Statistical methods for assessing agreement between two methods of clinical measurement

Solutions

Ernest Guevarra

30 November 2024

Contents

List of Figures

List of Tables

This document provides detailed solutions to the tasks set in the **Statistical methods for assessing agreement between two methods of clinical measurement** exercise set of the **Open and Reproducible Science in R** module of the **MSc in International Health and Tropical Medicine**.

1 Introduction to the exercise

The following tasks have been setup to help students get familiar with functional programming in R.

The students are expected to go through the tasks and appropriately write R code/script to fulfill the tasks and/or to answer the question/s being asked within the tasks. R code/script should be written inside a single R file named `ba.R` and saved in the project's root directory.

This exercise is based on:

> Bland, J. M. & Altman, DouglasG. Statistical Methods For Assessing Agreement Between Two Methods Of Clinical Measurement. Lancet 327, 307–310 (1986).

The dataset used in the paper can be accessed from the teaching_datasets repository. The URL to the `.dat` file is https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat.

The `ba.dat` dataset contains peak expiratory flow rate (PEFR) measurements (in litres per minute) taken with a Wright peak flow metre (`Wright` variable) and a Mini-Wright peak flow metre (`Mini` variable). This is the same data that is presented in the referenced Lancet article above.

2 Task 1: Read the dataset

This task is asking for the learner to create a function that would do the following:

1. Download the `ba.dat` file from the provided download link/URL; and,
2. Read that data into R.

To create this function, we can use already existing functions that does each step separately. For downloading files from the internet, there is a function called `download.file()`. For reading a text dataset, we can look at the `read.table()` family of functions.

2.1 Downloading files from the internet

The `download.file()` is the most straightforward function available in base R for downloading files from the internet. The basic syntax for the function is shown below:

```
download.file(
  url = "https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat", ①
  destfile = "data/ba.dat"                                                            ②
)
```

① Provide the download link/URL for your file. For this task, it would be https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat. Note that download link/URL should be enclosed in quotes.

② Provide the file path to where the file should be downloaded. The instructions specifically said that this should be downloaded into the `data` directory of the project and for the name, we just use the same name of the data file.

The arguments `url` and `destfile` are the minimum required arguments to specify. When you run this line of code, you should expect the `ba.dat` file to be downloaded into the `data` directory. You can check this by issuing the following command on the R console:

```
list.files("data")
```

```
[1] "ba.dat"
```

which shows that a file called `ba.dat` can be found inside the `data` directory of this project.

2.2 Reading text data

2.2.1 Reading text data after downloading

The `read.table()` function can be used to read the `ba.dat` file. After the data has been downloaded, it should be available in the `data` directory of our project. With this, we can use the `read.table()` function as follows:

```
read.table(
  file = "data/ba.dat",                                    ①
  header = TRUE,                                           ②
  sep = " "                                                ③
)
```

① Specify the file path to the dataset. In our case, the dataset has been downloaded into the `data` directory so the relative file path to the data is `data/ba.dat`.
② Specify whether the dataset being read has variable names or column names in the first row. In our case, the `ba.dat` dataset has variable names included (as described in the instructions). So, we specify `header = TRUE`.
③ Specify the value separators used in the dataset being read. The `ba.dat` dataset uses whitespace (`" "`) to separate values. So we specifcy `sep = " "`.

When we run this line of code, we get:

```
   Wright Mini
1     494  512
2     395  430
3     516  520
4     434  428
5     476  500
6     557  600
7     413  364
8     442  380
9     650  658
10    433  445
11    417  432
12    656  626
13    267  260
14    478  477
15    178  259
16    423  350
17    427  451
```

We get a `data.frame` with 17 rows and 2 columns of data.

2.2.2 Reading text data without downloading

The `read.table()` family of functions allow for reading of text data direct from a download link/URL without having to download the data first. The `file` argument in the `read.table()` family of functions accepts URL of a text data file. The function then reads the data directly from that URL. This functionality is useful for when downloading of data is not needed or if there are rules with regard to downloading and/or keeping of the required dataset. However, it should be noted that reproducibility using a link to read data directly into R depends on whether the URL link is a permanent link that will not change over time.

To read the `ba.dat` dataset directly into R without downloading first, we use the following code:

```
read.table(
  file = "https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat",
  header = TRUE,
  sep = " "
)
```

which gives:

```
   Wright Mini
1     494  512
2     395  430
3     516  520
4     434  428
5     476  500
6     557  600
7     413  364
8     442  380
9     650  658
10    433  445
11    417  432
12    656  626
13    267  260
14    478  477
15    178  259
16    423  350
17    427  451
```

This is the same output as the earlier approach.

2.2.3 Function to download and then read the dataset

Now that we know how to use the `download.file()` and the `read.table()` functions, we can now use them into creating a function that downloads the data and then reads it. A basic implementation of this function will look like this:

```
read_ba_data <- function(url, destfile) {                    ①
  download.file(url = url, destfile = destfile)              ②

  read.table(file = destfile, header = TRUE, sep = " ")      ③
}
```

① We set a function name of `read_ba_data()` and use arguments for `url` and `destfile` which are the two required arguments for `download.file()`. The `destfile` argument will then be used as the specification for the `file` argument for the path to the file to read in the `read.table()` function.
② Use the `url` and `destfile` argument specification to supply corresponding input values to the `download.file()` function.
③ Use the `destfile` argument specification to supply input value to the `read.table()` function. Specifications for `header` and `sep` arguments are hardcoded as it assumes that the `ba.dat` file has a header and uses `" "` as value separator.

We can then try this function to see if it gives us the expected outputs.

```
read_ba_data(
  url = "https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat",
  destfile = "data/ba.dat"
)
```

which produces the following result:

```
   Wright Mini
1     494  512
2     395  430
3     516  520
4     434  428
5     476  500
6     557  600
7     413  364
8     442  380
9     650  658
10    433  445
11    417  432
12    656  626
13    267  260
14    478  477
15    178  259
```

```
16    423  350
17    427  451
```

The same output is produced as earlier when the dataset was read into R. To check that the download component worked, we check if the `ba.dat` dataset is in the `data` directory of the project.

```
list.files("data")
```

```
[1] "ba.dat"
```

The `ba.dat` file is in the `data` directory. The `read_ba_data()` function works as expected.

