reproducible_figures_analysis

2023-12-02

Reproducible Figures Assignment

Abstract

This is an RMarkdown file to answer the questions in the coding assignment 'Reproducible Science and Figures in R'. I will be using the Palmer Penguin data set to first create a badly communicated plot and explain how it misleads the reader. I will then create a pipeline to analyse a specific part of the data set. Finally, I will upload my work to GitHub so that another person can check how reproducible it is.

Install packages if required

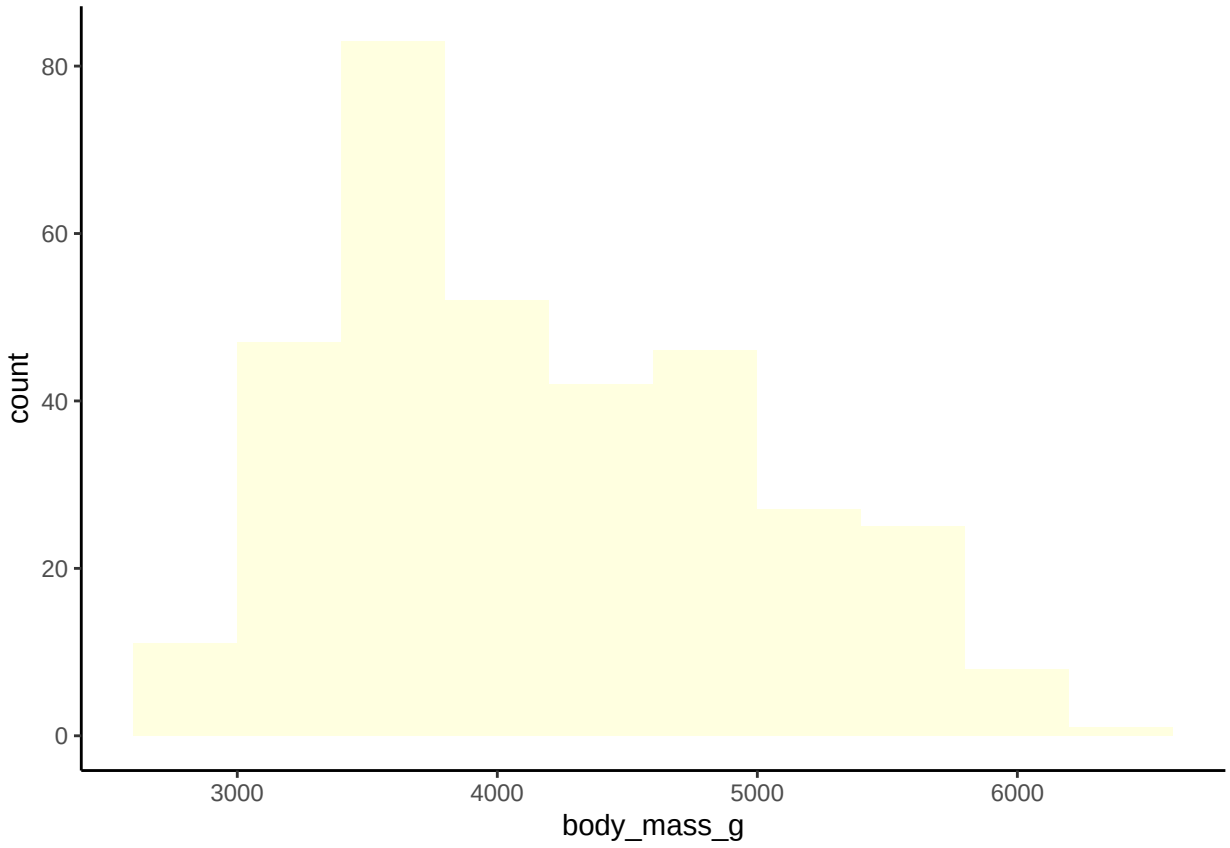
```
#if required install ggplot for graphing
if(!require("ggplot2", character.only = TRUE)) {
  install.packages("ggplot2")
}
#if required install the palmer penguins dataset
if(!require("palmerpenguins", character.only = TRUE)) {
  install.packages("palmerpenguins")
}
#if required install janitor for data cleaning
if(!require("janitor", character.only = TRUE)) {
  install.packages("janitor")
}
#if required install dplyr for piping
 if(!require("dplyr", character.only = TRUE)) {
  install.packages("dplyr")
 }
#if required install ragg for graphic devices
if(!require("ragg", character.only = TRUE)) {
  install.packages("ragg")
}
#if required install readr for reading delimitated files
if(!require("readr", character.only = TRUE)) {
  install.packages("readr")
}
#if required install tinytex to knit file to pdf
if(!require("tinytex", character.only = TRUE)) {
  install.packages("tinytex")
}
```

