulate the null (H_0) and alternative (H_A) hypotheses. Later, in the results section, remember to state whether the H_0 or H_A is accepted or rejected.
- **Graphical Support:** Support all descriptive and inferential statistical analyses with appropriate graphical representations of the data.
- **Presentation Format:** Structure each answer as a concise mini-paper, including the sections Introduction, Methods, Results, Discussion, and References. Though each answer is expected to span 2-3 pages, there are no strict page limits. [Does not apply to questions marked with an *]
  - Incorporate a **Preamble** section before the Introduction to detail preliminary analyses, figures, tables, and other relevant background information that doesn't fit into the main narrative of your paper. This section provides insight into the preparatory work and will not be considered part of the main evaluation.
  - The **Introduction** should set the stage by offering background information, establishing the relevance of the study, and clearly stating the research question or hypothesis.
  - The **Methods** section must specify the statistical methodologies applied, including how assumptions were tested and any additional data analyses performed. Emphasise the inferential statistics without delving into exploratory data analysis (EDA).

- In the **Results** section, focus solely on the findings pertinent to the hypotheses introduced in the Introduction. While assumption tests are part of the statistical analysis, they need not be highlighted in this section (unless they necessitated the decision to use a non-parametric test or a data transformation). Ensure that figure and/or table captions are informative and self-explanatory.
- The **Discussion** section is for interpreting the results, considering their significance, limitations, and implications, and suggesting avenues for future research. You may reference up to five pertinent studies in the Methods and Discussion sections.
- End with a consolidated **References** section, listing all sources cited across the questions.

- **Formatting:** Presentation matters. Marks are allocated for the visual quality of the submission. This includes the neatness of the document, proper use of headings, and adherence to coding conventions (e.g., spacing).
- **MARK ALLOCATION** Please see the Introduction Page for an explanation of the assessment approach that will be applied to these questions.

Submit the .html file wherein you provide answers to Questions 1–7 by no later than 21:00, Saturday, 19 July 2025. Label the script as follows:

BCB744_<Name>_<Surname>_BioStats_Exam_rewrite_2025, e.g.

BCB744_AJ_Smit_BioStats_Exam_rewrite_2025.html.

Email your answers to AJ Smit by no later than 21:00 on 19 July 2025.

1 Question 1: Effects of Mercury-Contaminated Fish Consumption on Chromosomes

1.1 Dataset Overview

The dataset `mercuryfish`, available in the R package **coin**, comprises measurements of mercury levels in blood, and proportions of cells exhibiting abnormalities and chromosome aberrations. This data is collected from individuals who consume mercury-contaminated fish and a control group with no such exposure. For detailed attributes and dataset structure, refer to the dataset's documentation within the package.

1.2 Objectives

Your analysis should aim to address the following research questions:

a. **Impact of Methyl-Mercury:** Is the consumption of fish containing methyl-mercury associated with an increased proportion of cellular abnormalities?

b. **Mercury Concentration and Cellular Abnormalities:** How does the concentration of mercury in the blood affect the proportion of cells with abnormalities? Moreover, is there a difference in this relationship between the control group and those exposed to mercury?

c. **Relationship Between Variables:** Does a relationship exist between the proportion of abnormal cells (`abnormal`) and the proportion of cells with chromosome aberrations (`ccells`)? This analysis should be conducted separately for the control and exposed groups to identify any disparities.

2 Question 2: Malignant Glioma Pilot Study

2.1 Dataset Introduction

The `glioma` dataset, found within the **coin** R package, originates from a pilot study focusing on patients with malignant glioma who underwent pretargeted adjuvant radioimmunotherapy using yttrium-90-biotin. This dataset includes variables such as patient sex, treatment group, age, histology (tissue study), and survival time.

2.2 Objectives

This analysis aims to investigate the following aspects:

a. **Sex and Group Interaction on Survival Time:** Determine whether there is an interaction between patient sex and treatment group that significantly impacts the survival time (`time`).

b. **Age and Histology Interaction on Survival Time:** Assess if age and histology interact in a way that influences the survival time of patients.

c. **Comprehensive Data Exploration:** Conduct an exhaustive graphical examination of the dataset to uncover any additional patterns or relationships that merit statistical investigation. Identify the most compelling and insightful observation, formulate a relevant hypothesis, and perform the appropriate statistical analysis.