2.2.4 Function to conditionally download dataset and then read the dataset

The `read_ba_data()` function is working the way we expect and need it to.

However, we can still consider adding some other features/functionalities that can make the function work more efficiently. For example, depending on our context and our requirements, we might want to just read the `ba.dat` dataset directly from the URL without downloading it. So, we probably would like to give the user of our function the ability to decide whether the file should be downloaded and then read or should just be read into R directly. For, this, we can refactor the earlier function as follows:

```
read_ba_data <- function(url,
                         download = TRUE,                    ①
                         destfile) {
  ## If download = TRUE, download dataset
  if (download) {                                            ②
    download.file(url = url, destfile = destfile)
  } else {
  ## If download = FALSE, set destfile as url link
    destfile <- url                                          ③
  }

  ## Read dataset
  read.table(file = destfile, header = TRUE, sep = " ")
}
```

① Use an argument called `download` which takes in logical values (TRUE or FALSE). If TRUE, then the function should download the dataset in the specified `destfile`. If FALSE, then the dataset is read directly from the `url`.

② Create a conditional `if()` statement based on whether `download` argument is TRUE or FALSE. If TRUE, then the function should download the dataset in the specified `destfile`.

③ If the `download` conditions is FALSE, then the `destfile` argument is specified with the `url` value so that the dataset is read directly from the `url`.

To test that this updated `read_ba_data()` function works as expected, we first remove the `ba.dat` file that we have already downloaded earlier by using the following commands in the R console:

```
file.remove("data/ba.dat")
```

```
[1] TRUE
```

Then, we check whether it has indeed been removed:

```
file.exists("data/ba.dat")
```

```
[1] FALSE
```

The `ba.dat` file doesn't exist in the `data` directory.

We then test the updated `read_ba_data()` function first with `download = FALSE` as follows:

```
read_ba_data(
  url = "https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat",
  download = FALSE                                                                  ①
)
```

① Set `download = FALSE` so that `ba.dat` is read directly into R without downloading. Since no download of data will be performed, there is no need to specify the `destfile` argument.

This code outputs the following:

```
   Wright Mini
1     494  512
2     395  430
3     516  520
4     434  428
5     476  500
6     557  600
7     413  364
8     442  380
9     650  658
10    433  445
11    417  432
12    656  626
```

```
13     267  260
14     478  477
15     178  259
16     423  350
17     427  451
```

which is the same output as the earlier function.

We then test whether or not the `ba.dat` file is present in the `data` directory. Our expectation is that the `ba.dat` file is not to be found there. We test as follows:

```
file.exists("data/ba.dat")
```

```
[1] FALSE
```

The `ba.dat` file is not found in the `data` directory. The updated `read_ba_data()` function works as expected.

2.2.5 Function to conditionally download dataset and then read the dataset - overwrite

The updated `read_ba_data()` function is working the way we expect and need it to.

However, we can still consider adding some conditionalities that can make the function work more efficiently. For example, once we have used the current `read_ba_data()` function with `download = TRUE` specification, the `ba.dat` file should already be in our `data` directory. Since this dataset is not expected to change anymore[1], we might want to update the current `read_ba_data()` function such that it will not overwrite an existing download of `ba.dat` if `download = TRUE`. This is useful as we don't have to repeat a download operation if the file is already present in our `data` directory. To implement this functionality, the `read_ba_data()` function can be updated as follows:

```
read_ba_data <- function(url,
                         download = TRUE,
                         destfile,
                         overwrite = FALSE) {                    ①
  ## If download = TRUE, download dataset
  if (download) {
    ## If overwrite = TRUE, download dataset
    if (overwrite) {                                             ②
      download.file(url = url, destfile = destfile)
    } else {
    ## If overwrite = FALSE, check if destfile exists before downloading
      if (!file.exists(destfile)) {                              ③
```

[1]This dataset is a teaching dataset and is provided by Bland and Altman in their paper. It is very reasonable to expect that no changes will happen to this dataset in the future.

```
        download.file(url = url, destfile = destfile)
      }
    }
  } else {
  ## If download = FALSE, set destfile as url link
    destfile <- url
  }

  ## Read dataset
  read.table(file = destfile, header = TRUE, sep = " ")
}
```

① Use `overwrite` argument that takes on logical (TRUE or FALSE) value so that user can specify whether they want the function to overwrite an existing file specified by `destfile`.

② If `download = TRUE` and `overwrite = TRUE`, download of the dataset is performed regardless of whether it is already present.

③ If `download = TRUE` and `overwrite = FALSE` (default), a check is performed whether the file path specified in `destfile` exists. If it doesn't exist, then download of the dataset is performed. If the file exists, then dataset is not downloaded.

We can now test whether the function works as expected. From the earlier example, we have the previous download of the `ba.dat` file and then ran the previous function with `download = FALSE`. So, no `ba.dat` file is currently found in the `data` directory. To test whether `overwrite = FALSE` works, we use the updated function as follows

```
read_ba_data(
  url = "https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat",
  download = TRUE,                                                                  ①
  destfile = "data/ba.dat",
  overwrite = FALSE
)
```

and then we get:

```
   Wright Mini
1     494  512
2     395  430
3     516  520
4     434  428
5     476  500
6     557  600
7     413  364
8     442  380
9     650  658
10    433  445
```

```
11     417  432
12     656  626
13     267  260
14     478  477
15     178  259
16     423  350
17     427  451
```

We get the data read into R and then we expect that the download process happened because `download = TRUE` even though `overwrite = FALSE` because the `ba.dat` file is not present in `data` directory. We check this by

```
file.exists("data/ba.dat")
```

```
[1] TRUE
```

The `ba.dat` file was found in the `data` directory of the project.