Load the packages

```
library(ggplot2)
library(palmerpenguins)
library(janitor)
library(dplyr)
library(readr)
library(ragg)
library(tinytex)
```

QUESTION 01: Data Visualisation for Science Communication

a) Create a figure using the Palmer Penguin dataset that is correct but badly communicates the data. Provide your figure here:



b) Write about how your design choices mislead the reader about the underlying data (200-300 words).

Although this figure is correct i.e., I have used a histogram to portray a sampling distribution with a continuous independent variable, it badly communicates the data in a number of ways:

(1) The lack of colours to denote species or sex may be mislead because it gives no information as to how the distribution of penguin body mass is affected by confounding variables. This is necessary so that

the reader does not assume mean body mass is ~3500g is uniform regardless of species/sex i.e., males and Adelie penguins make up the higher body masses! If this is not included, it can lead to incorrect interpretation of findings.

(2) The low number of bins i.e., the number of groups into which the data are divided, may also be misleading the reader about the underlying skew of the histogram. This is because the variability within the distribution is hidden by grouping/compressing lots of data together. Therefore, the number of bins should be increased (from 10 to 30) to more clearly show the shape of the distribution.

(3) The axis labels are misleading because they are vague, especially the Y axis 'counts'. Therefore the axes should be renamed to 'Body Mass (g)' and 'Number of penguins' to clearly tell the reader that I am using data on penguins, looking at the distribution of the number of individuals (Y axis) at different body masses in grams (X axis).

(4) Some minor things which may mislead the reader are a lack of title on this plot to communicate what it is about, the pale yellow bars are difficult to distinguish from the white background and the lack of a grid in the background theme which may cause values to be misread especially for bars far away from the Y axis.

References

- How to interpret histograms
- Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter
- Achieving graphical excellence: Suggestions and methods for creating high-quality visual displays of experimental data
- 1,500 scientists lift the lid on reproducibility
- Point of View: How open science helps researchers succeed

these link can be accessed by clicking on them

QUESTION 2: Data Pipeline

Write a data analysis pipeline in your .rmd RMarkdown file. You should be aiming to write a clear explanation of the steps, the figures visible, as well as clear code.

You will be marked on the following:

a) Your code for readability and functionality

b) Your figures for communication

c) Your text communication of your analysis

---

1. Introduction

I am going to use the Palmer Penguins data set to investigate if there is a correlation between culmen depth and body mass of Chinstrap penguins.

- First I will need to load and clean the data, before filtering is to just the variables I am interested in.

- I will them create an exploratory scatter plot to visualize the Chinstrap penguin data.
- After this I will use a statistical test and results figure to form a conclusion about my hypothesis and discuss the biological implications of this result.

NB// All the functions used in my code can be found in separate folders which can be accessed from my GitHub link in q3.

1.1 Load the data

```
#create a csv of raw data to save a copy
write.csv(penguins_raw, "data/penguins_raw.csv")

#view the first 6 rows of the raw data using the function head()
head(penguins_raw)

#view the column names of the raw data using the function names()
names(penguins_raw)
```

1.2 Clean the data

```
#load the cleaning functions
source("functions/cleaning_function.R")

#using pipes to clean and filter the data using script 'cleaning_functions.R' in the folder 'functions'
penguins_clean <- penguins_raw %>%
  #shorten species names
  shorten_species() %>%
  #make column names all lower case/snake case
  clean_column_names() %>%
  #remove columns starting with delta
  remove_delta_columns() %>%
  #remove empty columns or rows
  remove_empty_columns_rows()

#create a csv of clean data to save a copy
write.csv(penguins_clean, "data/penguins_clean.csv")

#view the new column names of clean data to check it has worked
names(penguins_clean)
```

```
##  [1] "study_name"        "sample_number"     "species"
##  [4] "region"            "island"            "stage"
##  [7] "individual_id"     "clutch_completion" "date_egg"
## [10] "culmen_length_mm"  "culmen_depth_mm"   "flipper_length_mm"
## [13] "body_mass_g"       "sex"               "comments"
```

