Data Summary in R

Sheherazade (R ladies Bogor)

02/07/2019

# Before!

* Always check which project and working directory you are working now

getwd()

## [1] "/Volumes/Sheherazade/syllabus/syllabus"

* All the dataset is in the folder used as a working directory
* All the materials are compiled from different sources:
  + McGarigal, K. Course Analysis of Environmental Data: Introduction to R. University of Massachusetts Amherst. www.umass.edu/landeco/teaching/ecodata/syllabus.ecodata.pdf
  + www.statisticssolutions.com/data-levels-of-measurement/
  + www.statmethods.net/stats/descriptives.html
  + www.r-bloggers.com/r-tutorial-series-summary-and-descriptive-statistics/
  + www.sthda.com/english/wiki/descriptive-statistics-and-graphics

# Data Summary in R

Variables -> named computer memory. For example, a list of species recorded stored in variable named ‘species\_name’.

pi <- 3.14 #real  
number\_species <- 327 #integer  
species\_name <- 'Panthera tigris' #character/string, should be entered in quotes  
rainforest <- TRUE #logical (only TRUE or FALSE)

## Data Structures

1. **vectors** : one-dimensional, a single data type (*e.g.* integer, real, and character/string).

example\_vector<-c(2,4,6,8,10)  
example\_vector

## [1] 2 4 6 8 10

(example\_vector<-c(2,4,6,8,10))

## [1] 2 4 6 8 10

example\_vector[2] #individal item is identified by the sequence (1-n)

## [1] 4

1. **matrices** : two dimensional, created by binding two or more vectors of the same type and length.

example\_vector\_2<-c(1,3,5,7,9)   
example\_matrix<-cbind(example\_vector,example\_vector\_2)  
example\_matrix

## example\_vector example\_vector\_2  
## [1,] 2 1  
## [2,] 4 3  
## [3,] 6 5  
## [4,] 8 7  
## [5,] 10 9

example\_matrix[4,2] #row 4, column 2

## example\_vector\_2   
## 7

example\_matrix[5,] #row 5

## example\_vector example\_vector\_2   
## 10 9

example\_matrix[,1] #column 1

## [1] 2 4 6 8 10

example\_matrix[] #all

## example\_vector example\_vector\_2  
## [1,] 2 1  
## [2,] 4 3  
## [3,] 6 5  
## [4,] 8 7  
## [5,] 10 9

example\_matrix[,1:2] #column 1 to 2

## example\_vector example\_vector\_2  
## [1,] 2 1  
## [2,] 4 3  
## [3,] 6 5  
## [4,] 8 7  
## [5,] 10 9

example\_matrix[,-1] #all except column 1. minus artinya kecuali

## [1] 1 3 5 7 9

1. **data frames** : one to multi-dimensional. Similar to spreadsheet. Column - field.

data1<-read.csv("data1.csv")  
head(data1)

## name nationality gender income age year\_firstboyfriend height  
## 1 a german female lower\_class 18 2010 189  
## 2 b china female lower\_class 25 2011 154  
## 3 c indonesia female lower\_class 32 2008 165  
## 4 d india female lower\_class 22 2009 166  
## 5 e malaysia female lower\_class 19 2010 16  
## 6 f indonesia female lower\_class 26 2019 153

1. **lists** : compound objects of associated data. It could consist of vector, matrix, and data frames.

example\_list<-list(example\_vector,data1)  
example\_list

## [[1]]  
## [1] 2 4 6 8 10  
##   
## [[2]]  
## name nationality gender income age year\_firstboyfriend height  
## 1 a german female lower\_class 18 2010 189  
## 2 b china female lower\_class 25 2011 154  
## 3 c indonesia female lower\_class 32 2008 165  
## 4 d india female lower\_class 22 2009 166  
## 5 e malaysia female lower\_class 19 2010 16  
## 6 f indonesia female lower\_class 26 2019 153  
## 7 g indonesia female lower\_class 18 2017 167  
## 8 h malaysia female lower\_class 18 2011 170  
## 9 i china female lower\_class 18 2011 156  
## 10 j german female lower\_class 34 2019 175  
## 11 k german female middle\_class 22 2016 180  
## 12 l british female middle\_class 22 2017 182  
## 13 m british female middle\_class 22 2010 173  
## 14 n australian female middle\_class 26 2019 185  
## 15 o malaysia female middle\_class 29 2009 173  
## 16 p malaysia male middle\_class 23 2014 150  
## 17 q malaysia male middle\_class 26 2010 167  
## 18 r malaysia male middle\_class 30 2011 168  
## 19 s china male middle\_class 18 2011 152  
## 20 t china male middle\_class 18 2011 161  
## 21 u china male upper\_class 18 2011 163  
## 22 v china male upper\_class 26 2011 167  
## 23 w indonesia male upper\_class 22 2017 156  
## 24 x malaysia male upper\_class 22 2018 171  
## 25 y indonesia male upper\_class 22 2018 170  
## 26 z indonesia male upper\_class 17 2019 168  
## 27 aa australian male upper\_class 34 2010 182  
## 28 bb australian male upper\_class 30 2010 179  
## 29 cc australian male upper\_class 34 2017 171  
## 30 dd australian male upper\_class 17 2019 184

