ISSN 2526-8910
Cad. Bras. Ter. Ocup., São Carlos, v. 27, n. 1, p. 217-230, 2019
https://doi.org/10.4322/2526-8910.ctoCB1625

Using Software R in research in occupational therapy

Maysa Marinho Antunes Ramos[a], **Pedro Luiz Ramos**[b], **Francisco Louzada Neto**[b], **Patrícia Carla de Souza Della Barba**[c]

[a] Universidade Federal de São Carlos – UFSCar, São Carlos, SP, Brasil.

[b] Universidade de São Paulo – USP, São Carlos, SP, Brasil.

[c] Departamento de Terapia Ocupacional, Universidade Federal de São Carlos – UFSCar, São Carlos, SP, Brasil.

Abstract: In this paper, it is presented a simple guide for researchers in occupational therapy to perform basic statistical analysis in a flexible and independent way, using the R software, that is a free open source software which its popularity has been increased considerably in many fields. We have presented a step-by-step guide about how to install such software, it is also discussed the necessary steps to include the data set and perform basic statistical analysis, such as the calculation of sample size, basic statistics, graphical presentation, hypothesis tests, and the linear correlation test. The dataset considered in this study comes from a research in occupational therapy and the topics considered are result of the common statistical procedures that were face in the course of the Post Graduation Program in Occupational Therapy at the Federal University of São Carlos (UFSCar), in which were possible to find the principal statistical procedures used by the researchers in applications.

Keywords: *Statistical Analysis, Software R, Occupational Therapy.*

Utilização do Software R em pesquisas na terapia ocupacional

Resumo: O presente artigo tem como objetivo fornecer subsídios para que pesquisadores em Terapia Ocupacional possam realizar procedimentos de estatística básica de maneira mais flexível e independente, a partir do sistema R, um software livre e gratuito, cuja popularidade vem aumentando consideravelmente no âmbito acadêmico. Apresentar-se-á, assim, o passo a passo de como instalar e utilizar o programa na realização de leitura e sumarização de dados, bem como no cálculo de estatísticas básicas, nas representações gráficas de dados quantitativos e qualitativos, no cálculo do tamanho amostral e na efetuação de testes de hipóteses e de teste de correlação linear. O banco de dados utilizado nas demonstrações apresentadas é oriundo de uma pesquisa em Terapia Ocupacional e as análises aqui realizadas resultam de uma demanda compreendida no decorrer de uma disciplina obrigatória do Programa de Pós-Graduação em Terapia Ocupacional da Universidade Federal de São Carlos (UFSCar), por meio da qual foi possível levantar os principais procedimentos estatísticos utilizados pelos pesquisadores em suas pesquisas.

Palavras-chave: *Análise Estatística, Software R, Terapia Ocupacional.*

Corresponding author: Maysa Marinho Antunes Ramos, Universidade Federal de São Carlos, Rodovia Washington Luís, s/n, CEP 13565-905, São Carlos, SP, Brasil, e-mail: maysamarinhoramos@gmail.com

Received on Jan. 14, 2018; 1st Revision on Apr. 24, 2018; 2nd Revision on Sept. 11, 2018; Accepted on Nov. 10, 2018.

1 Introduction

The responsibility of professionals to base their practices on scientific evidenceis one of the essences of research in occupational therapy since the choice of intervention tools and strategies depends on reliable precedents capable of ensuring a greater validity of the work to be performed (KIELHOFNER, 2006).

However, the systematic nature of scientific research implies a range of knowledge that goes beyond the knowledge common to each area, requiring the appropriation of external resources for the performance and understanding of sufficiently consistent studies (SAMPAIO; MANCINI; FONSECA, 2002).

Some of them are the statistical resources, strongly demanded and often avoided by researchers from other areas.

An article by Ottenbacher and Petersen (1985), published in *The American Journal of Occupational Therapy*, discussed the implications of the increasing use of quantitative procedures in the occupational therapy literature and revealed that

> the expansion of a research literature in the profession has been accompanied by an emerging sophistication in the use of research models and statistical analyzes (OTTENBACHER; PETERSEN, 1985, p.240).

