Effet of Missing Data in Analysis*

Qi Er (Emma) Teng

March 5, 2024

This essay studies the impact of Missing data on statistical inference, employing simulations to investigate how different data handling strategies affect analysis outcomes. In the different types of missingness, we will focus on Missing Completely at Random, and evaluate the imputed value with the actual data. Our results will underscore the significant influence of missing data. This essay concludes methods for managing MCAR data, emphasizing the importance of method selection to ensure reliable statistical analysis in the presence of incomplete data.

Table of contents

*Code is available at https://github.com/dwz92/Effect-of-Missing-Data-in-Analysis

Introduction

In this study, we delve into the effects of Missing data within the context of statistical analysis, which is a common challenge in research that can skew results if not properly addressed. We will focus on studying Missing Completely at Random (MCAR) data for our investigation. By using the `palmerpenguins` dataset (Horst, Hill, and Gorman 2020) as a foundation for simulation, we will artificially introducing MCAR conditions into the `penguins` dataset. We aim to explore the resultant biases and explore data handling techniques. This controlled examination seeks to illuminate the impact of MCAR data on common analytical processes, providing insights into the potential distortions in statistical conclusions. Through comparative analysis, this essay offers evidence-based recommendations for addressing MCAR data, enhance understanding of MCAR's implications and aid in the development of robust analytical practices for dealing with incomplete datasets.

Simulating Missing Data

In this section, we will simulate MCAR data to create a dataset of intended misses using R (R Core Team 2020), the `palmerpenguins` dataset (Horst, Hill, and Gorman 2020) and R libraries of dplyr (Wickham et al. 2023), and tidyverse (Wickham et al. 2019).

We will start with cleaning the original dataset.

```
penguin_bill_len <- penguins

glimpse(penguin_bill_len)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
$ sex               <fct> male, female, female, NA, female, male, female, male~
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

```
penguin_bill_len <- penguin_bill_len |>
  select("island", "bill_length_mm")
```

```r
head(penguin_bill_len)
```

```
# A tibble: 6 x 2
  island    bill_length_mm
  <fct>              <dbl>
1 Torgersen           39.1
2 Torgersen           39.5
3 Torgersen           40.3
4 Torgersen           NA
5 Torgersen           36.7
6 Torgersen           39.3
```

We are now interested in the key summary statistics of the cleaned dataset.

```r
summary(penguin_bill_len)
```

```
       island    bill_length_mm
 Biscoe   :168   Min.   :32.10
 Dream    :124   1st Qu.:39.23
 Torgersen: 52   Median :44.45
                 Mean   :43.92
                 3rd Qu.:48.50
                 Max.   :59.60
                 NA's   :2
```

We will now create several dataset that are missing islands completely at random. However, we will ensure rows with null `bill_length_mm` are removed so mean can be calculated successfully.

```r
sample_means <- tibble(seed = c(), mean = c(), islands_ignored = c())

unique(penguin_bill_len$island)
```

```
[1] Torgersen Biscoe    Dream
Levels: Biscoe Dream Torgersen
```

```r
for (i in c(3:5)) {
  set.seed(i)
```

```
  dont_get <- c(sample(x = unique(penguin_bill_len$island), size = 1))
  sample_means <-
    sample_means |>
    rbind(tibble(
      seed = i,
      mean =
        penguin_bill_len |>
          filter(!island %in% dont_get) |>
          summarise(mean = mean(bill_length_mm, na.rm = TRUE)) |>
          pull(),
      islands_ignored = str_c(dont_get, collapse = ", ")
    ))
}

sample_means |>
  kable(
    col.names = c("Seed", "Mean", "Ignored Islands"),
    digits = 2,
    format.args = list(big.mark = ","),
    booktabs = TRUE
  )
```

Seed	Mean	Ignored Islands
3	44.79	Torgersen
4	43.78	Dream
5	42.65	Biscoe

Understanding Missing Data

From the previous section, it is intuitive to notice the lack of significant difference in imputed mean compared to actual mean. This occurrence is normal considering we are simulating a Missing Completely at Random (MCAR) data. Per the definition of MCAR, the observations of a dataset are missing independently of any other variables. Therefore, it has least impact on summary statistics and inference within the three types of Missing data.

Another factor to consider for the lack of difference in comparative analysis is that `island` has little influence on `bill_length_mm`. In the context of our dataset, geographical value has little influence on the bill length of the penguins.

Handling Missing Data

A straightforward way to handle MCAR data is through imputation methods. Due to the nature of MCAR, the missing data can be replaced with the mean, median, or mode of the observed values of that variable. Although this method is straightforward, it can underestimate variability and affect the distribution of the data.

Thus, we can implement multiple imputation to handle these concerns. By generating several imputed datasets separately and combine the results through averaging, we are able to implement multiple imputation. MOreove, multiple imputation accounts for the uncertainty of the imputed values, as it can provide more accurate standard errors and confidence intervals.

Conclusion

Our study on the Missing Completely at Random (MCAR) data within the `palmerpenguins` dataset underscores that MCAR has a negligible effect on statistical analyses, maintaining the validity of results even with data missing at random. The efficiency of imputation methods, especially multiple imputation, is a robust solution for handling MCAR, allowing for accurate and reliable research outcomes. This investigation reinforces the importance of appropriate data handling techniques to counteract the challenges posed by incomplete datasets in statistical research.

Appendix

Feedback

Yuanyi (Leo) Liu provided the following feedback for the initial draft of this paper: "MAR data are missing at random, which means they are missing from the dataset in a way that is related to other variables in the dataset. Your simulations follow MCAR, which randomly misses from the dataset."

References

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. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.