3 Question 3: Risk factors associated with low infant birth weight

3.1 Dataset Introduction

Package **MASS**, dataset `birthwt`: This dataframe has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass. during 1986.

3.2 Objectives

State three hypotheses and test them. Make sure one of the tests makes use of the 95% confidence interval approach rather than a formal inferential methodology.

4 Question 4: The lung capacity data

4.1 Objectives

a. Using the Lung Capacity data provided, please calculate the 95% CIs for the `LungCap` variable as a function of:

   i. `Gender`

ii. `Smoke`

iii. `Caesarean`

b. Create a graph of the mean ± 95% CIs and determine if there are statistical differences in `LungCap` between the levels of `Gender`, `Smoke`, and `Caesarean`. Do the same using a *t*-test. Are your findings the same using these two approaches?

c. Produce all the associated tests for assumptions – i.e. the assumptions to be met when deciding whether to use a *t*-test or its non-parametric counterpart.

d. Create a combined tidy dataframe (observe tidy principles) with the estimates for the 95% CI for the `LungCap` data (`LungCap` as a function of `Gender`), estimated using both the traditional and bootstrapping approaches. Create a plot comprising two panels (one for the traditional estimates, one for the bootstrapped estimates) of the mean, median, scatter of raw data points, and the upper and lower 95% CI.

e. Undertake a statistical analysis that factors in the effect of `Age` together with one of the categorical variables on `LungCap`. What new insight does this provide?

5 Question 5: Piglet data

5.1 Objectives

Here are some fictitious data for pigs raised on different diets (make up an equally fictitious justification for the data and develop hypotheses around that):

```
feed_1 <- c(60.8, 57.0, 65.0, 58.6, 61.7)
feed_2 <- c(68.7, 67.7, 74.0, 66.3, 69.8)
feed_3 <- c(102.6, 102.1, 100.2, 96.5, 110.3)
feed_4 <- c(87.9, 84.2, 83.1, 85.7, 90.3)

bacon <- data.frame(cbind(feed_1, feed_2, feed_3, feed_4))
```

6 Question 6: Investigating the Impact of Biochar on Crop Growth and Nutritional Value

6.1 Overview of Dataset

In this analysis, we will explore the effects of biochar application on the growth and elemental composition of four key crops: carrot, lettuce, soybean, and sweetcorn. The dataset for this study is sourced from the US Environmental Protection Agency (EPA) and is available at EPA's Biochar Dataset. To gain a comprehensive understanding of the dataset and its implications, it is highly recommended to review two pertinent research papers linked on the dataset page. These papers not only provide valuable background information on the studies conducted but also offer critical insights and methodologies for data analysis that may be beneficial for this project.

6.2 Research Goals

The primary aim of this project is to analyse the impact of biochar on plant yield and identify the three most significant nutrients that influence human health. Your task is to:

1. Determine whether biochar treatments vary in effectiveness across the different crops.
2. Provide evidence-based recommendations on how to tailor biochar application for each specific crop to optimise the production of nutrients beneficial to human health and achieve the best possible yield.

In the Introduction section, it is crucial to justify the selection of the three nutrients you will focus on, explaining their importance to human nutrition. Through detailed data analysis, this project seeks to offer actionable insights on biochar application strategies that enhance both the nutritional value and the biomass of the crops by the end of their growth period.

7 Question 7*

7.1 Objectives

a. For each line of the script, below, write an English explanation for what the code does.

```
ggplot(points, aes(x = group, y = count)) +
  geom_boxplot(aes(colour = group), size = 1, outlier.colour = NA) +
  geom_point(position = position_jitter(width = 0.2), alpha = 0.3) +
  facet_grid(group ~ ., scales = "free") +
    labs(x = "", y = "Number of data points") +
  theme(legend.position = "none",
    strip.background = element_blank(),
    strip.text = element_blank())
```

b. Using the `rnorm()` function, generate some fictitious data that can be plotted using the code, above. Make sure to assemble these data into a dataframe suitable for plotting, complete with correct column titles.

c. Apply the code *exactly as stated* to the data to demonstate your understanding of the code and convince the examiner of your understanding of the correct data structure.

7.2 The end. Thank you for playing along, and have a happy weekend...

Bibliography