According to Sampaio, Mancini and Fonseca (2002), occupational therapists need to be critical producers and consumers of information. However, there is often discouragement when there is a statistical topic since many of them reveals ignoring the section of statistical analysis when reading scientific articles (KIELHOFNER, 2006). Thus, whenever possible it is essential to engageto learn how to deal with these resources since they are available to help researchers and professionals in this arduous but essential journey to consolidate the profession.

Currently, software capable of generating statistics quickly and easily is found (KIELHOFNER, 2006). Also, more than practicality, it can be freedom and gratuitousness in the accomplishment of such procedures, since the technological barrier with the private software also prevents a greater contact with this essential step to many researches.

Recently, the use of free software has been intensified, including by the constant governmental incentive. Besides focusing on reducing costs, increasing competition and generating jobs, the government sees, above all, greater independence and collaboration in the production and dissemination of knowledge needed for the country's technological development. Data from the Federal Data Processing Service (Serpro) point to an economy of the Federal Government of approximately R$ 370 million with the use of free software in recent years and this number becomes significant when there are numerous other demands to be met (COSTA, 2009).

However, although there is a movement in favor of this software, still a lot of money is spent on licenses that, besides being expensive, they have expiration term. In the academic field, there is a strong dependence of the researchers in software paid for the accomplishment of statistical analyzes. In addition to the high cost attributed to them, its use is restricted to a few computers and spaces in laboratories that often cannot be accessed routinely by all.

As the options that have been disseminated in the scientific community to replace paid software, R (R CORE TEAM, 2018), stands out here a free, multiplatform and expandable software, which is gaining popularity in the academic field, and may exceed in the coming years, the use of paid software such as SAS, SPSS, Statistica, Minitab, among others. However, nothing prevents an institution or researcher from using paid software if they wish, but having other possibilities with more obvious advantages at their disposal and choose the one that best suits their needs.

Since there is no available literature that addresses the use of R in the scope of occupational therapy, this article aims to instruct researchers in the area to use this software to obtain basic statistics, giving them greater independence and scientific flexibility.

The database used in the demonstrations presented here comes from a research in occupational therapy titled *"The Ages and Stages Questionnaires Brazil (ASQ-BR) as an instrument for screening development in the context of early childhood education"* by Della Barba (2014). The main objective of this study was to analyze the performance of children attending early childhood education in a municipality in the interior of the state of São Paulo in an American development screening instrument. The analyzes performed here (reading and summarizing data, sample size calculation, hypothesis tests,and linear correlation test) are the result of a demand comprised in the course of a subject of the Post-Graduation Program in Occupational Therapy of the Federal University of São Carlos (UFSCar).

Therefore, the article is organized as follows:

In Chapter 1, the Software R, the steps for its installation and its layout will be shown. Chapter 2 will demonstrate how to read data in R and add other information. In Chapter 3, the data summarization will be treated, from the calculation of descriptive statistics (variance and standard deviation) to the development of graphical representations of both quantitative and qualitative data. In Chapter 4,the calculation of the sample size of simple random samples will be explained. Chapter 5 shows the commands for performing a hypothesis test in two populations. Finally, Chapter 6 will demonstrate how to perform calculations to verify correlations between the variables.

1.1 Software R

R was developed a priori by Ross Ihaka and Robert Gentleman and later added by collaborators from other parts of the world. R is a computational program directed to statistical and graphical operations widely demanded for the treatment, systematization, and dissemination of informative data (R CORE TEAM, 2018).

Since there are other programs with the same purpose, it is necessary to list the advantages attributed to the use of R to be an option differentiated from the other competitors.

First, it is free software, allowing the researcher to propose new subroutines and implement new methods of analysis according to their need. Second, it is free of charge, therefore it has no expiration time and can be used with more flexibility. Third, because it is multiplatform, it can be run by Windows, Macintosh, and Unix/Linux. Fourth, it is expandable, since it offers from the most basic to the most complex services, for example, new statistical techniques that are published in journals are usually accompanied by packages with functions implemented in R, enabling the access to such methodologies and apply them easily.