names(example\_list)<-c('vector','data frame')  
example\_list

## $vector  
## [1] 2 4 6 8 10  
##   
## $`data frame`  
## name nationality gender income age year\_firstboyfriend height  
## 1 a german female lower\_class 18 2010 189  
## 2 b china female lower\_class 25 2011 154  
## 3 c indonesia female lower\_class 32 2008 165  
## 4 d india female lower\_class 22 2009 166  
## 5 e malaysia female lower\_class 19 2010 16  
## 6 f indonesia female lower\_class 26 2019 153  
## 7 g indonesia female lower\_class 18 2017 167  
## 8 h malaysia female lower\_class 18 2011 170  
## 9 i china female lower\_class 18 2011 156  
## 10 j german female lower\_class 34 2019 175  
## 11 k german female middle\_class 22 2016 180  
## 12 l british female middle\_class 22 2017 182  
## 13 m british female middle\_class 22 2010 173  
## 14 n australian female middle\_class 26 2019 185  
## 15 o malaysia female middle\_class 29 2009 173  
## 16 p malaysia male middle\_class 23 2014 150  
## 17 q malaysia male middle\_class 26 2010 167  
## 18 r malaysia male middle\_class 30 2011 168  
## 19 s china male middle\_class 18 2011 152  
## 20 t china male middle\_class 18 2011 161  
## 21 u china male upper\_class 18 2011 163  
## 22 v china male upper\_class 26 2011 167  
## 23 w indonesia male upper\_class 22 2017 156  
## 24 x malaysia male upper\_class 22 2018 171  
## 25 y indonesia male upper\_class 22 2018 170  
## 26 z indonesia male upper\_class 17 2019 168  
## 27 aa australian male upper\_class 34 2010 182  
## 28 bb australian male upper\_class 30 2010 179  
## 29 cc australian male upper\_class 34 2017 171  
## 30 dd australian male upper\_class 17 2019 184

## Data types

Understanding classification of data is a prerequisite for data summary and visualization, including to determine the correct statistical analysis. There are four levels of data measurement:

**1. Nominal data**  
Labels for a particular category, usually referred as categorical data. For example:

* Gender: male and female
* Religion: moslem, christian, hindu, budha, etc.
* Occupation: fisherman, businessman, etc.

In R, categorical data is represented as factors that contain levels.

You can check the structure of data using ‘str’ function.

data1<-read.csv("data1.csv")  
names(data1)

## [1] "name" "nationality" "gender"   
## [4] "income" "age" "year\_firstboyfriend"  
## [7] "height"

str(data1$gender)

## Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...

Gender is a factor, with two levels: male and female.

**2. Ordinal data**  
This is similar to nominal, but it has orders. For example:

* Level of income: lower class, middle class, and upper class
* Level of agreement: strongly disagree, disagree, neutral, agree, and strongly agree

In R, function factor() assigns order to nominal data.

factor\_income<-factor(data1$income,order=TRUE,levels=c("lower\_class","middle\_class","upper\_class"))  
str(factor\_income)

## Ord.factor w/ 3 levels "lower\_class"<..: 1 1 1 1 1 1 1 1 1 1 ...

Nominal and ordinal data are also reffered to as qualitative data.

**3. Interval data**

* has values of equal intervals that mean something (20-30 C and 40-50 C have 10 degree differences)
* no absolute zero

For example: Temperature (Celcius, Fahrenheit)

**4. Ratio data**

* like interval data but has absolute zero

For example: height, weight (no 0 kg, but there is 0 C)

Interval and ratio data are usually called as quantitative data, which may be classified as **discrete** (integer in R) and **continuous** (real in R).