Now that the `ba.dat` file is back in the `data` directory, we can test the `overwrite = FALSE` argument again. We expect that the `read_ba_data()` will not re-download the `ba.dat` dataset. Instead, it will just read what is already available in the `data` directory.

```
read_ba_data(
  url = "https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat",
  download = TRUE,                                                                    ①
  destfile = "data/ba.dat",
  overwrite = FALSE
)
```

We get the data read into R and during the running of the code, we notice that no console message relating to download were showing with the whole process taking relatively quicker than earlier.

```
   Wright Mini
1     494  512
2     395  430
3     516  520
4     434  428
5     476  500
6     557  600
7     413  364
8     442  380
9     650  658
10    433  445
11    417  432
12    656  626
```

```
13    267  260
14    478  477
15    178  259
16    423  350
17    427  451
```

Then we can test the `overwrite = TRUE` argument. We expect that the `read_ba_data()` will download the `ba.dat` dataset regardless of whether the `ba.dat` is found in the `data` directory.

```
read_ba_data(
  url = "https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat",
  download = TRUE,                                                                  ①
  destfile = "data/ba.dat",
  overwrite = TRUE
)
```

```
   Wright Mini
1     494  512
2     395  430
3     516  520
4     434  428
5     476  500
6     557  600
7     413  364
8     442  380
9     650  658
10    433  445
11    417  432
12    656  626
13    267  260
14    478  477
15    178  259
16    423  350
17    427  451
```

We get the data read into R and during the running of the code, we notice that on the console we see messages relating to the download process were showing (see Console output 1) with the whole process taking relatively longer than earlier.

The updated `read_ba_data()` function is working as expected. We use it to create a data object called `ba_data`.

```
ba_data <- read_ba_data(
  url = "https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat",
  destfile = "data/ba.dat"
)
```

```
trying URL 'https://raw.githubusercontent.com/OxfordIHTM/teaching_datasets/main/ba.dat'
Content type 'text/plain; charset=utf-8' length 166 bytes
==================================================
downloaded 166 bytes
```

Console output 1: Showing downloading process ongoing

which results in:

```
ba_data
```

```
   Wright Mini
1     494  512
2     395  430
3     516  520
4     434  428
5     476  500
6     557  600
7     413  364
8     442  380
9     650  658
10    433  445
11    417  432
12    656  626
13    267  260
14    478  477
15    178  259
16    423  350
17    427  451
```

3 Task 2: Calculate the metrics needed for a Bland and Altman plot

The different metrics used to construct a Bland and Altman plot are:

- Mean of the per subject measurements made by the Wright and the Mini-Wright;
- Difference between the per subject measurements made by the Wright and the Mini-Wright;
- Mean of the difference between the per subject measurements made by the Wright and the Mini-Wright; and,
- Lower and upper limits of agreement between the per subject measurements made by the Wright and the Mini-Wright.

We can approach this task in five different ways.

3.1 Function to calculate Bland and Altman metrics - vectorised approach

The vectorised approach calculates each metric as vectors[2] and then concatenates them into a list[3]. The following function shows this approach:

```
calculate_ba_metrics <- function(ba_data) {                              ①
  ## Get per row mean of measurements
  mean_values <- (ba_data$Wright + ba_data$Mini) / 2                     ②

  ## Get per row difference of measurements
  differences <- ba_data$Wright - ba_data$Mini                           ③

  ## Mean of the differences of measurements
  mean_differences <- mean(differences)                                  ④

  ## Upper and lower limits of agreement
  upper_limit <- mean_differences + 1.96 * sd(differences)               ⑤
  lower_limit <- mean_differences - 1.96 * sd(differences)

  ## Concatenate metrics into a named list
  list(                                                                  ⑥
    mean_values = mean_values,
    differences = differences,
    mean_differences = mean_differences,
    upper_limit = upper_limit,
    lower_limit = lower_limit
  )
}
```

① Name the function `calculate_ba_metrics()` and use an argument `ba_data` which is the data.frame of the dataset we downloaded and read in the first task.
② Calculate the per row mean of the two measurements and assign it to the `mean_values` object.
③ Calculate the per row difference between the two measurements and assign it to the `differences` object.
④ Calculate the mean of the per row differences and assign it to the `mean differences` object.
⑤ Calculate the upper and lower limits of agreement and assign them to the `upper_limit` and `lower_limit` objects respectively.
⑥ Concatenate all the calculated values into a named list.

[2]A vector is essentially a collection of elements of the same data type, arranged in a one-dimensional array.
[3]A list is a data structure that can store elements of different types. Unlike vectors, which are homogenous (all elements must be of the same type), lists are heterogeneous and can hold elements such as numbers, strings, vectors, other lists, and even functions.

We use this function as follows:

```
calculate_ba_metrics(ba_data)
```

which gives the following output:

```
$mean_values
 [1] 503.0 412.5 518.0 431.0 488.0 578.5 388.5 411.0 654.0 439.0 424.5 641.0
[13] 263.5 477.5 218.5 386.5 439.0

$differences
 [1] -18 -35  -4   6 -24 -43  49  62  -8 -12 -15  30   7   1 -81  73 -24

$mean_differences
[1] -2.117647

$upper_limit
[1] 73.86201

$lower_limit
[1] -78.0973
```

We see the named list structure of the output with the `mean_values` and the `differences` being vectors whilst the rest of the metrics are single values.

3.2 Function to calculate Bland and Altman metrics - data.frame approach

The data.frame approach outputs each metric as vectors and concatenates them into the input dataset. The function using this approach can look like this:

```
calculate_ba_metrics <- function(ba_data) {
  ## Get per row mean of measurements and add to ba_data
  ba_data$mean_values <- (ba_data$Wright + ba_data$Mini) / 2              ①

  ## Get per row difference of measurements and add to ba_data
  ba_data$differences <- ba_data$Wright - ba_data$Mini                    ②

  ## Mean of the differences of measurements
  ba_data$mean_differences <- mean(ba_data$differences)                  ③

  ## Upper and lower limits of agreement
  ba_data$upper_limit <- ba_data$mean_differences + 1.96 *               ④
    sd(ba_data$differences)
  ba_data$lower_limit <- ba_data$mean_differences - 1.96 *
    sd(ba_data$differences)
```

```
  ## Return ba_data
  ba_data
}
```

① The per row mean of measurements are calculated and added as a new column in the input dataset.
② The per row difference in measurements are calculated and added as a new column in the input dataset.
③ The mean of the difference in measurements is calculated and added as a new column in the input dataset. Because this value is a single value, it is repeated per row of the input data.
④ The upper and lower limits of agreement in measurements are calculated and are each added as new columns in the input dataset. Because these values correspond to single values, they are repeated per row of the input data.