2. Hypothesis

2.1 Null Hypothesis

- There is no correlation between culmen depth (mm) and body mass (g) of Chinstrap penguins (r=0)

2.2 Alternative Hypothesis

- There is a positive correlation between culmen depth (mm) and body mass (g) of Chinstrap penguins (r>0)

3. Exploratory figure

3.1 Filter the data

```
#load the cleaning functions file from the functions folder
source("functions/cleaning_function.R")

#filter the data to just focus on culmen depth and body mass of Chinstrap penguins
penguins_filtered <- penguins_clean %>%
  filter_by_species("Chinstrap") %>%
  subset_columns(c("body_mass_g", "culmen_depth_mm")) %>%
  remove_NA()

#create a csv of filtered data to save a copy
write.csv(penguins_filtered, "data/penguins_filtered.csv")

#view the new column names of filtered data to check it has worked
names(penguins_filtered)
```

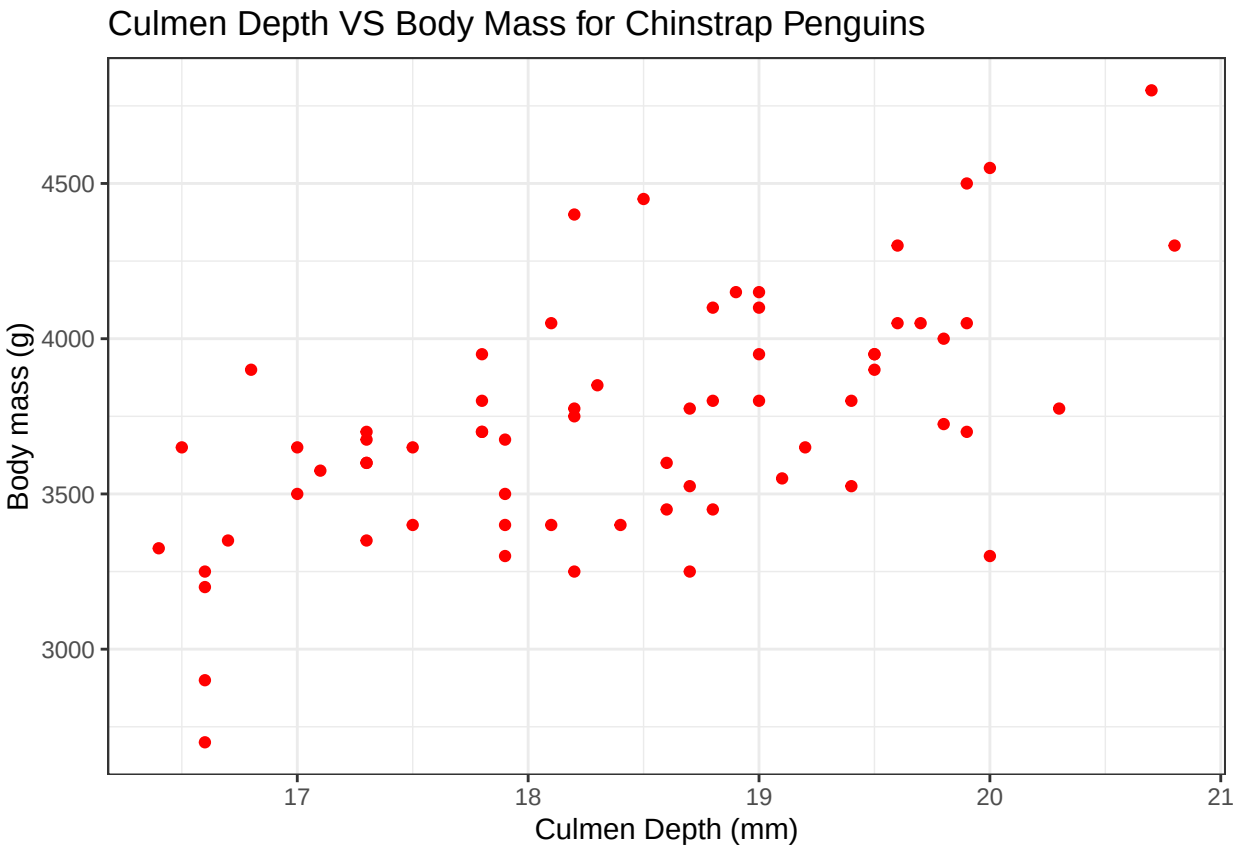
```
## [1] "body_mass_g"     "culmen_depth_mm"
```

