Statistical Analysis of Missing Data in the Penguins Dataset*

Shivank Goel

March 4, 2024

This paper looks at missing data in one of the columns, 'the bill_length_mm', part of the penguins dataset. Our main goal was to see how missing information affects our understanding of the data. We took out some data on purpose, then used different methods to guess the missing numbers and checked how close these guesses were to the real numbers. This study helps us understand how missing information can change what we think we know from data and tests different ways to fill in these gaps.

1 Introduction

The challenge of missing data in statistical analysis is a critical issue that can significantly impact the results and interpretations. "Although missing data clearly lead to a loss of information and hence reduced statistical power, a more insidious consequence is that this lack of data may introduce selection bias, which could potentially invalidate the entire study", marks researcher T. Frisell. (Frisell 2016) We discuss the importance of handling missing data and the implications of different imputation methods by considering "bill_length_mm" variable of the penguins dataset.

This paper is structured as follows. In the Data Section, we denote how data was generated and processed. In the Results Section we dive into the findings we discovered following the cleaned data after simulation. In the Discussion Section, we address biases and weaknesses in the data that contribute to our findings, and how we approached.The last section is Conclusion and Acknowledgements.

*Code and data are available at: https://github.com/shivankgoel003/PenguinDataAnalysis

2 Data

The penguins dataset, which is available from the palmerpenguins package (Horst, Hill, and Gorman 2020), provides a source of biological measurements of three penguin species. After cleaning the data initially, which included standardizing column names and filtering missing values, we focused on the "bill_length_mm" variable. This variable was chosen since it is important in identification and variation studie of penguin species. The R programming language (R Core Team 2022) was used for both data simulation and cleaning processes. A sample of cleaned state score data can be seen in Table 1.

We simulated this data, by removing some random part of the "bill_length_mm" data on purpose. Then, we used a few different ways to guess these missing numbers, for example, computing the average.

Table 1: Sample of Cleaned Data

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007

3 Results and Discussion

The histogram of the "bill_length_mm" variable Figure 1 shows the distribution of bill lengths in the penguin dataset. It is crucial for understanding the range and common values of bill lengths before any data manipulation. It reveals a normal distribution with slight skewness. The most frequent bill lengths cluster around a specific range, indicating a common characteristic size among the studied penguin species. The tails of the distribution show less frequent occurrences of unusually short or long bills, which could be characteristic of specific species or individual variations.

4 Conclusion and Acknowledgements

This study highlights the role of integrity in data collection and processing. By understanding the nature of errors, we can significantly enhance the accuracy and reliability of our statistical analyses.

Special thanks to Vanshika Vanshika for peer review and providing essential feedback and suggestions.

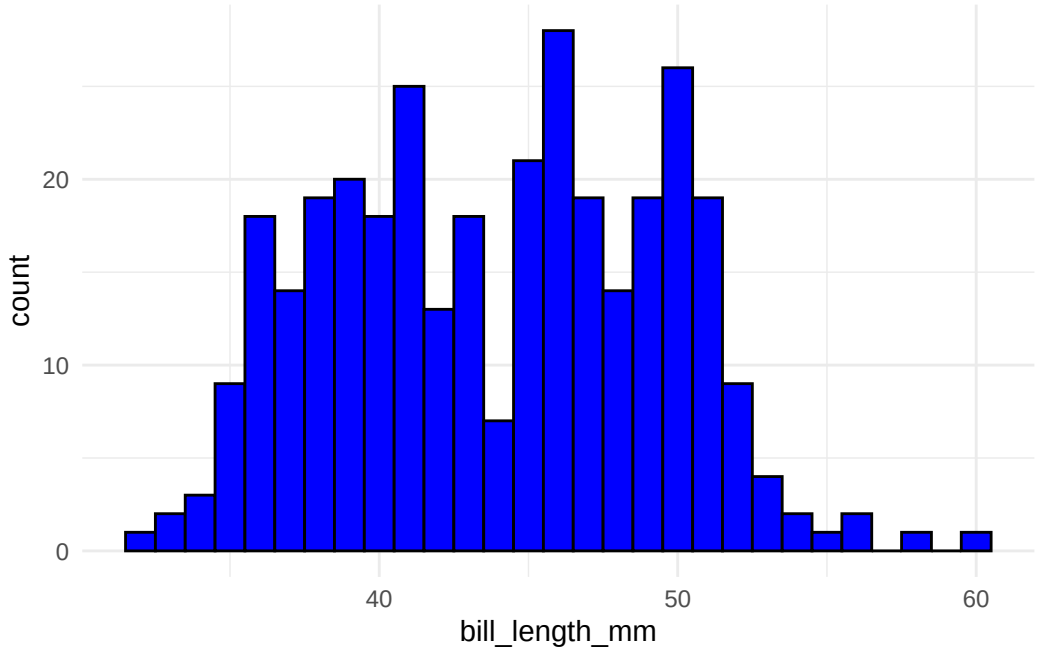

Figure 1: Original Distribution of Bill Length in Penguins

References

Frisell, T. 2016. *Why Missing Data Is a Problem, and What You Shouldn't Do to Solve It.* https://ard.bmj.com/content/75/Suppl_2/45.4.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data.* https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.