We use this function as follows:

```
calculate_ba_metrics(ba_data)
```

which gives the following output:

```
   Wright Mini mean_values differences mean_differences upper_limit lower_limit
1     494  512       503.0         -18        -2.117647    73.86201    -78.0973
2     395  430       412.5         -35        -2.117647    73.86201    -78.0973
3     516  520       518.0          -4        -2.117647    73.86201    -78.0973
4     434  428       431.0           6        -2.117647    73.86201    -78.0973
5     476  500       488.0         -24        -2.117647    73.86201    -78.0973
6     557  600       578.5         -43        -2.117647    73.86201    -78.0973
7     413  364       388.5          49        -2.117647    73.86201    -78.0973
8     442  380       411.0          62        -2.117647    73.86201    -78.0973
9     650  658       654.0          -8        -2.117647    73.86201    -78.0973
10    433  445       439.0         -12        -2.117647    73.86201    -78.0973
11    417  432       424.5         -15        -2.117647    73.86201    -78.0973
12    656  626       641.0          30        -2.117647    73.86201    -78.0973
13    267  260       263.5           7        -2.117647    73.86201    -78.0973
14    478  477       477.5           1        -2.117647    73.86201    -78.0973
15    178  259       218.5         -81        -2.117647    73.86201    -78.0973
16    423  350       386.5          73        -2.117647    73.86201    -78.0973
17    427  451       439.0         -24        -2.117647    73.86201    -78.0973
```

We see the `data.frame` structure of the output with new columns/variables for each of the calculated metric and the repeated per row values for the `mean_differences`, `upper_limit`, and `lower_limit`.

3.3 Function to calculate Bland and Altman metrics - combined approach

The structure of the output (either `list` or `data.frame`) each has their pros and cons. A `list` structure is relatively more flexible and can be transformed in many ways whilst the `data.frame` structure is ideal for use in functions that require a `data.frame` input (i.e., `ggplot2` for plotting). With this in mind, a combined approach that gives users an option between a vectorised or a data.frame approach can be useful and have a broader and more generalised usage.

This function can be implemented as follows:

```
calculate_ba_metrics <- function(ba_data, type = c("list", "df")) {          ①
  type <- match.arg(type)                                                    ②

  mean_values <- (ba_data$Wright + ba_data$Mini) / 2                         ③
  differences <- ba_data$Wright - ba_data$Mini
  mean_differences <- mean(differences)
  upper_limit <- mean_differences + 1.96 * sd(differences)
  lower_limit <- mean_differences - 1.96 * sd(differences)

  if (type == "list") {                                                      ④
    list(
      mean_values = mean_values,
      differences = differences,
      mean_differences = mean_differences,
      upper_limit = upper_limit,
      lower_limit = lower_limit
    )
  } else {
    data.frame(
      ba_data, mean_values, differences, mean_differences,
      upper_limit, lower_limit
    )
  }
}
```

① Use an argument called `type` which can be specified as either `list` or `df` (short for `data.frame`).
② The function `match.arg()` matches the input value for `type` to the choices allowed and checks whether it is specified correctly. If `type` not specified, `match.arg()` uses the first value (`list`) as the default.
③ Calculate each of the metrics like before.
④ Based on what `type` is, the metrics are concatenated into a `list` or into a `data.frame`.

We use this function as follows to output a `list`:

```
calculate_ba_metrics(ba_data)
```

which gives the following output:

```
$mean_values
 [1] 503.0 412.5 518.0 431.0 488.0 578.5 388.5 411.0 654.0 439.0 424.5 641.0
[13] 263.5 477.5 218.5 386.5 439.0

$differences
 [1] -18 -35  -4   6 -24 -43  49  62  -8 -12 -15  30   7   1 -81  73 -24

$mean_differences
[1] -2.117647

$upper_limit
[1] 73.86201

$lower_limit
[1] -78.0973
```

We use this function as follows to output a `data.frame`:

```
calculate_ba_metrics(ba_data, type = "df")
```

which gives the following output:

```
   Wright Mini mean_values differences mean_differences upper_limit lower_limit
1     494  512       503.0         -18        -2.117647    73.86201    -78.0973
2     395  430       412.5         -35        -2.117647    73.86201    -78.0973
3     516  520       518.0          -4        -2.117647    73.86201    -78.0973
4     434  428       431.0           6        -2.117647    73.86201    -78.0973
5     476  500       488.0         -24        -2.117647    73.86201    -78.0973
6     557  600       578.5         -43        -2.117647    73.86201    -78.0973
7     413  364       388.5          49        -2.117647    73.86201    -78.0973
8     442  380       411.0          62        -2.117647    73.86201    -78.0973
9     650  658       654.0          -8        -2.117647    73.86201    -78.0973
10    433  445       439.0         -12        -2.117647    73.86201    -78.0973
11    417  432       424.5         -15        -2.117647    73.86201    -78.0973
12    656  626       641.0          30        -2.117647    73.86201    -78.0973
13    267  260       263.5           7        -2.117647    73.86201    -78.0973
14    478  477       477.5           1        -2.117647    73.86201    -78.0973
15    178  259       218.5         -81        -2.117647    73.86201    -78.0973
16    423  350       386.5          73        -2.117647    73.86201    -78.0973
17    427  451       439.0         -24        -2.117647    73.86201    -78.0973
```

3.4 Function to calculate Bland and Altman metrics - modular approach

The modular approach creates multiple functions that calculates each component metric of the Bland and Altman plot and then assembles them into one overall function. Function 1, Function 2, Function 3, Function 4 each calculate one of the metrics needed for the Bland and Altman plot.

```
calculate_mean_values <- function(ba_data) {
  (ba_data$Wright + ba_data$Mini) / 2
}
```

Function 1: Calculate per row mean of measurements

```
calculate_diff_values <- function(ba_data) {
  ba_data$Wright - ba_data$Mini
}
```

Function 2: Calculate per row difference of measurements

```
calculate_mean_diff <- function(ba_data) {
  mean(calculate_diff_values(ba_data))
}
```

Function 3: Calculate the mean of differences of the measurements

Each of these functions require the same arguments: `m1` which is a vector of numeric values for the first measurement and `m2` which is a vector of numeric values for the second measurement. Function 1 and Function 2 are very simple functions performing basic row-wise vectorised operations. Function 3 and Function 4 use the first 2 functions to perform the calculations.

Then, using all these four functions, Function 5 calculates all the Bland and Altman metrics and provides the option for the user to output a `list` or a `data.frame`.