3.2 Create an exploratory plot

```
#load the potting function file from the functions folder
source("functions/plotting_function.R")

#create an exploratory scatter plot figure
culmendepth_bodymass_exploratory <- penguins_filtered %>%
  plot_exp_fig()

#print the exploratory figure
culmendepth_bodymass_exploratory
```



This exploratory plot suggests that there might be some positive relationship between body mass and culmen depth but it is hard to be sure as there is no regression line.

3.3 Save the exploratory plot

```
#use ragg package to save a .png file of exploratory plot in 'figures' folder
#RUN WHOLE CHUNK
agg_png("figures/culmendepth_bodymass_exploratory.png",
        width = 22,
        height = 16,
        units = "cm",
        res = 300,
        scaling = 1)
culmendepth_bodymass_exploratory
dev.off()
```

```
## pdf
##   2
```

4. Statistical Methods

4.1 Testing the assumptions for a Pearsons Correlation Coefficient

- Before proceeding with the statistical test, we must first check that the following assumptions have been met:

  (1) Linearity: we can see that this assumption is met by looking at the exploratory scatter plot above.

  (2) Normally distributed: to test this we can use a Shapiro-Wilk normality test as shown in the code below. If the p-values are greater than 0.05 we can assume that the data for each variable does not differ significantly from a normal distribution.

```
#test if culmen depth is normally distributed
shapiro.test(penguins_filtered$culmen_depth_mm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  penguins_filtered$culmen_depth_mm
## W = 0.97274, p-value = 0.1418
```

```
#test if body mass is normally distributed
shapiro.test(penguins_filtered$body_mass_g)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  penguins_filtered$body_mass_g
## W = 0.98449, p-value = 0.5605
```

- The p-value for culmen depth is 0.1418 which is greater than 0.05; therefore we can assume the data is not significantly different to a normal distribution.
- The p-value for body mass is 0.5605 which is greater than 0.05; therefore we can assume the data is not significantly different to a normal distribution.
- The assumptions of linearity and normal distribution have been met, therefore, we can proceed with the statistical test.

4.2 Pearsons Correlation Coefficient

```
#calculate the correlation coefficient between culmen depth and body mass of Chinstrap penguins
chinstrap_cor <- cor.test(penguins_filtered$culmen_depth_mm, penguins_filtered$body_mass_g)
chinstrap_cor
```

```
##
##  Pearson's product-moment correlation
##
## data:  penguins_filtered$culmen_depth_mm and penguins_filtered$body_mass_g
## t = 6.1649, df = 66, p-value = 4.795e-08
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.4277193 0.7367390
## sample estimates:
##       cor
## 0.6044983
```

- This output of this statistical test tells us several things:

  (1) t = 6.1649 tells us that the correlation between culmen depth and body mass is fairly strong and in a positive direction

  (2) df = 66 tells us the number of degrees of freedom

  (3) p-value = 4.795e-08 tells us the probability of finding the current result if the null hypothesis were true (r=0)

  (4) 95% CI = [0.4277193, 0.7367390] tells us the confidence interval i.e., we can be 95% confidence that the true correlation coefficient between culmen depth and body mass lies between 0.428 and 0.737.

  (5) sample estimate: cor = 0.6044 tells us the value of the correlation coefficient. If it is between 0.5 and 1, there is a strong positive correlation.