## Why is this important?

* t.test -> continuous data (sugar concentration in tomatoes)
* ANOVA -> continuous data (mice weight compared among control, treatment 1, and treatment 2)
* linear modelling -> determine distribution (binary categorical -> binomial, count discrete -> poisson)

**So, understand your data first!**

## Other basic operations in R

### Missing values

example<-c(1,3,5,7,9,11)   
#delete number 9 or replace it with NA (if delete 9 does not work) for this set of codes  
example<-as.numeric(example)  
str(example)

## num [1:6] 1 3 5 7 9 11

str(example)

## num [1:6] 1 3 5 7 9 11

is.na(example) #to identify missing values

## [1] FALSE FALSE FALSE FALSE FALSE FALSE

example\_2<-example[!is.na(example)] #to use all of the vector except the missing values  
(example\_2<-example[!is.na(example)]) #or

## [1] 1 3 5 7 9 11

example\_3<-na.omit(example)  
example\_3

## [1] 1 3 5 7 9 11

(mean<-mean(example))

## [1] 6

(mean\_1<-mean(example,na.rm=TRUE))

## [1] 6

### Creating subsets

head(data1)

## name nationality gender income age year\_firstboyfriend height  
## 1 a german female lower\_class 18 2010 189  
## 2 b china female lower\_class 25 2011 154  
## 3 c indonesia female lower\_class 32 2008 165  
## 4 d india female lower\_class 22 2009 166  
## 5 e malaysia female lower\_class 19 2010 16  
## 6 f indonesia female lower\_class 26 2019 153

data\_new<-data1[,5:7] #select all rows and columns 5 through 7  
head(data\_new)

## age year\_firstboyfriend height  
## 1 18 2010 189  
## 2 25 2011 154  
## 3 32 2008 165  
## 4 22 2009 166  
## 5 19 2010 16  
## 6 26 2019 153

data\_new<-data1[,-c(1:4)]  
head(data\_new)

## age year\_firstboyfriend height  
## 1 18 2010 189  
## 2 25 2011 154  
## 3 32 2008 165  
## 4 22 2009 166  
## 5 19 2010 16  
## 6 26 2019 153

data\_new<-data1[1:6,] #select rows 1 through 6  
head(data\_new)

## name nationality gender income age year\_firstboyfriend height  
## 1 a german female lower\_class 18 2010 189  
## 2 b china female lower\_class 25 2011 154  
## 3 c indonesia female lower\_class 32 2008 165  
## 4 d india female lower\_class 22 2009 166  
## 5 e malaysia female lower\_class 19 2010 16  
## 6 f indonesia female lower\_class 26 2019 153

data\_new<-subset(data1,nationality=="german") #one of the most important functions  
head(data1)

## name nationality gender income age year\_firstboyfriend height  
## 1 a german female lower\_class 18 2010 189  
## 2 b china female lower\_class 25 2011 154  
## 3 c indonesia female lower\_class 32 2008 165  
## 4 d india female lower\_class 22 2009 166  
## 5 e malaysia female lower\_class 19 2010 16  
## 6 f indonesia female lower\_class 26 2019 153

head(data\_new)

## name nationality gender income age year\_firstboyfriend height  
## 1 a german female lower\_class 18 2010 189  
## 10 j german female lower\_class 34 2019 175  
## 11 k german female middle\_class 22 2016 180

write.table(data\_new,'german.csv',row.names = FALSE,sep=',')  
write.table(data\_new,'german2.csv',row.names = TRUE,sep=',')

### Packages in R

library(tools)  
library(HSAUR3)  
data("USairpollution",package = "HSAUR3")  
head(USairpollution)

## SO2 temp manu popul wind precip predays  
## Albany 46 47.6 44 116 8.8 33.36 135  
## Albuquerque 11 56.8 46 244 8.9 7.77 58  
## Atlanta 24 61.5 368 497 9.1 48.34 115  
## Baltimore 47 55.0 625 905 9.6 41.31 111  
## Buffalo 11 47.1 391 463 12.4 36.11 166  
## Charleston 31 55.2 35 71 6.5 40.75 148

str(USairpollution)

## 'data.frame': 41 obs. of 7 variables:  
## $ SO2 : int 46 11 24 47 11 31 110 23 65 26 ...  
## $ temp : num 47.6 56.8 61.5 55 47.1 55.2 50.6 54 49.7 51.5 ...  
## $ manu : int 44 46 368 625 391 35 3344 462 1007 266 ...  
## $ popul : int 116 244 497 905 463 71 3369 453 751 540 ...  
## $ wind : num 8.8 8.9 9.1 9.6 12.4 6.5 10.4 7.1 10.9 8.6 ...  
## $ precip : num 33.36 7.77 48.34 41.31 36.11 ...  
## $ predays: int 135 58 115 111 166 148 122 132 155 134 ...