Thus, R's popularity with other programs is justified. The figure below shows that along with the decline of SPSS, there is the rise of R in the academic field (Figure 1).

1.1.1 Installation

Step 1 – Access the link available in FIOCRUZ (2019).

Step 2 - Choose the platform on which to run R.

Step 3 - Click on *"install R for the first time"*.

Step 4- Click on *"Download R 3.3.3 for"*.

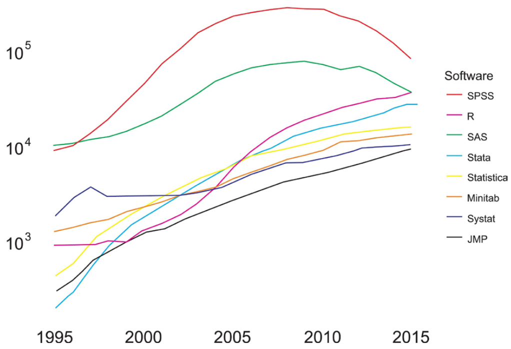


Figure 1. Number of accesses to different statistical software in google scholar.

After performing the procedures for installing the software, it is ready for use (Figure 2). It is important to note that when performing step 4, an updated version may be available. The software is constantly updated to accommodate new technologies; however, the procedures discussed here do not change for any version.

1.1.2 Layout

The R layout has a window entitled "console". It is a space in which the user will insert, change or save the data and the analysis codes to be performed (Figure 3).

It is suggested to open a new window called "new script" to facilitate the operational process so the data, as well as the commands, can be organized and transported to the console immediately, without typing errors, by the combination Ctrl R. To open a companion window just go to "file" and click on the "new script" option.

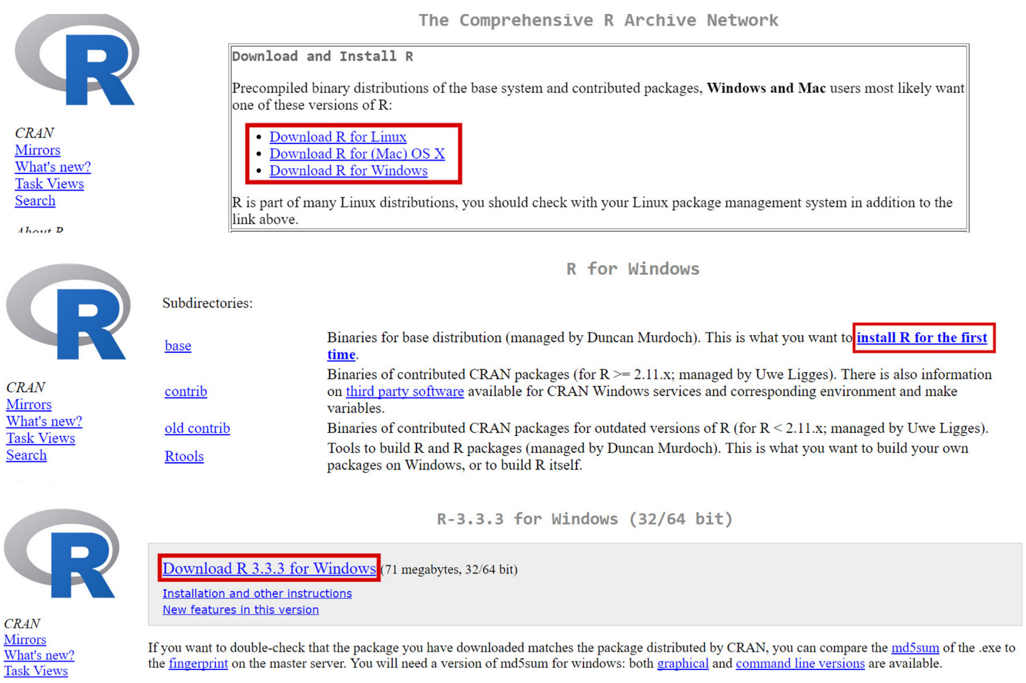
2 Data Reading

Here, how to read data in R will be shown (Table 1).