5. Results & Discussion

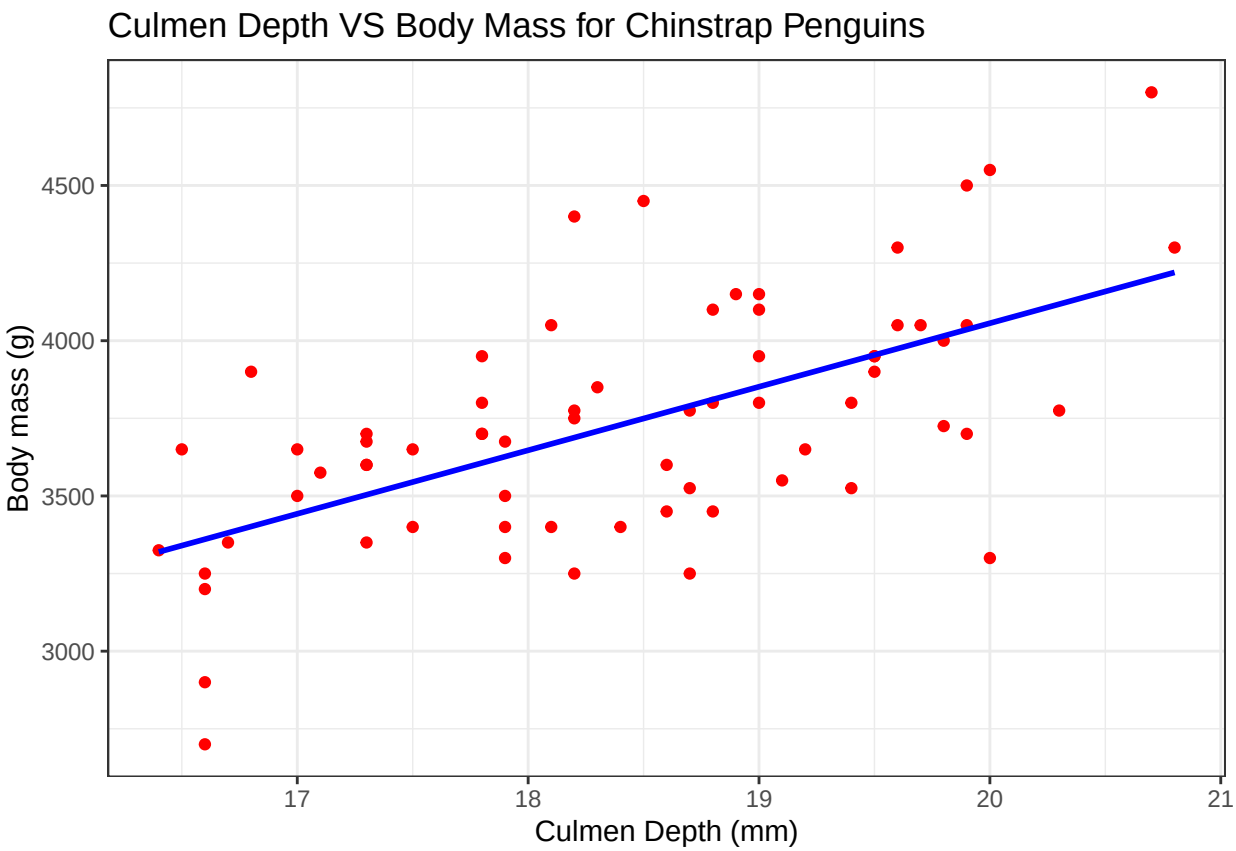
5.1 Plot the results figure

- The correlation coefficient (r) was found to be 0.604 (3 d.p.) which is a fairly strong positive correlation and this can be shown by adding a regression line (with a gradient representative of 0.604) to the scatter plot.

```
#load the plotting function file from the functions folder
source("functions/plotting_function.R")

#create a scatter plot result figure
culmendepth_bodymass_results <- penguins_filtered %>%
  plot_res_fig()

#print the results figure
culmendepth_bodymass_results
```



5.2 Save the results plot

```
#using the ragg package to save a .png file of my 'culmendepth_bodymass_results' in 'figures' folder (R
agg_png("figures/culmendepth_bodymass_results.png",
        width = 22,
        height = 16,
        units = "cm",
        res = 300,
        scaling = 1)
culmendepth_bodymass_results
dev.off()
```

```
## pdf
##   2
```

5.3 Discussion

- Since my value of r(0.604) is greater than 0 and my p-value(4.795e-08) is less than 0.05, I can reject my null hypothesis at the 95% confidence level which indicates that the observed correlation is statistically significant.
- Therefore, I can conclude that there is a fairly strong positive correlation between culmen depth and body mass in Chinstrap penguins.

