Project Proposal

Penguin Measurement Analyses and Prediction using Machine Learning

Gurdeep Panag (30101520)

Harjot Dhaliwal (30051859)

Lukas Escoda (30211208)

Shabbir Khandwala (30219011)

University of Calgary

DATA 606: Statistical Methods in Data Science

Instructor: Dr. Greenberg

Feb 2, 2024

Introduction

In the vast and dynamic ecosystem of Antarctica, penguins emerge not only as key species but also as fascinating subjects of scientific study. Understanding their life patterns, physical adaptations, and population dynamics is essential to unraveling the mysteries of these resilient beings and their interactions with the environment. This report focuses on exhaustive statistical analysis of a detailed dataset of penguins, encompassing multiple species and diverse physical and biological measurements.

Through advanced statistical techniques, including various types of sampling, statistical techniques and logistic regression, we aim to unravel the complex interactions between the physical characteristics of the penguins, their environmental context, and species. This research not only seeks to contribute to the existing academic body on Antarctic wildlife but also to highlight the importance of applying rigorous statistical methods to better understand the animal. With meticulous analysis, our goal is to provide insights that are fundamental for the conservation and management of this animal, emphasizing how data science and statistical techniques can be powerful tools in understanding biodiversity and various animal species. This report is a step towards the profound exploration of the secrets that penguins have guarded for millennia, opening new avenues for research and appreciation of these emblematic creatures.

Through some exploratory research, we have identified that the only categorical factors that would likely influence the physical and biological measurements of penguins would be the species (Animal Corner, 2023) and sex of the penguin (Horvath A., 2024). Through this research we have identified the best sampling method to choose would be stratified sampling based on the species of penguin, as we believe this has the largest impact on the measurements.

Guiding Questions

Some of the crucial guiding questions we identified for our analyses:

1. How well does a sample chosen using stratified sampling on penguin species approximate the population mean?
2. Using our sample from stratified sampling, which 2 measurement variables are most correlated with each other, and can then be used to predict each other using ratio estimation?
3. Does the average penguin in the population have a height of 3 feet, according to our stratified sample, using bootstrapping?
4. Is it possible to predict the species of a penguin based on its physical and biological measurements using logistic regression?

Dataset

This dataset is from Kaggle and was sourced from the original public Github page (Amy, 2024). It can be used according to the MIT license which means we have the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the dataset without limitation (Amy, 2024).

Column Name	**Description**
Species	Species of the penguin
Island	Island the penguin belongs to
Culmen_length_mm	Length of the penguin
Culmen_depth_mm	Depth of the penguin
Flipper_length_mm	Flipper length of the penguin
Body_mass_g	Body mass of the penguin
Sex	Sex of the penguin

Table 1: Column Descriptions

Objective

Our first objective would be to check to see if stratified sampling based on penguin species is a good predictor of the actual population means of the flipper and height measurements of penguins. Our second objective in this project is to see if any physical and biological measurements of penguins are highly correlated with one another, and if so, are we able to use ratio estimation to obtain estimates of the other variable, in order to simplify the data gathering process for future measurements. Our third objective is to conduct a t-test of penguin heights to see if penguin heights on average are different than 3 meters, using the stratified sample from earlier, combined with bootstrapping in order to ensure the presence of a small sample is not hindering us from testing multiple samples. Our last objective would be to create an accurate Logistic Regression algorithm (supervised machine learning) in order to predict the species of a penguin given various physical and biological measurements.

Methodology

Our analytical methodology revolves around the utilization of Python in a Jupyter notebook environment. To facilitate our analysis, we make use of essential libraries such as NumPy, Pandas, Seaborn, SciPy, and Scikit-learn. We commence our analysis with exploratory data analysis, aiming to uncover valuable insights from the penguin dataset and present them through compelling visualizations. Following the initial exploration, we proceed with meticulous data pre-processing. This stage involves handling null values, formatting strings, and transforming categorical variables, such as sex, into binary representations. By ensuring the data is clean and properly formatted, we lay the foundation for accurate analysis.

The sampling phase plays a crucial role in our analysis. Here, we compare stratified sampling sample mean to the actual population means to see how well stratified sampling is able to estimate the population means for specific measurements. This allows us to gain a deeper understanding of the penguin population and its characteristics. Once robust correlations between variables are identified, we expand our exploration by implementing a ratio estimator. This estimator enables us to predict the values of different numerical features, and evaluate the accuracy and feasibility of our predictions.

In the next phase of our analysis, we plan on conducting t-tests with the bootstrapping method using the stratified sample. We would be testing the hypothesis that average penguin heights are about 3 meters tall, as we had a rigorous debate on what the average height of a penguin is in actuality.

The culmination of our analysis involves the development of a logistic regression model using Scikit-learn for supervised machine learning. This model is designed to predict the species of a penguin based on its features. We rigorously assess the accuracy of the model to ensure its reliability and effectiveness. It is important to emphasize that our methodology is adaptive and flexible. We are open to making adjustments based on emerging insights and analysis outcomes. By following this structured approach, we anticipate gaining valuable insights into the various characteristics of penguins.

References

Amy. (2024). *Penguin sizes dataset*. Kaggle. https://www.kaggle.com/datasets/amulyas/penguin-size-dataset

Animal Corner. (2023). *Penguin Size Comparison-How Big are Penguins?* Animal Corner. https://animalcorner.org/blog/penguin-size-comparison/

Horvath A. (2024). How do Penguins find their mate in a sea of tuxedos? The University of Melbourne. https://pursuit.unimelb.edu.au/articles/how-do-penguins-find-their-mate-in-a-sea-of-tuxedos