We can apply this overall function as follows:

```
calculate_ba_metrics(ba_data)
```

which gives the following output

```
$mean_values
 [1] 503.0 412.5 518.0 431.0 488.0 578.5 388.5 411.0 654.0 439.0 424.5 641.0
[13] 263.5 477.5 218.5 386.5 439.0

$differences
 [1] -18 -35  -4   6 -24 -43  49  62  -8 -12 -15  30   7   1 -81  73 -24
```

```r
calculate_diff_limits <- function(ba_data) {
  differences <- calculate_diff_values(ba_data)
  mean_differences <- calculate_mean_diff(ba_data)

  upper_limit <- mean_differences + 1.96 * sd(differences)
  lower_limit <- mean_differences - 1.96 * sd(differences)

  c(upper_limit, lower_limit)
}
```

Function 4: Calculate the upper and lower limits of agreement

```
$mean_differences
[1] -2.117647

$upper_limit
[1] 73.86201

$lower_limit
[1] -78.0973
```

To output a `data.frame`, we use the function as follows:

```r
calculate_ba_metrics(ba_data, type = "df")
```

which gives the following output

```
   Wright Mini mean_values differences mean_differences upper_limit lower_limit
1     494  512       503.0         -18        -2.117647    73.86201    -78.0973
2     395  430       412.5         -35        -2.117647    73.86201    -78.0973
3     516  520       518.0          -4        -2.117647    73.86201    -78.0973
4     434  428       431.0           6        -2.117647    73.86201    -78.0973
5     476  500       488.0         -24        -2.117647    73.86201    -78.0973
6     557  600       578.5         -43        -2.117647    73.86201    -78.0973
7     413  364       388.5          49        -2.117647    73.86201    -78.0973
8     442  380       411.0          62        -2.117647    73.86201    -78.0973
9     650  658       654.0          -8        -2.117647    73.86201    -78.0973
10    433  445       439.0         -12        -2.117647    73.86201    -78.0973
11    417  432       424.5         -15        -2.117647    73.86201    -78.0973
12    656  626       641.0          30        -2.117647    73.86201    -78.0973
13    267  260       263.5           7        -2.117647    73.86201    -78.0973
14    478  477       477.5           1        -2.117647    73.86201    -78.0973
15    178  259       218.5         -81        -2.117647    73.86201    -78.0973
16    423  350       386.5          73        -2.117647    73.86201    -78.0973
17    427  451       439.0         -24        -2.117647    73.86201    -78.0973
```

```r
calculate_ba_metrics <- function(ba_data, type = c("list", "df")) {
  type <- match.arg(type)

  mean_values <- calculate_mean_values(ba_data)
  differences <- calculate_diff_values(ba_data)
  mean_differences <- calculate_mean_diff(ba_data)
  limits <- calculate_diff_limits(ba_data)

  if (type == "list") {
    list(
      mean_values = mean_values,
      differences = differences,
      mean_differences = mean_differences,
      upper_limit = limits[1],
      lower_limit = limits[2]
    )
  } else {
    data.frame(
      ba_data, mean_values, differences, mean_differences,
      upper_limit = limits[1], lower_limit = limits[2]
    )
  }
}
```

Function 5: Calculate all the Bland and Altman metrics

3.5 Function to calculate Bland and Altman metrics - universal approach

All the above approaches for creating a function to calculate Bland and Altman metrics have been designed specific for the `ba.dat` dataset. What this means is that the functions work appropriately but only for a dataset that has variables called ***"Wright"*** and ***"Mini"***. Hence, these functions will be useful for this specific exercise only. Once you have a dataset with two values of measurements similar to the `ba.dat` dataset which, in theory, can be analysed using the Bland and Altman method, these functions cannot be used for that dataset.

It would be ideal that we try to make this function as universally applicable as possible such that it can be used with any dataset that has data to which the Bland and Altman method can be applied to. This will make the function usable beyond just this exercise.

It is relatively easy to improve our current function to make it universal. Instead of making the function require a specific `data.frame` input as its main argument, we can instead make the function require vectors of values for the first and second measurements instead as its arguments. This is demonstrated by the functions below.

Function 6, Function 7, Function 8, Function 9 are variations of the earlier per metric

```r
calculate_mean_values <- function(m1, m2) {
  (m1 + m2) / 2
}
```

Function 6: Calculate per row mean of measurements

```r
calculate_diff_values <- function(m1, m2) {
  m1 - m2
}
```

Function 7: Calculate per row difference of measurements

functions. Instead of a `data.frame` as the required input, we use two arguments `m1` and `m2` for vectors of values for the first and second measurements respectively. This is a very minor adjustment but it makes a big difference in making the functions more universal. By doing this, we let the user decide the kind of input to provide to the function based on their data.

Now, for the final overall function shown below, we set an argument for a data.frame input that contains measurements values that can be analysed using the Bland and Altman approach and then a further two arguments for the names of the variables for the first and second measurements that are being assessed.

```r
calculate_ba_metrics <- function(df,                                    ①
                                 m1, m2,                                ②
                                 type = c("list", "df")) {
  type <- match.arg(type)

  m1 <- df[ , m1]
  m2 <- df[ , m2]

  mean_values <- calculate_mean_values(m1 = m1, m2 = m2)
  differences <- calculate_diff_values(m1 = m1, m2 = m2)
  mean_differences <- calculate_mean_diff(m1 = m1, m2 = m2)
  limits <- calculate_diff_limits(m1 = m1, m2 = m2)

  if (type == "list") {
    list(
      mean_values = mean_values,
      differences = differences,
      mean_differences = mean_differences,
      upper_limit = limits[1],
      lower_limit = limits[2]
    )
  } else {
    data.frame(
```

```r
calculate_mean_diff <- function(m1, m2) {
  mean(calculate_diff_values(m1 = m1, m2 = m2))
}
```

Function 8: Calculate the mean of differences of the measurements

```r
calculate_diff_limits <- function(m1, m2) {
  differences <- calculate_diff_values(m1 = m1, m2 = m2)
  mean_differences <- calculate_mean_diff(m1 = m1, m2 = m2)

  upper_limit <- mean_differences + 1.96 * sd(differences)
  lower_limit <- mean_differences - 1.96 * sd(differences)

  c(upper_limit, lower_limit)
}
```

Function 9: Calculate the upper and lower limits of agreement

```r
      df, mean_values, differences, mean_differences,
      upper_limit = limits[1], lower_limit = limits[2]
    )
  }
}
```

① A data.frame object containing measurement values that can be analysed using the Bland and Altman approach.
② Two additional arguments for character values of the names of the variables in the `data.frame` input for the values for the first and second measurements respectively.

