Exercise day 2 with solutions

Alessandra Meddis

2023-08-15

Exercise A: Reshaping the data

For this exercise we will work with a subset of "follicle" data, collected from patients with cancer that had OTC (ovarian tissue cryopreservation). Follicles were cultured for 8 days and the diameter was collected every 2 days. The aim of the study was to compare the follicles growth among different treatment groups.

Data set includes:

- Patient: patient ID
- Number: follicle ID
- Disease: type of disease
- Age: age at day 0
- Treatment: treatment group
- Day_x: follicle diameter at Day x

Question 1: Download the data and load them into R. Visualize the first lines of the data and a summary of the data.

1. How many observations?
2. what is the data type of each variable?

```
db_follicle<-read.csv("~/Desktop/KVN2021/Course/IntrotoR/data_exercise/follicle.csv")
knitr::kable(head(db_follicle))
```

Patient	Number	Disease	Age	Treatment	Day0	Day2	Day4	Day6	Day8
1	1	Breast_cancer	31.6	FBS	101.4590	112.1605	135.1980	160.2315	161.500
1	2	Breast_cancer	31.6	FBS	89.8315	141.3770	165.4925	NA	NA
1	3	Breast_cancer	31.6	FBS	90.2835	116.9870	122.6500	127.7305	129.447
1	4	Breast_cancer	31.6	FBS	120.3145	148.8840	166.9970	170.0245	170.740
1	5	Breast_cancer	31.6	FBS	93.0085	112.3135	120.8550	120.9000	120.940
2	6	Breast_cancer	29.4	FBS	83.9085	100.3520	112.6875	NA	NA

```
summary(db_follicle)
```

```
##     Patient         Number        Disease              Age
##  Min.   : 1.0   Min.   : 1.00   Length:70          Min.   :19.30
##  1st Qu.: 4.0   1st Qu.:18.25   Class :character   1st Qu.:29.40
##  Median : 7.5   Median :35.50   Mode  :character   Median :31.40
##  Mean   : 7.5   Mean   :35.50                      Mean   :29.74
##  3rd Qu.:11.0   3rd Qu.:52.75                      3rd Qu.:32.30
##  Max.   :14.0   Max.   :70.00                      Max.   :37.10
##
##   Treatment              Day0             Day2             Day4
##  Length:70          Min.   : 48.32   Min.   : 59.87   Min.   : 62.97
```

```
##  Class :character   1st Qu.: 62.74   1st Qu.: 81.50   1st Qu.: 96.96
##  Mode  :character   Median : 83.91   Median :105.03   Median :120.51
##                     Mean   : 84.70   Mean   :106.85   Mean   :127.01
##                     3rd Qu.: 95.56   3rd Qu.:118.05   3rd Qu.:147.24
##                     Max.   :194.47   Max.   :263.01   Max.   :299.65
##                     NA's   :3        NA's   :5        NA's   :11
##       Day6             Day8
##  Min.   : 71.34   Min.   : 72.48
##  1st Qu.:102.64   1st Qu.:103.78
##  Median :126.42   Median :129.45
##  Mean   :139.46   Mean   :146.24
##  3rd Qu.:172.66   3rd Qu.:176.93
##  Max.   :304.94   Max.   :318.88
##  NA's   :24       NA's   :29
```

We have a total of 70 observations and 11 Variables. Patient,Number,Age and Day_x are quantitative and are either numerical or integer, while Disease and Treatment are categorical variables and are listed as characters.

Question 2

Treatment and Disease are *characters*, can we understand from the summary how many different diseases are in the data?

1. Would it be better if Treatment and Disease were of a different type? If yes, which one?
2. Transform them into factor. You can use the function **factor()** in **R**.
3. Print the summary of the data, can you see any difference?