Notes:

* HSAUR 3 is a functions, data sets, analyses and examples from the third edition of the book “A Handbook of Statistical Analyses Using R” (Torsten Horthorn and Brian S. Everitt, Chapman & Hall/CRC, 2014). [link](https://cran.r-project.org/web/packages/HSAUR3/index.html)
* USairpollution is the annual mean concentration of sulphur dioxide, in micrograms per cubic metre, is a measure of the air pollution of the city. This dataset is from packages HSAUR3 [link](https://www.rdocumentation.org/packages/HSAUR3/versions/1.0-9/topics/USairpollution)

# Descriptive statistics

This is a summary statistic that describe and summarize our data. It is different to inferential statistic (e.g. testing hypothesis). Summary statistics:

* mean
* var -> a measure of variability, the greater the variability in the data, the greater will be our uncertaintiy in the values of parameter.
* min
* max
* median
* quantile

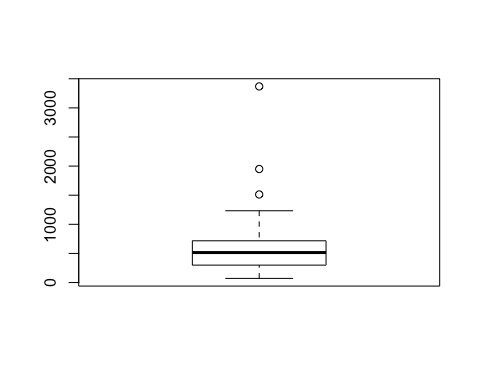
head(USairpollution)

## SO2 temp manu popul wind precip predays  
## Albany 46 47.6 44 116 8.8 33.36 135  
## Albuquerque 11 56.8 46 244 8.9 7.77 58  
## Atlanta 24 61.5 368 497 9.1 48.34 115  
## Baltimore 47 55.0 625 905 9.6 41.31 111  
## Buffalo 11 47.1 391 463 12.4 36.11 166  
## Charleston 31 55.2 35 71 6.5 40.75 148

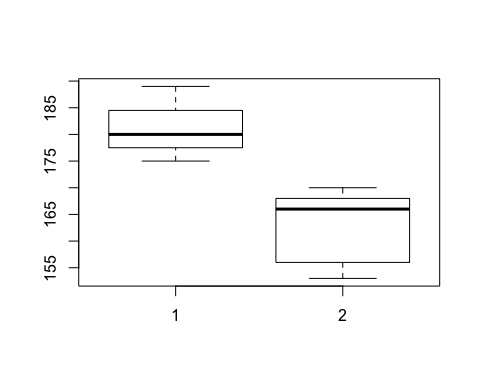
summary(USairpollution)

## SO2 temp manu popul   
## Min. : 8.00 Min. :43.50 Min. : 35.0 Min. : 71.0   
## 1st Qu.: 13.00 1st Qu.:50.60 1st Qu.: 181.0 1st Qu.: 299.0   
## Median : 26.00 Median :54.60 Median : 347.0 Median : 515.0   
## Mean : 30.05 Mean :55.76 Mean : 463.1 Mean : 608.6   
## 3rd Qu.: 35.00 3rd Qu.:59.30 3rd Qu.: 462.0 3rd Qu.: 717.0   
## Max. :110.00 Max. :75.50 Max. :3344.0 Max. :3369.0   
## wind precip predays   
## Min. : 6.000 Min. : 7.05 Min. : 36.0   
## 1st Qu.: 8.700 1st Qu.:30.96 1st Qu.:103.0   
## Median : 9.300 Median :38.74 Median :115.0   
## Mean : 9.444 Mean :36.77 Mean :113.9   
## 3rd Qu.:10.600 3rd Qu.:43.11 3rd Qu.:128.0   
## Max. :12.700 Max. :59.80 Max. :166.0

boxplot(USairpollution$popul)



data\_german<-subset(data1,nationality=="german")  
data\_indonesia<-subset(data1,nationality=="indonesia")  
boxplot(data\_german$height,data\_indonesia$height)



sapply(USairpollution,mean,na.rm=TRUE) #mean

## SO2 temp manu popul wind precip   
## 30.048780 55.763415 463.097561 608.609756 9.443902 36.769024   
## predays   
## 113.902439

sapply(USairpollution,sd,na.rm=TRUE) #sd, different to se?