The overall function can be used as follows:

```r
calculate_ba_metrics(df = ba_data, m1 = "Wright", m2 = "Mini")
```

which gives the following `list` output

```
$mean_values
 [1] 503.0 412.5 518.0 431.0 488.0 578.5 388.5 411.0 654.0 439.0 424.5 641.0
[13] 263.5 477.5 218.5 386.5 439.0

$differences
 [1] -18 -35  -4   6 -24 -43  49  62  -8 -12 -15  30   7   1 -81  73 -24

$mean_differences
[1] -2.117647

$upper_limit
```

```
[1] 73.86201

$lower_limit
[1] -78.0973
```

or

```
calculate_ba_metrics(df = ba_data, m1 = "Wright", m2 = "Mini", type = "df")
```

which gives the following `data.frame` output

```
   Wright Mini mean_values differences mean_differences upper_limit lower_limit
1     494  512       503.0         -18        -2.117647    73.86201    -78.0973
2     395  430       412.5         -35        -2.117647    73.86201    -78.0973
3     516  520       518.0          -4        -2.117647    73.86201    -78.0973
4     434  428       431.0           6        -2.117647    73.86201    -78.0973
5     476  500       488.0         -24        -2.117647    73.86201    -78.0973
6     557  600       578.5         -43        -2.117647    73.86201    -78.0973
7     413  364       388.5          49        -2.117647    73.86201    -78.0973
8     442  380       411.0          62        -2.117647    73.86201    -78.0973
9     650  658       654.0          -8        -2.117647    73.86201    -78.0973
10    433  445       439.0         -12        -2.117647    73.86201    -78.0973
11    417  432       424.5         -15        -2.117647    73.86201    -78.0973
12    656  626       641.0          30        -2.117647    73.86201    -78.0973
13    267  260       263.5           7        -2.117647    73.86201    -78.0973
14    478  477       477.5           1        -2.117647    73.86201    -78.0973
15    178  259       218.5         -81        -2.117647    73.86201    -78.0973
16    423  350       386.5          73        -2.117647    73.86201    -78.0973
17    427  451       439.0         -24        -2.117647    73.86201    -78.0973
```

We now have a function that we can use for any data that can be analysed using the Bland and Altman method/approach.

We should now create an object called `ba_metrics` to store the output of this function for use in the next step (plotting).

```
ba_metrics <- calculate_ba_metrics(df = ba_data, m1 = "Wright", m2 = "Mini")
```

which gives us

```
ba_metrics
```

```
$mean_values
 [1] 503.0 412.5 518.0 431.0 488.0 578.5 388.5 411.0 654.0 439.0 424.5 641.0
[13] 263.5 477.5 218.5 386.5 439.0
```

```
$differences
 [1] -18 -35  -4   6 -24 -43  49  62  -8 -12 -15  30   7   1 -81  73 -24

$mean_differences
[1] -2.117647

$upper_limit
[1] 73.86201

$lower_limit
[1] -78.0973
```

4 Task 3: Create a Bland and Altman plot

The Bland and Altman method lends itself for useful graphical representation of outputs. This is why the Bland and Altman plot is commonly used to identify this analytical approach.

The Bland and Altman plot, also known as a ***Tukey mean-difference plot***, is a graphical method used to compare two different measurement techniques or instruments by analyzing the agreement between them. It is commonly used in clinical research and other scientific disciplines where comparing two methods is essential.

4.1 Key Features of the Bland-Altman Plot

4.1.1 Axes

- The **x-axis** represents the mean of the two measurements for each observation.
- The **y-axis** represents the difference between the two measurements for each observation.

4.1.2 Data Points

Each point on the plot corresponds to one observation, showing the relationship between the average measurement and the difference between the two methods.

4.1.3 Central Line

A horizontal line at the mean difference (**bias**) indicates the average level of disagreement between the two methods.

4.1.4 Limits of Agreement (LoA)

Two horizontal lines represent the limits within which most differences between the two methods will fall.

These limits approximate a 95% confidence interval for the differences.

4.2 Interpretation

- If the differences cluster around zero and within the limits of agreement, the two methods are considered to agree well.
- Systematic bias is indicated by a consistent deviation from zero.
- A pattern in the differences (e.g., increasing or decreasing differences) might suggest proportional bias or other issues with one or both methods.

4.3 Creating a Bland and Altman plot

4.3.1 Basic Bland and Altman plot

The following code creates a basic Bland and Altman plot:

```
plot_ba <- function(ba_metrics) {                                    ①
  plot(x = ba_metrics$mean_values, y = ba_metrics$differences)       ②

  abline(h = ba_metrics$mean_differences)                            ③
  abline(h = ba_metrics$upper_limit)
  abline(h = ba_metrics$lower_limit)
}
```

① Use `ba_metrics` output from previous function calculating Bland and Altman metrics.
② Create a scatter plot using the base `plot()` function in R with the per row mean of measurements in the x-axis and the per row differences of measurements in the y-axis.
③ Create horizontal lines for the mean differences (**bias**), upper, and lower limits of agreement of the plot.

We use this function as follows:

```
plot_ba(ba_metrics)
```

which produces the following plot (Figure 1):

In this output plot, we see that no title is given to the plot. The x- and y-axis labels defaults to the input values for the `x` and `y` arguments in the `plot()` function. The horizontal lines for the bias and limits of agreement are solid.