We discussed that for easier representation it is better to have factor for categorical variables, thus we can transform Disease and Treatment into factors.

```
db_follicle$Treatment<-factor(db_follicle$Treatment)
db_follicle$Disease<-factor(db_follicle$Disease)

summary(db_follicle)
```

```
##     Patient         Number                          Disease        Age
##  Min.   : 1.0   Min.   : 1.00   Brain_cancer            : 5   Min.   :19.30
##  1st Qu.: 4.0   1st Qu.:18.25   Breast_cancer           :40   1st Qu.:29.40
##  Median : 7.5   Median :35.50   Chronic_myeloid_leukemia: 5   Median :31.40
##  Mean   : 7.5   Mean   :35.50   Mb_Hodgkin              : 5   Mean   :29.74
##  3rd Qu.:11.0   3rd Qu.:52.75   Neurological_cancer     : 5   3rd Qu.:32.30
##  Max.   :14.0   Max.   :70.00   Rheumatoid_arthritis    : 5   Max.   :37.10
##                                 Sarcoma                 : 5
##  Treatment      Day0             Day2             Day4             Day6
##  FBS:55    Min.   : 48.32   Min.   : 59.87   Min.   : 62.97   Min.   : 71.34
##  hPL: 5    1st Qu.: 62.74   1st Qu.: 81.50   1st Qu.: 96.96   1st Qu.:102.64
##  HSA:10    Median : 83.91   Median :105.03   Median :120.51   Median :126.42
##            Mean   : 84.70   Mean   :106.85   Mean   :127.01   Mean   :139.46
##            3rd Qu.: 95.56   3rd Qu.:118.05   3rd Qu.:147.24   3rd Qu.:172.66
##            Max.   :194.47   Max.   :263.01   Max.   :299.65   Max.   :304.94
##            NA's   :3        NA's   :5        NA's   :11       NA's   :24
##       Day8
##  Min.   : 72.48
##  1st Qu.:103.78
##  Median :129.45
```

```
##  Mean   :146.24
##  3rd Qu.:176.93
##  Max.   :318.88
##  NA's   :29
```

The summary now shows the different levels of Treatment and Disease with the number of observations in that category.

Question 3:

1. How many patients are included in the study? How many follicles are collected for each patient? **Hint**: You can use the function *table()*

Looking at the summary we could see that the max ID for patient is 14, thus we would guess that there are 14 patients in the study. Moreover, we have 70 observations, thus 70 follicles, and if the number of follicles is the same for each patient we will have 70/14=5 follicles. We can check this by using the table function.

```
#number of observations
nrow(db_follicle)
```

```
## [1] 70
```

```
table(db_follicle$Patient)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14
##  5  5  5  5  5  5  5  5  5  5  5  5  5  5
```

- 5 observations (follicles) for each patient, for 14 patients*.

Question 4:

For each follicle the diameter was measured at day 0,2,4,6,8.

1. Are the data in a wide or long format?
2. Convert data from wide to long (or viceversa).
3. How many rows would we expect for each patient? Is it correct?

Data are in a wide format because we have one row for each follicle and several columns to indicate the measurement at different timepoints. We can transform it in a long format using the function ***reshape****.*

```
db_long<-reshape(db_follicle, direction="long",
                 idvar="Number",
                 varying=c("Day0","Day2","Day4","Day6","Day8"),
                 timevar="Day",v.names=c("diameter"))

db_long$Day<-factor(db_long$Day, levels=1:5, labels=c("0","2","4","6","8"))


table(db_long$Patient)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14
## 25 25 25 25 25 25 25 25 25 25 25 25 25 25
```

Question 5

We are interested in the follicle growth over time. We can calculate the diameter difference from time 0 at each time point:

1. Create a data.frame with Number (follicle ID) and the diameter at Day 0.
2. Rename the variable of diameter into *diameter0*
3. Do a join between this data.frame and the long version of your data set on *Number*
4. Create a new variable for the difference of diameter at each time point.

```
day0<-subset(db_long, Day==0)
day0<-day0[, c("Number","diameter")]
colnames(day0)<-c("Number","diameter0")

db_join=merge(db_long,day0, by="Number")

db_join$diam.change<-db_join$diameter-db_join$diameter0
```

Exercise B: descriptives