MECH481A6: Engineering Data Analysis in R

Chapter 10 Homework: Measurement

Flynn Nyman

06 November, 2024

Load packages

Chapter 10 Homework

This homework will give you practice at working with a measurement dataset: `airlift_mass_repeatability.csv`. This data set represents repeated measures of "blank" air sampling filters.

A couple notes to consider when reporting answers in response to questions. The microbalance used to make these measurements reads out to the nearest microgram (μg), which is 0.000001 g or 0.001 mg. Thus, be careful when reporting descriptive statistics so as not to overstate your **precision**. Use the `round()` function to avoid reporting more than 0.1 μg of precision (or 0.0001 mg). Here is some example code that uses the `across()` function from `dplyr::` to round numeric output to just four digits (appropriate for mg units in this exercise):

```
dplyr::mutate(across(.cols = where(is.numeric), .fns = round, 3))
```

Question 1

Import the `airlift_mass_repeatability.csv` file into a data frame called `blanks` and perform the following data wrangling in a single pipe:

- retain only the first 3 columns of data;
- rename the columns with the names `date`, `id`, and `mass_mg`;
- convert the `date` column vector into a date class object using `lubridate::`
- convert the `id` variable to a class `factor` (this can be accomplished using `base::as.factor()` or `purrr::as_factor()`)
- create a new column vector named `mass_mg` by rescaling the `mass_g` data (i.e., convert g to mg by multiplying `mass_g` by 1000)

Question 2:

2a. Are there any NAs present in the data frame?
There are no missing values in the data set.
2b. How many unique filter IDs are present in this data frame?
There are 5 unique filter IDs.
2c. How many samples are present for each filter ID? Hint: look up the `dplyr::count()` function.
There are 78 occurrences of IDs 41666, 41667, 41668, and 41671. There are 76 occurrences of ID 41669.
2d. Over how long of a period were these blank measurements made? Hint: this can be done in base R with a `max() - min()` or with `lubridate::interval() %>% as.duration()`. There is a span of 35 days between the earliest and latest measurement.

```
## [1] FALSE
```

```
## [1] 41669 41671 41667 41666 41668
## Levels: 41666 41667 41668 41669 41671
```

```
##      id  n
## 1 41666 78
## 2 41667 78
## 3 41668 78
## 4 41669 76
## 5 41671 78
```

```
## [1] 35
```

Question 3

Group the `blanks` data frame by `id` and calculate mean, median, and standard deviations for each filter id. Hint: use `group_by() %>% summarise()` to do this efficiently.

```
## # A tibble: 5 x 4
##   id     mean median       sd
##   <fct> <dbl>  <dbl>    <dbl>
## 1 41666  98.3   98.3 0.000767
## 2 41667  95.5   95.5 0.000534
## 3 41668  98.0   98.0 0.000834
## 4 41669  97.8   97.8 0.00113
## 5 41671  97.6   97.6 0.000834
```

Question 4

Calculate the limit of detection (LOD) for this measurement method. Note: you will need to calculate standard deviations for each filter `id` (as done in question 3) and then estimate LOD from $LOD = 3 \cdot \sigma_b$ where σ_b is calculated for each filter `id`.

```
## # A tibble: 5 x 1
##     LOD
##   <dbl>
## 1  98.3
## 2  95.5
## 3  98.0
## 4  97.8
## 5  97.6
```

Appendix

```
# set global options for figures, code, warnings, and messages
knitr::opts_chunk$set(fig.width=6, fig.height=4, fig.path="../figs/",
                      echo=FALSE, warning=FALSE, message=FALSE)
# load packages for current session
library(tidyverse)
library(gridExtra)
library(MASS)
library(lubridate)
library(dplyr)
airlift_data <- read.csv("./AIRLIFT_mass_repeatability.csv") %>%
  dplyr::select(Date, Filter.ID, Mass..g.) %>%
  rename("date" = "Date", "id" = "Filter.ID", "mass_g" = "Mass..g.") %>%
  mutate(date = dmy(date)) %>%
  mutate(id = as.factor(id)) %>%
  mutate(mass_mg = mass_g*1000)
NA_present <- any(is.na(airlift_data)) %>%
  print()
## There are no missing values in the data set.

unique_IDs <- unique(airlift_data$id) %>%
  print()
## There are 5 unique filter IDs in the data set.

airlift_data %>%
  count(id)
## There are 78 occurrences of IDs 41666, 41667, 41668, and 41671. There are 76 occurrences of ID 41669

measurement_span <- as.numeric(difftime(max(airlift_data$date, na.rm = TRUE), min(airlift_data$date, na
  print()
## There is a span of 35 days between the earliest and latest measurement.
summary <- airlift_data %>%
  group_by(id) %>%
  summarise(mean(mass_mg), median(mass_mg), sd(mass_mg)) %>%
  rename('mean' = 'mean(mass_mg)',
         'median' = 'median(mass_mg)',
         'sd' = 'sd(mass_mg)') %>%
  print()
LOD <- summary %>%
  mutate(LOD = mean + 3*sd) %>%
  dplyr::select(!c(id, mean, median, sd)) %>%
  print()
```