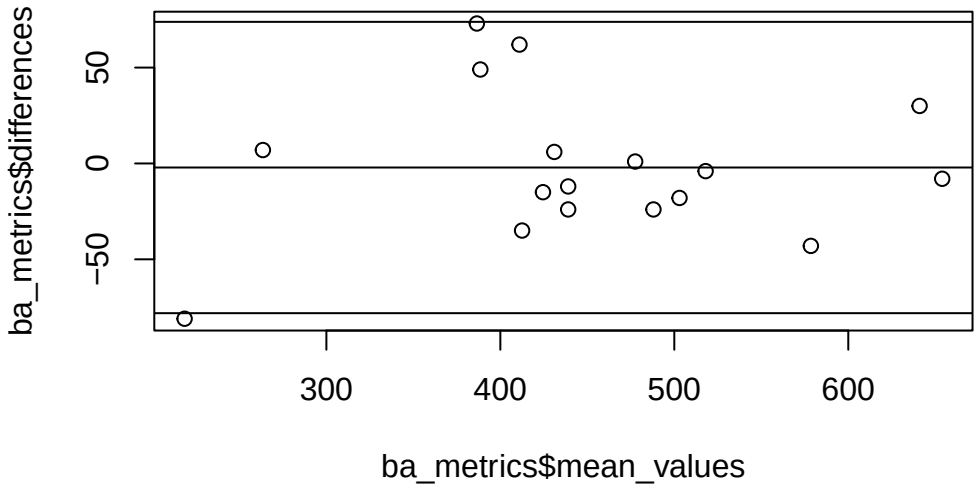

Figure 1: Bland and Altman plot - basic

4.3.2 Bland and Altman plot with labels

We can improve the basic plot by adding relevant labels. These are:

- A title for the plot;
- Axis labels for x- and y-axis; and,
- Additional styling and labels for the bias line and the limits of agreement lines.

The following function shows how this can be done:

```
plot_ba <- function(ba_metrics,
                    title = "Bland and Altman plot",                    ①
                    xlab = NULL, ylab = NULL,                           ②
                    limits_lab = TRUE) {                                ③
  plot(
    x = ba_metrics$mean_values,
    y = ba_metrics$differences,
    main = title,                                                       ④
    xlab = xlab, ylab = ylab                                            ⑤
  )

  abline(h = ba_metrics$mean_differences, lty = 2, lwd = 0.7)           ⑥
  abline(h = ba_metrics$upper_limit, lty = 2, lwd = 0.7)
  abline(h = ba_metrics$lower_limit, lty = 2, lwd = 0.7)
```

```
  if (limits_lab) {                                                    ⑦
    ## Label for mean differences line
    text(                                                              ⑧
      x = max(ba_metrics$mean_values), y = ba_metrics$mean_differences,
      labels = paste0(
        "Mean difference: ", round(ba_metrics$mean_differences, digits = 1)
      ),
      pos = 2, cex = 0.70
    )

    ## Label for upper limit of agreement
    text(
      x = max(ba_metrics$mean_values), y = ba_metrics$upper_limit,
      labels = paste0(
        "Upper limit: ", round(ba_metrics$upper_limit, digits = 1)
      ),
      pos = 2, cex = 0.70
    )

    ## Label for lower limit of agreement
    text(
      x = max(ba_metrics$mean_values), y = ba_metrics$lower_limit,
      labels = paste0(
        "Lower limit: ", round(ba_metrics$lower_limit, digits = 1)
      ),
      pos = 2, cex = 0.70
    )
  }
}
```

① Use an argument `title` for the title of the plot. A default title is given which will be used if the user doesn't supply a title value.

② Use `xlab` and `ylab` arguments for the x- and y-axis labels respectively. This is set to `NULL` by default. If user doesn't supply values for these arguments, the function will use the default labels used in the `plot()` function.

③ Use `limits_lab` argument which takes a logical (TRUE or FALSE) input. If TRUE (default), then labels for the bias, upper, and lower limits of agreement will be added to the plot. If FALSE, no labels will be added.

④ The value supplied to the `title` argument is used in the `main` argument of the `plot()` function.

⑤ The values supplied to `xlab` and `ylab` arguments are used in the `xlab` and `ylab` arguments of the `plot()` function.

⑥ The lines for the bias and the limits of agreement are stylised into dashed lines by setting the `lty` argument of the `abline()` function to ***2*** which will use a dashed line. The lines are also made a little bit thinner by specifying a value of ***0.7*** to the `lwd` argument

of the `abline()` function. This style is used to the lines so that when the labels are added onto the lines, the label will be more visible and the lines blend more to the background.

⑦ Create a condition using the `if` function that will check whether the `limits_lab` is TRUE or FALSE which will create the labels for the lines if TRUE.

⑧ If `limits_lab` is TRUE, then the bias and limits of agreement lines are created using the `text()` function. The labels are positioned to the left (`pos = 2`) of the maximum value of the mean values of the measurements (`x`) and at the values of the bias, upper, and lower limits of agreement (`y` respectively). The label for the text identifies what the line represents and the value it represents. The text is given a character expansion (`cex`) of ***0.7***.

We can use this function as follows:

```
plot_ba(
  ba_metrics,
  title = "Wright vs Mini-Wright",
  xlab = "Mean PEFR (per subject)",
  ylab = "Difference in PEFR (per subject)"
)
```

which outputs this plot (see Figure 2):

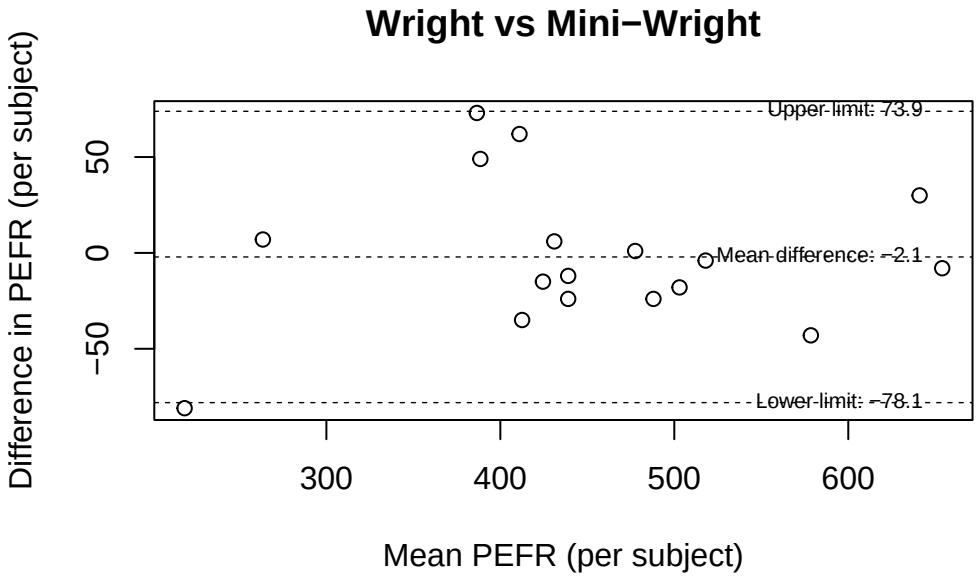

Figure 2: Bland and Altman plot with customised labels

The output plot has a customised title, the x- and y-axis have appropriate and informative labels, the bias line and the limit of agreement lines are styled appropriately and labeled more informatively.

