An Exercise in the Missing Data at Random methodology conducted on penguin bill length (in mm)*

Owais Zahid

March 5, 2024

In this paper, we examine the effect of removing data at random with respect to the bill length variable in the penguins dataset, on the mean bill length value. We then draw a comparison to see whether this imputed value is a good estimate for the mean.

1 Loading and previewing the data

In this step we load the necessary packages, (Mice (van Buuren and Groothuis-Oudshoorn 2011), Tidyverse (Wickham et al. 2019) and Palmerpenguins (Horst, Hill, and Gorman 2020)) and obtain a preview of the penguins dataset to see what kind of data we are working with. Additionally as a quick disclaimer, we are going to be using the R programming language (R Core Team 2023), as well as RStudio (Posit team 2023) for the remainder of the paper.

Table 1. Penguins Table Header

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           39.1          18.7               181        3750
2 Adelie  Torgersen           39.5          17.4               186        3800
3 Adelie  Torgersen           40.3          18                 195        3250
4 Adelie  Torgersen           NA            NA                  NA          NA
5 Adelie  Torgersen           36.7          19.3               193        3450
```

*Code and data are available at: https://github.com/FFFiend/penguins_bill_length

```
6 Adelie  Torgersen           39.3          20.6               190        3650
# i 2 more variables: sex <fct>, year <int>
```

We observe that there are 6 columns in total, and our variable of interest is the bill_length_mm column for the Missing At Random (MAR) methodology.

2 Constructing MAR table

In order to remove entries with respect to the bill length, we replace all entries with bill length between 34 and 40 mm (inclusive) with NA, and construct the following table. The range 34-40 inclusive seemed suitable as most bill length values seemed to be in or around that range. We then obtain the following table.

Table 2. New penguins table with bill length between 34 and 40 MAR

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           NA            18.7                181        3750
2 Adelie  Torgersen           NA            17.4                186        3800
3 Adelie  Torgersen           40.3          18                  195        3250
4 Adelie  Torgersen           NA            NA                   NA          NA
5 Adelie  Torgersen           NA            19.3                193        3450
6 Adelie  Torgersen           NA            20.6                190        3650
# i 2 more variables: sex <fct>, year <int>
```

3 Imputation and Comparison

Next, we calculate the mean bill length from the original dataset, as well as the mean bill length for the penguins_MAR dataset after which we perform multiple imputation on the former table, to finally obtain the table below.

Table 3. Comparing the imputed value of bill length with the original mean bill length

Method	Value
Drop missing	99.00000
Input mean	46.49224

Method	Value
Multiple imputation	44.83866
Actual	43.92193

Here we see that the difference between the imputed mean and actual mean is less than the difference between the input mean (i.e the mean bill length from Table 2) and the original. We also note that a total of 99 entries were removed (set to NA) in the process which led to a higher mean value.

Hence, we can conclude that the imputed mean is a good estimate of the true mean bill length for the penguins 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.

Posit team. 2023. *RStudio: Integrated Development Environment for r*. Boston, MA: Posit Software, PBC. http://www.posit.co/.

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 Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.