The X represents the male groups, the Y represents the female group and the addends 1, 2, 3, 4 and 5 are the categories: Communication, Broad Motor Coordination, Fine Motor Coordination, Problem Solving and Personal/Social, respectively.

In Brazil, the standard is to use the comma as the separation of decimal places, while the international standard is given by a dot. In this case, R uses the dot to define decimals and the comma to distinguish the elements.

For data reading, it is necessary to first insert them "manually" in the supplementary page:



The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Task Views
Search

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

R for Windows

CRAN
Mirrors
What's new?
Task Views
Search

Subdirectories:

base	Binaries for base distribution (managed by Duncan Murdoch). This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.11.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.11.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.

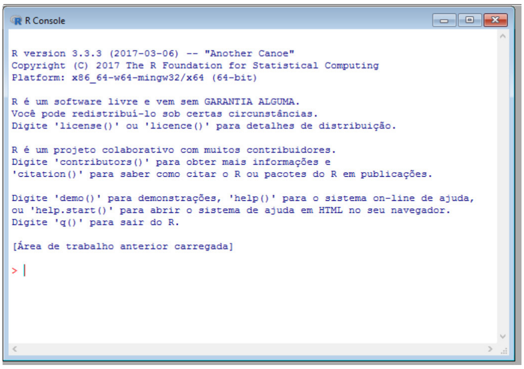
R-3.3.3 for Windows (32/64 bit)

CRAN
Mirrors
What's new?
Task Views

Download R 3.3.3 for Windows (71 megabytes, 32/64 bit)
Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.

Figure 2. Installation of R.



```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

[Área de trabalho anterior carregada]

> |
```

Figure 3. R Console.

```
x1=c(25,60,50,50,40,45,55,60,50,40,60)
x2=c(50,55,50,60,50,35,60,60,60,50,55)
x3=c(20,30,50,40,30,40,55,50,55,45,50)
x4=c(35,40,45,60,50,45,60,50,60,30,60)
x5=c(30,60,40,45,50,25,40,40,45,60,60)
y1=c(25,60,55,55,60,60,55,60,55,60)
y2=c(60,60,55,55,60,60,60,15,60,55)
y3=c(55,40,30,30,55,60,45,60,60,50)
y4=c(50,35,40,55,55,50,50,50,60,60)
y5=c(50,50,50,60,60,55,50,60,60,45)
```

To put additional information in the command, simply enter # and the information. The additional information is intended to signal, complement or differentiate the presented data and should be used whenever there is a need to specify something, for example through a title or a subtitle.

```
#Boys Communication Group 16-27 months old
#Girls Communication Group 16-27 months old
```

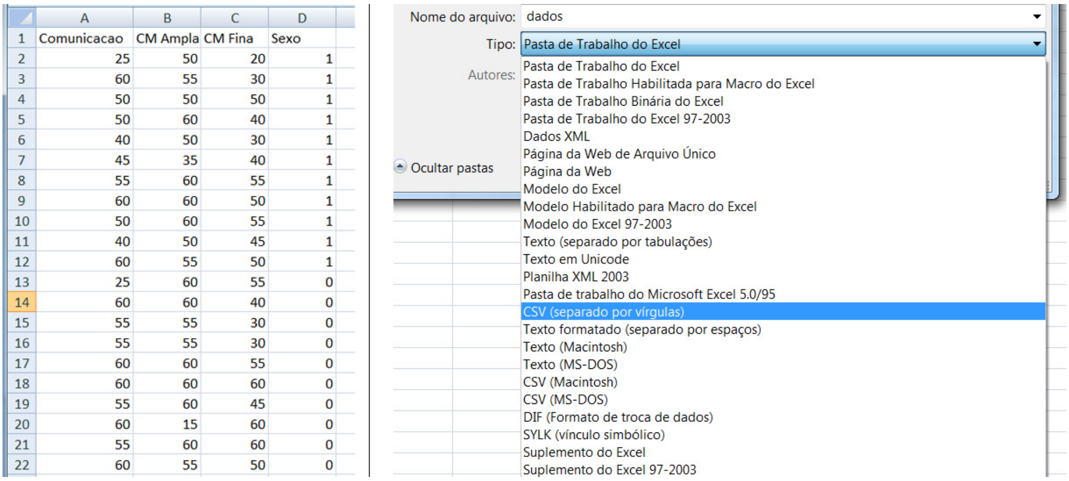
R also allows direct software reading such as Excel, Minitab, SPSS, among others. For example, supposing the interest is to read the data inserted in Excel (see Figure 4 left panel). A simple way to insert data without the need to install new packages is to save the data in .csv format. Excel will allow saving the data in this format (see Figure 4 right panel).

