Solution Intro-2

Dr. Osama Mahmoud

05/07/2019

Exercise 1:

a) Set your working directory to a folder named `Practical Intro-2` stored at a path of your choice. Make sure that your working directory contains the data set file named `lbw.csv`.

```
# to set the working directory to the parent (up-level) folder, use setwd("../")
setwd("PATH_of_YOUR_Practical Intro-2_folder")
```

b) Load the `lbw` dataset into your R session:

```
lbw = read.csv("lbw.csv", header = T)
```

c) The data set `lbw` is a modified version of the Hosmer and Lemeshow data on birthweight. Its variables are as follows:

Variable	Description
id	Subject identification code
low	Logical: birthweight < 2500g
age	Maternal age at delivery (years)
lwt	Weight at last menstrual period (Kg)
race	Race
smoke	Logical: smoked during pregnancy
ptl	Premature labour history (count)
ht	Logical: maternal history of hypertension
ui	Logical: maternal uterine irritability
ftv	number of visits to physician during 1st trimester
bwt	birthweight (grams)

Now, show variable names of the `lbw` data frame.

```
names(lbw)
```

```
##  [1] "id"    "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"
##  [9] "ui"    "ftv"   "bwt"
```

d) Extract data for the first 180 subjects with only the `id`, maternal age at delivery, maternal smoking and the birth weight variables included. Assign the resulted subset to an object named `Reduced.lbw`

```
Reduced.lbw = lbw[1:180, c("id", "age", "smoke", "bwt")]
```

e) Add a new variable to `Reduced.lbw`, named as `yage`, to identify mothers with young age (< 19 years) at delivery.

```
Reduced.lbw$yage = Reduced.lbw$age < 19
```

f) Find out how many mothers identified as young and smoker.

```
table(Reduced.lbw$yage, Reduced.lbw$smoke)
```

```
##
##          0  1
##   FALSE 93 55
##   TRUE  18 14
```

g) Write down the names of the statistics or the role that each of these functions calculates or performs, by looking at the help file of the function:

```
attach(Reduced.lbw)
mean(bwt)
```

```
## [1] 2967.778
```

```
sd(bwt)
```

```
## [1] 739.2645
```

```
min(age)
```

```
## [1] 14
```

```
max(age)
```

```
## [1] 45
```

```
median(bwt)
```

```
## [1] 3037.5
```

```
IQR(bwt)
```

```
## [1] 1144.75
```

```
range(bwt)
```

```
## [1]  709 4990
```

```
summary(bwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     709    2406    3038    2968    3551    4990
```

```
detach(Reduced.lbw)
```

`attach()` and `detach()` add and remove a data object to and from the R search path making it unnecessary and compulsory respectively to specify the dataset name whenever\n you need to call one of its variables

h) Use one or more of the functions mentioned in the previous question to summarise the birth weight for smoker mothers and non-smoker mothers

```
tapply(Reduced.lbw$bwt, INDEX = Reduced.lbw$smoke, summary)
```

```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1021    2630    3175    3076    3640    4990
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     709    2353    2906    2794    3303    4238
```

i) calculate the correlation coeficient between (`age` and `bwt`)

```
cor(Reduced.lbw[["age"]], Reduced.lbw$bwt)
```

```
## [1] 0.07914953
```

Exercise 2:

a) Generate a vector `rand` consisting of 100 normally distributed values with a mean of 11 and standard deviation of 5.5 (`rand = rnorm(100, 11, 5.5)`). Make sure your vector is reproducible (i.e.repeated execution of your script should lead to identical vectors).

```
set.seed(110011)
rand = rnorm(100, 11, 5.5)
```

b) Find out, which (if any) elements of `rand` that are less than or equal the first quartile of the emprical distribution are positive.

```
(Index = which(rand <= quantile(rand, probs = 0.25) & rand > 0))
```

```
##  [1] 14 19 24 25 27 31 35 38 50 54 55 63 65 70 72 77 84 85 89 90 95 99
```

c) What are the values of the elements you identified in the previous question, if any?

```
rand[Index]
```

```
##  [1] 4.5457671 0.7246930 6.1289974 2.5755407 7.6722830 6.2240446 7.1654233
##  [8] 4.8385267 0.8003473 7.7767905 6.1681353 3.0817669 8.1052245 6.5033540
## [15] 2.4917633 4.1328145 7.3511675 8.0877797 6.5157785 3.3740306 5.2225340
## [22] 4.1972885
```