

# MLW / KUHeS Statistics and R short course

## Session 1 - Practical

Evaristar Kudowa, Marc Henrion

2026-01-19

### Session 1 - Practical (Solutions)

Go to the course website on GitHub:

[https://github.com/mlw-stats/R\\_And\\_Statistics\\_Training\\_2026/Session1](https://github.com/mlw-stats/R_And_Statistics_Training_2026/Session1)

From here, in the **dataAndSupportDocs** folder, download the following files:

`btTBreg.csv`  
`btTBregHospitals.csv`  
`btTBreg_info.txt`

1. Load the `btTBreg.csv` data table into R.
2. The variables `cd41`, `cd42` and `cd41.sk`, `cd42.sk` measure the same variables (`cd4` and `cd4.sk` respectively) in the same individuals at two different time point. This means the data are in wide format. Reformat to long format.
3. Save the reformatted data into a file called `btTBregLong.tab` in such a way that
  - i. Columns are tab-separated.
  - ii. Column names are saved.
  - iii. No row number is saved in the resulting file.
4. Copy the code below to generate some wide-format data. We will assume this dataset contains observations of 2 biomarkers, `ferritin` and `rbp4` for 10 study participants at 2 different timepoints, `day1` and `day90`.

```

set.seed(123)

df<-data.frame(
  id=paste(sep="", "P", 1:10),
  ferritin_day1=rexp(10,rate=1/195),
  rbp4_day1=rexp(10,rate=1/2.5)
) %>%
  mutate(
    ferritin_day90=rnorm(10,mean=ferritin_day1+5,sd=4),
    rbp4_day90=rbp4_day1+rexp(10,rate=1/0.25)
  )

```

This is what this data table looks like:

	<b>id</b>	<b>ferritin_day1</b>	<b>rbp4_day1</b>	<b>ferritin_day90</b>	<b>rbp4_day90</b>
1	P1	164.474166	2.5120751	169.251422	3.021712
2	P2	112.439003	1.2005368	114.301474	1.552174
3	P3	259.165699	0.7025341	261.231686	0.829188
4	P4	6.157585	0.9427946	10.294123	1.007684
5	P5	10.961140	0.4707101	14.621489	1.119933
6	P6	61.717737	2.1244653	62.374941	2.431722
7	P7	61.274322	3.9080088	65.932629	4.105679
8	P8	28.327027	1.1969010	37.609469	1.354221
9	P9	531.616111	1.4773371	536.034536	1.790997
10	P10	5.684922	10.1025293	6.022743	10.249700

Reformat this to long format, i.e. so that you have 4 columns: **id**, **time**, **ferritin** and **rbp4**.

5. Load the **btTBregHospitals.csv** data table. Join the data frames storing **btTBreg.csv** and **btTBregHospitals.csv**.
6. Compute the average patient age and the proportion of male patients for each hospital.

Useful functions for this are **aggregate()** and **group\_by()**. You can however also do it manually.

7. Write an R function that computes the following summary statistics, then, using your custom function, compute these for the **bmi**, **cd41**, **cd42** columns:
  - i. mean
  - ii. median
  - iii. interquartile range
  - iv. minimum

- v. maximum
- vi. number of missing values

8. Do the same now, but only for female patients. Repeat for only male patients.