Finally, when saving the file inside a destination folder, for example, D:/, the reading is done as follows:

```
dados=read.csv(“D:/dados.csv”,header = TRUE,
sep=”;”)
> dados
Communication CM.Wide CM.Fine Gender
1 25 50 20 1
2 60 55 30 1
....
21 60 55 50 0
t1=data$Communication
> t1
[1] 25 60 50 50 40 45 55 60 50 40 60 25 60 55 55
60 60 55 60 55 60
x1=data$Communication[data$Gender==1]
y1=data$Communication[data$Gender==0]
> x1
[1] 25 60 50 50 40 45 55 60 50 40 60
> y1
[1] 25 60 55 55 60 60 55 60 55 60
```

Table 1. Scores in different performance areas obtained by children aged 16 to 27 months old in a municipality in the interior of São Paulo.

Gender	Category	Scores										
M	X1	25	60	55	55	60	60	55	60	55	60	
	X2	60	60	55	55	60	60	60	15	60	55	
	X3	55	40	30	30	55	60	45	60	60	50	
	X4	50	35	40	55	55	50	50	50	60	60	
	X5	50	50	50	60	60	55	50	60	60	45	
F	Y1	25	60	50	50	40	45	55	60	50	40	60
	Y2	50	55	50	60	50	35	60	60	60	50	55
	Y3	20	30	50	40	30	40	55	50	55	45	50
	Y4	35	40	45	60	50	45	60	50	60	30	60
	Y5	30	60	40	45	50	25	40	40	45	60	60



	A	B	C	D
1	Comunicacao	CM Ampla	CM Fina	Sexo
2	25	50	20	1
3	60	55	30	1
4	50	50	50	1
5	50	60	40	1
6	40	50	30	1
7	45	35	40	1
8	55	60	55	1
9	60	60	50	1
10	50	60	55	1
11	40	50	45	1
12	60	55	50	1
13	25	60	55	0
14	60	60	40	0
15	55	55	30	0
16	55	55	30	0
17	60	60	55	0
18	60	60	60	0
19	55	60	45	0
20	60	15	60	0
21	55	60	60	0
22	60	55	50	0

Nome do arquivo: dados

Tipo: Pasta de Trabalho do Excel

Autores:
Ocultar pastas
Pasta de Trabalho do Excel
Pasta de Trabalho Habilitada para Macro do Excel
Pasta de Trabalho Binária do Excel
Pasta de Trabalho do Excel 97-2003
Dados XML
Página da Web de Arquivo Único
Página da Web
Modelo do Excel
Modelo Habilitado para Macro do Excel
Modelo do Excel 97-2003
Texto (separado por tabulações)
Texto em Unicode
Planilha XML 2003
Pasta de trabalho do Microsoft Excel 5.0/95
CSV (separado por vírgulas)
Texto formatado (separado por espaços)
Texto (Macintosh)
Texto (MS-DOS)
CSV (Macintosh)
CSV (MS-DOS)
DIF (Formato de troca de dados)
SYLK (vínculo simbólico)
Suplemento do Excel
Suplemento do Excel 97-2003

Figure 4. Data insertion in Excel.

