Statistical analysis of breath VOC metabolite data in R (BIOL72031)

2025-12-16

Aim: Learn basic R commands to explore breath VOC metabolite data, make plots, and compare groups using basic statistical analysis.

Data: Phase I clinical trial data to investigate changes in breath VOCs during circadian rhythmicity in asthma. Some metadata was collected through questionnaires during the clinical visit. VOC peak areas were measured using GC-MS from participant breath samples and instrument blank (i.e. background) samples.

Download R and data

Download RStudio (https://posit.co/download/rstudio-desktop/) or use the online version of RStudio by creating a free account on https://posit.cloud/

If you are new to RStudio, check out this user guide https://docs.posit.co/ide/user/ide/get-started/

1. Open a new script in RStudio and copy and paste the code in the grey boxes below. You can highlight a chunk of code and click Run (or CTRL + ENTER) to run it.

Install and load the relevant packages in R. Packages are collections of useful functions so you don't need to write your own. These specific packages below are widely used for statistical analysis

```
install.packages("tidyverse")
library(tidyverse)

install.packages("ggpubr")
library(ggpubr)
```

Explore and summarise data

Download the data file of the VOC data from the GitHub server. Once data is loaded into the R "environment" window, the output will give a short summary.

The assign operator <- will copy the file as "data" in our R environment.

```
data <-
read_csv("https://raw.githubusercontent.com/Waqar54/Rtutorial/main/BIOL72031.
csv")
```

Open and explore the data table. Running the code below will give us the top 10 rows. Each row represents a sample, and each column represents a variable (e.g. sample metadata or VOC peak areas).

```
data
```

Summarise the different sample types using the count() function. The pipe operator %>% just means "then do this", so we can use multiple functions in one chunk of code. It does not save or overwrite any data files in the environment.

```
data %>% count(Sample_type)
```

Look at what drinks participants had before their clinical visit. We will not include "blank" samples using the *filter* function. We will also use the *sort* function to sort drink count from highest to lowest.

Which is the most popular drink?

```
data %>%
  filter(Sample_type == "Breath") %>%
  count(Drink, sort = TRUE)
```

Lets try the same code but change the variable to the "Medication" column instead of "Drinks"

How many did not use inhalers?

```
data %>%
  filter(Sample_type == "Breath") %>%
  count(Medication, sort = TRUE)
```

Prepare data for statistical analysis

In the same chunk of code we will:

a. structure only the VOC numerical data into a long format so that each row represents a data point.
b. normalise all samples so they are comparable with each other.
c. log transform for easy visual comparison.

For a new variable we will add a column by using the "mutate" function

Notice how we create a new dataframe called "data_std", so we don't overwrite the original data.

Can you identify where in the code we have performed all three tasks?

```
data_std <- data %>%
  pivot_longer(
    cols = -(Sample_ID:ISTD),
    names_to = "voc",
    values_to = "area"
  ) %>%
  mutate(intensity = log(area / ISTD))
```

Plot and interpret data

First we will generate a histogram plot for each VOC to compare breath and blank samples. For plotting data we will use ggplot, a common function used for plotting data in R. **Which VOC has similar peak intensities in both breath and blank samples?**

```
data_std %>%
  ggplot(aes(x = intensity, fill = Sample_type)) +
  geom_histogram() +
  facet_wrap(~ voc, scales = "free")
```

Next we will statistically compare medication using a box-and-whisker plot. Here we will use "stat_compare_means" to perform statistical comparisons for us. **What does the p-value tell you?**

```
data_std %>%
  filter(Sample_type == "Breath",
         Medication %in% c("Clenil", "None")) %>%
  ggplot(aes(x = Medication, y = intensity, fill = Medication)) +
  geom_boxplot() +
  facet_wrap(~ voc, scales = "free") +
  stat_compare_means()
```

Lets also statistically compare drinks. **Which drink shows the largest difference between VOCs?**

```
data_std %>%
  filter(Sample_type == "Breath") %>%
  group_by(Drink) %>%
  filter(n() > 30) %>%
  ggplot(aes(x = Drink, y = intensity, fill = Drink)) +
  geom_boxplot() +
  facet_wrap(~ voc, scales = "free") +
  stat_compare_means()
```

Lets now add pairwise comparisons between drinks. First we need to define which variables we want to compare by creating a list called "drink_comparisons"

```
drink_comparisons <- list(
  c("Orange Juice", "Water"),
  c("Orange Juice", "Tea")
)
data_std %>%
  filter(Sample_type == "Breath") %>%
  group_by(Drink) %>%
  filter(n() > 30) %>%
  ggplot(aes(x = Drink, y = intensity, fill = Drink)) +
  geom_boxplot() +
  facet_wrap(~ voc, scales = "free") +
  stat_compare_means(comparisons = drink_comparisons)
```

Now that we've explored some Breath VOC data from real clinical trial, we can start to build some conclusions and formulate new hypotheses.

Do VOCs change with drug treatment?

Are these VOCs metabolites of the drug or are they from some confounding exogenous factors?

Do you think the type of drink you have could interfere with potentially new biomarkers?

How would you design a new clinical study based on these results?