MECH481A6: Engineering Data Analysis in R

Chapter 10 Homework: Measurement

Brad Portouw

30 November, 2022

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?
2b. How many unique filter IDs are present in this data frame?
2c. How many samples are present for each filter ID? Hint: look up the `dplyr::count()` function.
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()`.

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

```
## Time difference of 35 days
```

Seemingly there are no NA entries for the blanks data frame. The data from the raw airlift_mass_repeatability.csv did have many columns with NA's present, however those columns were removed for this data analysis.

The vector x, lists all the unique id's for the different sensors used with 5 sensors total being:

41666, 41667, 41668, 41669, 41671

. Line 73 counts the number of data points for each sensor, placing them in the data frame above. Four sensors have 78 datapoints, with sensor 41669 having 76.

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_mass_mg 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 2
##   id        LOD
##   <fct>   <dbl>
## 1 41666 0.00230
## 2 41667 0.00160
## 3 41668 0.00250
## 4 41669 0.00340
## 5 41671 0.00250
```