The read.csv command is used to read the file, “D:/data.csv” is the folder and the name of the saved file, header = TRUE is case where the table has a header if there is no switch to header = FALSE,esep=”;” displays the form that separates the information from the .csv file (Excel saves it like this). If the interest is to declare the information separately as described in the previous example, the command is x1=data$NAME, where NAME is the name of the variable defined in the header.

It is important to highlight that here the variables were not initially separated between male and female and that the Gender column is used to discriminate the gender. If the interest is to work with x1 and y1 separately, the term data$NAME[data$Gender==1] can be used to select only male and change 1 by 0 to select only female.

3. Data Summarization

The process of summarizing and describing the results is important because, from this process, the data will be organized and presented to the reader. To think about this stage implies the use of increasingly understandable resources since communication in the scientific field is fundamental for the dissemination of information and for the construction of new knowledge (KIELHOFNER, 2006).

3.1 Measures of central tendency and variability

Basic statistics such as mean, variance and standard deviation are the most used since they allow having a measure of central tendency and a measure of dispersion. It should be noted that the commands are linked to the English term, for example, “var” is linked to the English term “variance”, and “sd” is linked to the term “standard deviation”.

Commands to calculate such statistics:

```
mean(x1)
[1] 48.63636
var(x1)
[1] 115.4545
sd(x1)
[1] 10.74498
```

Also, the minimum, 1stquartile, median, 3rd quartile, and the maximum can be obtained:

```
summary(x1)
Min. 1st Qu. Median Mean 3rd Qu. Max.
25.00 42.50 50.00 48.64 57.50 60.00
```

3.2 Chart representation

Getting charts in the R is very simple, whether they are generated from quantitative or qualitative data. However, there are appropriate charts for each case, but only the most common ones will be presented here.

3.2.1 Quantitative data

For quantitative data, the histogram and boxplot are considered.

Command to generate the histogram:

```
#Left Figure
hist(x1)
#Right Figure
hist(x1,main=”Communication Group 16-27
months old”,xlab =”Score”, ylab=”Frequency”,
col=”deepskyblue4”)
```

In the first case, using the command hist(x1), the graph of interest was obtained. However, customizations are possible by adding information such as title (main), information on the axis x (*xlab*), or the frequency on the axis y (*ylab*) and different colors (col). To change the color of the charts just change the name in the command. The name of the available colors can be accessed at Columbia University (2019).

For the boxplot, on the horizontal axis, there is the factor (s) of interest and on the vertical axis, there is the variable to be analyzed. The boxplot is a very informative graph since it allows locating the distribution of the data, the variability, the symmetry or asymmetry, it provides a criterion for the identification of extreme values and, finally, it allows the comparison of these results for different groups. The first line of the box represents the first quartile, the second line is the median, and the third line is the third quartile. The vertical line connected to the top represents the maximum and the lower line is the minimum. Thus, the boxplot can be used to visualize whether or not certain sets of data have equivalence.

Command to generate the boxplot:

```
#Left Figure
boxplot(x3,y3)
#Right Figure
boxplot(x3,y3,col=c(”deepskyblue4”,”gray70”),
main=”Boxplot for
Fine Motor Coordination”, names=c(”Male”,
“Female”), ylab=”Score”,
horizontal=FALSE)
```

To generate the boxplot figure it is necessary to use the *boxplot()*command (Figure 5). If the interest is viewing only variable *x*3, the*boxplot(x3)* can be used, and *boxplot(x3, y3)*is used for two variables. Colors can be customized through *col*, while the *names* discriminate which variables are under study. Finally, the *horizontal* argument is used to inform if there is a need the *boxplot* to be displayed vertically (*horizontal*=FALSE) or horizontally (*horizontal*=TRUE).

3.2.2 Qualitative data

For qualitative variables, consider the *pie* chart and the *bar* chart.