## SO2 temp manu popul wind precip   
## 23.472272 7.227716 563.473948 579.113023 1.428644 11.771550   
## predays   
## 26.506419

library(pastecs)  
stat.desc(USairpollution)

## SO2 temp manu popul  
## nbr.val 41.0000000 41.0000000 41.00000 4.100000e+01  
## nbr.null 0.0000000 0.0000000 0.00000 0.000000e+00  
## nbr.na 0.0000000 0.0000000 0.00000 0.000000e+00  
## min 8.0000000 43.5000000 35.00000 7.100000e+01  
## max 110.0000000 75.5000000 3344.00000 3.369000e+03  
## range 102.0000000 32.0000000 3309.00000 3.298000e+03  
## sum 1232.0000000 2286.3000000 18987.00000 2.495300e+04  
## median 26.0000000 54.6000000 347.00000 5.150000e+02  
## mean 30.0487805 55.7634146 463.09756 6.086098e+02  
## SE.mean 3.6657530 1.1287796 87.99985 9.044226e+01  
## CI.mean.0.95 7.4087633 2.2813487 177.85432 1.827906e+02  
## var 550.9475610 52.2398780 317502.89024 3.353719e+05  
## std.dev 23.4722722 7.2277160 563.47395 5.791130e+02  
## coef.var 0.7811389 0.1296139 1.21675 9.515342e-01  
## wind precip predays  
## nbr.val 41.0000000 41.0000000 41.0000000  
## nbr.null 0.0000000 0.0000000 0.0000000  
## nbr.na 0.0000000 0.0000000 0.0000000  
## min 6.0000000 7.0500000 36.0000000  
## max 12.7000000 59.8000000 166.0000000  
## range 6.7000000 52.7500000 130.0000000  
## sum 387.2000000 1507.5300000 4670.0000000  
## median 9.3000000 38.7400000 115.0000000  
## mean 9.4439024 36.7690244 113.9024390  
## SE.mean 0.2231167 1.8384072 4.1396072  
## CI.mean.0.95 0.4509358 3.7155596 8.3664582  
## var 2.0410244 138.5693840 702.5902439  
## std.dev 1.4286442 11.7715498 26.5064189  
## coef.var 0.1512769 0.3201485 0.2327116

summary(data1) #whcih type of data that allow for mean and sd calculation?

## name nationality gender income age   
## a : 1 australian:5 female:15 lower\_class :10 Min. :17.0   
## aa : 1 british :2 male :15 middle\_class:10 1st Qu.:18.0   
## b : 1 china :6 upper\_class :10 Median :22.0   
## bb : 1 german :3 Mean :23.6   
## c : 1 india :1 3rd Qu.:26.0   
## cc : 1 indonesia :6 Max. :34.0   
## (Other):24 malaysia :7   
## year\_firstboyfriend height   
## Min. :2008 Min. : 16.0   
## 1st Qu.:2010 1st Qu.:161.5   
## Median :2011 Median :168.0   
## Mean :2013 Mean :163.8   
## 3rd Qu.:2017 3rd Qu.:174.5   
## Max. :2019 Max. :189.0   
##

# Homework

* Tidy up your own data into a dataframe
* Identify data type of each variable
* Run summary statistic
* Determine a research question that you want to answer using your data
* Create a plot of your data using the ggplot2

Send it to me [ssheherazade@wcs.org](mailto:ssheherazade@wcs.org) before August 5th, 2019!

An example:

**Name**: Sheherazade

**Description of data and research**: This research is about durian pollination by bats in Sulawesi.

**Dataframe**:

|  |  |
| --- | --- |
| treatment | no\_durian |
| control | 13 |
| control | 11 |
| treatment\_1 | 30 |
| treatment\_2 | 2 |

**Summary statistic**: mean, sd, and var at least

**mean\_control**: 12.5

**Research questions**:

1. Do bats pollinate durian? how treatment affects the number of durian produced?

**Plot**

Thank you!