

Missing Data from Palmerpenguins*

Kyungrok Park

2024-03-04

Palmerpenguins Data

```
penguins_bill_length |>
  summary()
```

| | species | bill_length_mm |
|------------|---------|----------------|
| Adelie | :152 | Min. :32.10 |
| Chinstrap: | 68 | 1st Qu.:39.23 |
| Gentoo | :124 | Median :44.45 |
| | | Mean :43.92 |
| | | 3rd Qu.:48.50 |
| | | Max. :59.60 |
| | | NA's :2 |

The above summary statistic indicates that there are 3 different species of palmerpenguins (Adelie, Chinstrap, Gentoo) , and their mean bill length as 43.92 mm. However, for the bill length (mm), I observed that there are two missing values from the original data set.

Missing Data from Palmerpenguins

```
penguins_bill_length <- penguins_bill_length[complete.cases
  (penguins_bill_length), ]
```

To explore the imputation for missing values, we first drop two NA values mentioned above from the original data set.

*Code and data are available at: <https://github.com/KyungrokP/Missing-data-exercise.git>

MCAR

```
set.seed(213)
penguins_mcar <-
  penguins_bill_length |>
    mutate(bill_length_mm = replace(bill_length_mm, sample(row_number(), size = 3,
                                                             replace = FALSE), NA_real_))
penguins_mcar |>
  summary()
```

| species | bill_length_mm |
|---------------|----------------|
| Adelie :151 | Min. :32.1 |
| Chinstrap: 68 | 1st Qu.:39.2 |
| Gentoo :123 | Median :44.4 |
| | Mean :43.9 |
| | 3rd Qu.:48.5 |
| | Max. :59.6 |
| | NA's :3 |

Out of three cases for missing values, I chose to simulate for “Missing Completely At Random (MCAR)” case. I randomly selected 3 indexes from the data set, and make the bill length (mm) of those 3 selected indexes as NA values.

Imputation

```
set.seed(232)
multiple_imputation <-
  mice(
    penguins_mcar,
    print = FALSE
  )
mice_estimates <-
  complete(multiple_imputation) |>
  as_tibble()
```

By using mice() function, we can do the multiple imputation.

Table 1: Comparing the imputed values of bill length(mm) for three missing penguins and the mean of each species

| Index | Species | Input Mean | Multiple Inputation | Actual Value |
|---------|-----------|------------|---------------------|--------------|
| 225 | Gentoo | 38.79139 | 44.90 | 46.50 |
| 272 | Gentoo | 47.48926 | 44.90 | 50.40 |
| 329 | Chinstrap | 48.92836 | 46.40 | 42.50 |
| Overall | | 43.91000 | 43.91 | 43.92 |

For the mean replacement, I calculated the mean of bill length for each species and replaced missing values with those different mean length based on species.

As Table 1 shows, both of multiple imputation and mean replacement do not show notable difference from the actual mean of bill length (mm).

Instead, we can observe quite differences between each value imputed using multiple imputation and each value imputed using mean imputation compared to the real values. This fact implies that if the number of missing values is not three but rather one-third of the data, it can lead to significant discrepancies in the overall mean value observed.

References

- Arel-Bundock, Vincent. 2022. “modelssummary: Data and Model Summaries in R.” *Journal of Statistical Software* 103 (1): 1–23. <https://doi.org/10.18637/jss.v103.i01>.
- 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. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. “mice: Multivariate Imputation by Chained Equations in r.” *Journal of Statistical Software* 45 (3): 1–67. <https://doi.org/10.18637/jss.v045.i03>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://yihui.org/knitr/>.