- There are a number of reason by culmen depth and body mass may be positively correlated Chinstrap penguins.

  (1) It could be because of genetics i.e., individuals with genes that code for bigger culmen depth have an adaptive advantage because they can catch bigger fish which would increase their body mass and enhance their survival and reproduction by staying warm themselves and keeping their egg warm.

  (2) It could be related to age i.e., older individuals may have bigger culmen depth and may also reach higher body masses because of their experience at foraging. I did not include age in my study but this may be something to incorporate in the future.

- Other limitations of the analysis include only looking at one species of penguin and not controlling for sex. Therefore, the analysis could be repeated looking at Gentoo and Adelie penguins and filtering the data for males and females to see if the same positive correlation is found.

6. Conclusion

- This analysis has shown a statistically significant strong positive correlation between culmen depth and body mass in Chinstrap penguins.

QUESTION 3: Open Science

a) GitHub

*Upload your RProject you created for **Question 2** and any files and subfolders used to GitHub. Do not include any identifiers such as your name. Make sure your GitHub repo is public.*

GitHub link: [https://github.com/otter456/reproducible_figures_assignment]

You will be marked on your repo organisation and readability.

b) Share your repo with a partner, download, and try to run their data pipeline.

Partner's GitHub link: [https://github.com/ReproducibleScienceAssessment/Reproducible_figures]

*You **must** provide this so I can verify there is no plagiarism between you and your partner.*

c) Reflect on your experience running their code. (300-500 words)

- *What elements of your partner's code helped you to understand their data pipeline?*

The hash-tagged comments within their chunks of code were helpful to help me understand exactly what they were doing and why. I also thought that the way in which they split up their chucks of code was clear.

- *Did it run? Did you need to fix anything?*

The only thing I needed to fix when running my partners code was getting it to knit to a pdf. To do this I hash-tagged the install.packages() lines of code as I already had these packages installed in my Rstudio. My partner also had two chunks of code both called 'Statistics' so I needed to change these to 'Statisics1' and 'Statistics2' for it to knit. The outputs of my partners code were fine and I was able to see all the plots.

- *What suggestions would you make for improving their code to make it more understandable or reproducible, and why?*

To make the code more reproducible, it would be helpful to have separate folders for cleaning, plotting etc. and in these folders R scripts could be included each containing the functions my partner used. This is helpful because each function has a sensible name and only does one thing so the person reading the code can still understand without seeing all the underlying detail in the main .rmd.

- *If you needed to alter your partner's figure using their code, do you think that would be easy or difficult, and why*

I think it would be quite easy for me to alter my partners figure using their code because they have used functions to break the code into more manageable chunks, that can be modified, while still showing whats going on.

d) Reflect on your own code based on your experience with your partner's code and their review of yours. (300-500 words)

- *What improvements did they suggest, and do you agree?*

My partner suggested that it would be helpful to have the code for plotting figures in the main document, rather than in in separate folders. I do agree that this would make it easier to quickly make a connection between the plotting code and outputted graph, however, I think that using folders for functions makes the code clearer and less cluttered. The second suggestion my partner made was to reduce the explanation within the lines of code. Again, I agree that it would more clear to have the explanation in one chunk, however, for a less experienced user of R it might be helpful to see the explanation line by line so they can understand exactly what is going on.

- *What did you learn about writing code for other people?*

I have learnt that writing code for other people is a lot more difficult than writing code for yourself mainly because of having to explain things in the clearest way possible. However, this assignment taught me lots of things to help make my code as reproducible as possible.

Firstly, I now know how to use pipelines in my code to make it more succinct and readable. Secondly, I now know how to create a function and why it's important to put these functions into separate R scripts and collate them into a folder called functions. Similarly, a data folder can be used to store copies of the raw and clean data, this is important so that nothing is overwritten.

Overall, I think one of the most useful things I have learnt about writing code for other people is that I can use GitHub to upload my code to a repository. This is helpful because it allows collaboration and changes to be tracked. This allows my code to be clear, consistent and understandable while also being open and sharing my work. These things are really important when making code for other people because it enhances reproducibility and replicability of scientific research.