4.3.3 Bland and Altman plot with colours and additional styling

We can further improve on the labeled version of the plot added styling and colour to the points of the scatter plot.

The points of the scatter plot use default points style of a hollow point in black outline. We can add styling and colour with this version of the function:

```
plot_ba <- function(ba_metrics,
                    title = "Bland and Altman plot",
                    xlab = NULL, ylab = NULL, limits_lab = TRUE,
                    pch = NULL, col = NULL, bg = NULL, cex = NULL) {      ①
  plot(
    x = ba_metrics$mean_values,
    y = ba_metrics$differences,
    main = title,
    xlab = xlab, ylab = NULL,
    pch = pch,                                                            ②
    col = ifelse(is.null(col), "black", col),                             ③
    bg = bg,                                                              ④
    cex = cex                                                             ⑤
  )

  abline(h = ba_metrics$mean_differences, lty = 2, lwd = 0.7)
  abline(h = ba_metrics$upper_limit, lty = 2, lwd = 0.7)
  abline(h = ba_metrics$lower_limit, lty = 2, lwd = 0.7)

  if (limits_lab) {
    ## Label for mean differences line
    text(
      x = max(ba_metrics$mean_values), y = ba_metrics$mean_differences,
      labels = paste0(
        "Mean difference: ", round(ba_metrics$mean_differences, digits = 1)
      ),
      pos = 2, cex = 0.70
    )

    ## Label for upper limit of agreement
    text(
      x = max(ba_metrics$mean_values), y = ba_metrics$upper_limit,
      labels = paste0(
        "Upper limit: ", round(ba_metrics$upper_limit, digits = 1)
      ),
      pos = 2, cex = 0.70
    )
```

```
    ## Label for lower limit of agreement
    text(
      x = max(ba_metrics$mean_values), y = ba_metrics$lower_limit,
      labels = paste0(
        "Lower limit: ", round(ba_metrics$lower_limit, digits = 1)
      ),
      pos = 2, cex = 0.70
    )
  }
}
```

① Use `pch`, `col`, `bg`, and `cex` arguments for the function to style and colour the points of the plot.
② Use `pch` argument to specify an integer value for the character type to use for the points of the plot. See `?pch` for details. Default is NULL which will use default `pch` value used by R (which is ***1*** for a hollow circle).
③ Use `col` argument to provide a colour specification to use for colouring the points in the scatter plot. Default is NULL which sets the colour to default colour used by R (***black***). Note that for hollow points (`pch` from 0 to 14) and for points with a fill element (`pch` from 21 to 25), `col` will colour the outline of the point. For `pch` values from 15 to 19, `col` will colour the whole point. To colour the fill element of points specified by `pch` from 21 to 25, see `bg` argument.
④ Use `bg` argument to provide a colour specification to use for colouring the fill element of points with a fill element (`pch` from 21 to 25). Default is NULL which sets the fill element to a ***light gray***.
⑤ Use `cex` argument to provide a character expansion numeric value for the points of the scatter plot. Default is NULL which will use the default size of points used by R which is a value of ***1***.

We can use this function as follows:

```
plot_ba(
  ba_metrics,
  title = "Wright vs Mini-Wright",
  xlab = "Mean PEFR (per subject)",
  ylab = "Difference in PEFR (per subject)",
  pch = 21,
  col = "darkblue",
  bg = "lightblue",
  cex = 1.2
)
```

which outputs this plot (see Figure 3):

The output plot has much bigger points than the previous plot and because we used one of the points with a fill element, the points have a dark blue outline and a light blue fill.

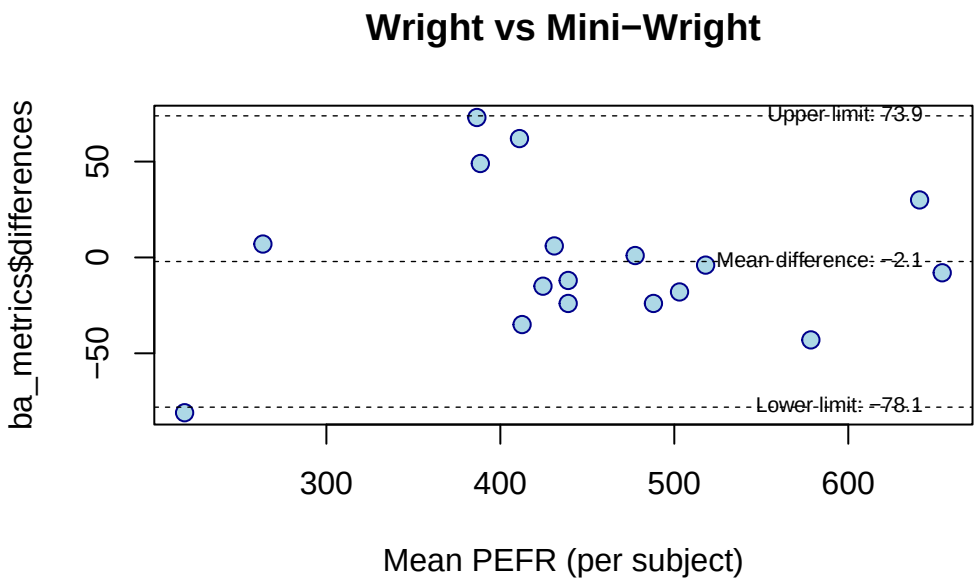

Figure 3: Bland and Altman plot with customised points styles and colours

4.4 Some guidance on plotting functions

Data visualisation using plots in R is a very individual-preference type of approach. Hence, the general recommendation on making plotting functions is to be a little bit opinionated with the plot style (colours, line types and widths, text labels, sizes, etc.) based on your individual style. Because you are making a function intended for your use, you want that function to implement the style and look that you like. What this means, however, is that your function is most likely something that will be used only by you or others who might like your style and preference.