F. FABRI
Practical Tests (R Software)
Promotion 2022 - ISAFE

Programming with R: datasheet 2
The basics of R programming computing

I The data.frame type

The data.frame class is a class used in R to contain data that is not necessarily of the same type. A data.frame type object can mix numeric, logical or character strings variables (quantitative and qualitative variables).

	UE	population	superficie (km^2)
France	Oui	66259012	643427
Suisse	Non	8061516	41277
Espagne	Oui	47737941	505370
Norvège	Non	5147792	323802
Belgique	Oui	11239755	30528
Allemagne	Oui	80996685	357022

• Let construct this table under R:
tableau=data.frame(UE=c("Oui",…,"Oui"),population=c(66259012, …,80996685),superficie=c(643427,…,357022),row.names=c("France",…,"Allemagne")) ; tableau

⏱ We notice that the individuals are classified by rows and the variables by columns. The call of an element of the table is done as for an element of a matrix table [2,3]

⏱ We can also call several elements of the table by means of a vector table [c (2,4), 1]

⏱ You can also call elements with the names given to variables or individuals.

tableau["France","UE"] ou *tableau["France",]* ou *tableau[,"UE"]*. Une commande équivalente à *tableau[,"superficie"]* consiste à écrire : *tableau$superficie.*

• We can add a column (a vector x) to a df object of type data.frame with:
x=seq(6)
df=data.frame(tableau, "nom_nouvelle_colonne"=x) ; df

• You can delete rows or columns from a df object of type data.frame with:
df=df[-2,] ou df=df[,-2] ou df=df[-c(2,3),]

⏱ You can use the function *subset(df,condition,col)* to find part of the data.frame. The *df* argument will indicate the data.frame variable to use. The condition argument, of type Boolean, will give a restriction on the individuals to select in *df* and the col argument will be a vector indicating the variables / columns to select in df. The col argument can be the number of columns to select or their name (*2:3* ou *c(2,3)* or *c("population","superficie")*), by default all columns are kept.

subset(tableau,tableau$UE=="Oui",1:2)

Applications :

1) What is the population of Spain?
2) What are the population and area of the first three countries?
3) What are all the data for France and Germany?
4) Show the population of countries in the EU with an area greater than 50,000km^2?
5) Add a column named "*Num*", made up of the first six integers to array.
6) Select in "table" only columns 2, 3 and 4 of the countries in the EU.

II Importing a text file

Copy the data from the Excel file in a text file.

Tmp=read.table("w:/ file_path / filename ",header=T,dec=",")

The option *header=T* allows you to specify that the first row corresponds to the name of the variables. The option *dec=","* allows you to specify the separator between the integer and decimal part of a number (usually it will be "," or "."). The option *sep* =” the field separator character. Values on each line of the file are separated by this character. The sep = "" (the default for read.table) the separator is ‘white space’, that is one or more spaces, tabs, newlines or carriage returns”.

Errors that do not allow data to be imported into R come from the fact that database is not cleaned. To clean the base, it will be necessary to

- remove spaces;
- change commas in numbers (-> unrecognized numeric variable);
- deal with missing values (replace with NA);
- …

Applications :

1) Test if the variable Taille (size) is numeric with the command *is.numeric(Tmp$Taille).*
2) Calculate the minimum given by the students for the 5th Group (Nombre5) number between 0 and 99.
3) Type the command *min(Tmp$Nombre5,na.rm=TRUE).*
4) Calculate the maximum weight (poids) given by the students.
5) What data is transmitted by M. LEGRAND ?
6) Give the characteristics of the teacher (*Poids* and *Taille*) transmitted by all students of the male D1 group (*Demi*). We will also indicate the name of the student.
7) Give the teacher weight (*poids*) estimates for all students in group D1.

III Basic functions in one-variable descriptive statistics

Functions	**Descriptions**
mean(x)	Returns the mean of the elements of vector x
var(x)	Returns the estimated variance of the population from the sample composed of the elements of the vector x
sd(x)	Returns the estimated standard deviation of the population from the sample composed of the elements of vector x
median(x)	Returns the median from the sample composed of the elements of vector x
quantile(x,p)	Returns the lower-order quantiles of all the probabilities of vector p from the sample composed of the elements of vector x
summary(x)	Return of the characteristics (minimum, first quartile, second quartile, mean, third quartile, maximum) for the series composed of the elements of the vector x (when x is a vector of numeric data)
summary(x)	Returns the counts of all modalities of factor x (when x is a vector of non-numeric data)
summary(df)	Return for all the variables composing the data.frame df information relating to a numeric or non-numeric variable.

cut(x,breaks=y)	Returns the membership class of the element of the series x according to the divisions contained in y

PS : The limits in following breaks *(a_1,a_2,a_3,a_4)* form closed open intervals. $a_1;a_2,a_2;a_3,a_3;a_4$
Data without a class will be counted in an NA class..

Applications :

1) Calculate the means of the estimates of the teacher's weight (poids) in group D1 (then D2).
2) Calculate the average of the first random numbers (*Nombre1*) of the students in group D2?
3) Estimate the variance of the size (*Taille*) of the teacher, we will take all the students.
4) Calculate the variance of the sample of the teacher's weight (*poids*), we will take all the students.
5) Calculate the median of the sample consisting of the first random number (*Nombre1*) between 0 and 99 given by the students (Men=*M*).
6) Calculate the deciles of the sample consisting of the first random number (*Nombre1*) between 0 and 99 given by the students (Men).
7) Calculate the coefficients of variation, skewness and kurtosis of the teacher's heights (taille) of all students.
8) Summarize the series by these main characteristics (function *summary*).
9) Display the division of the series made up of the second random number (*Nombre2*) between 0 and 99 given by the students (Men =M and Women=M) according to the following breakdown:

]-1 ; 9],]9 ; 19],]19 ; 29],]29 ; 39],]39 ; 49],]49 ; 59],]59 ; 69],]69 ; 79],]79 ; 89],]89 ; 99].

We can add class names (*"c0","c1",…)* by adding the argument *labels=c("c0","c1",…)* in the function *cut*. It will be necessary to respect that the number of classes obtained with breaks corresponds to the number of class names.
table(cut(…))

IV Basic functions in descriptive statistics with two variables

Functions	**Descriptions**
table(x,y)	Returns the contingency table between the x and y factors
cov(x,y)	Returns the covariance between x and y
cor(x,y)	Returns the correlation coefficient between x and y

Applications :

1) Give the contingency table for the factors Group (*Demi*) and Sex.
2) Calculate the covariance between the first random number (*Nombre1*) and the second (*Nombre2*).
3) Calculate the correlation coefficient between the first random number (*Nombre1*) and the second (*Nombre2*).