Command to generate the pie chart:

```
pie(c(length(x1),length(y1)))
pie(c(length(x1),length(y1)),
col=c(“deepskyblue4”,”gray70”),
labels=c(”Male”, “Female “), main=” Gender “)
```

In this case, there are the following commands: generate pie chart (*pie*), select the colors (*col*) and define the labels (*labels*) and the title of the figure (*main*). The *length* command is used to calculate the frequency of occurrence of the variable; however, the values can be enter manually if interested. Lastly, only two categories have been used here, but can be extended to as many as necessary, including a comma after the last item of each argument followed by their information.

If the data were entered directly by Excel and the interest was to work directly with *data$Gender*, the chart could be obtained as follows:

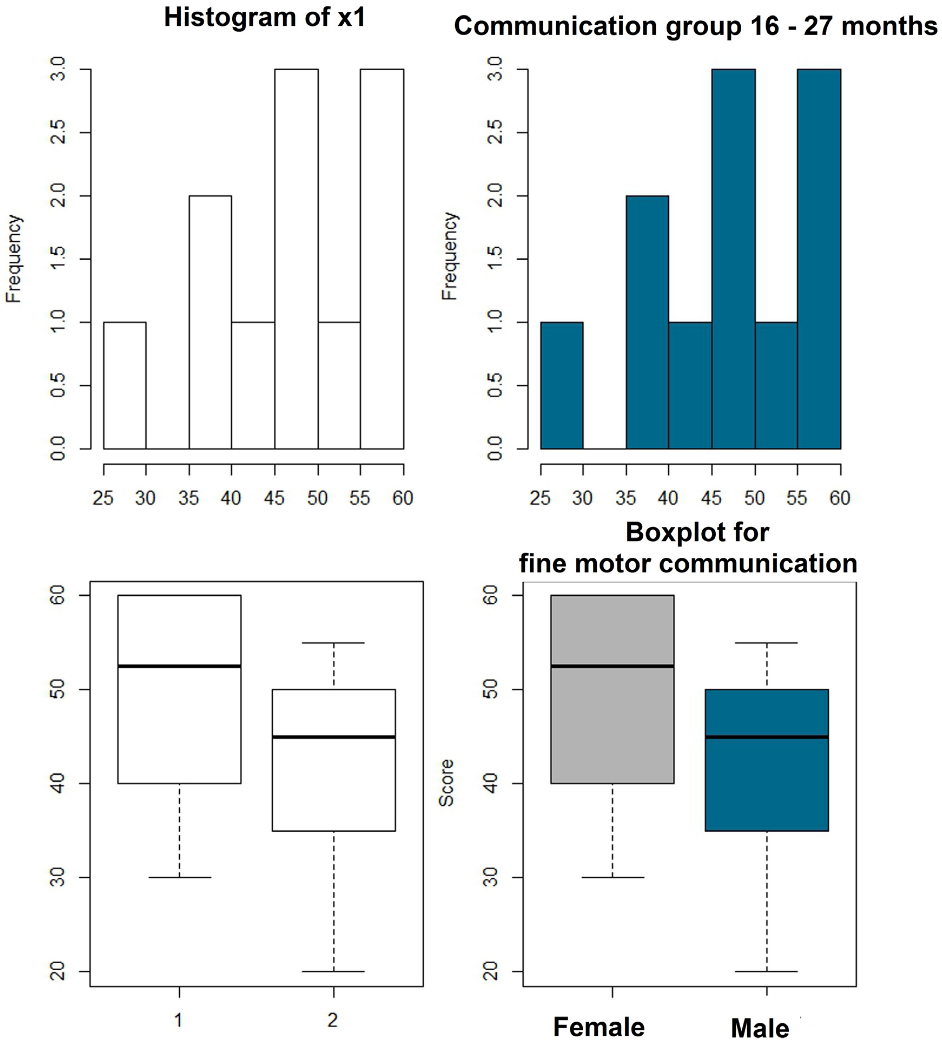


Figure 5. Histogram and Boxplot of the score Fine motor coordination for the boys and girls groups 16-17 months old.

```
pie(c(sum(data$Gender),length(data$Gender)-
sum(data$Gedner)))
pie (c(sum(data$Gender),length(data$Gender)-
sum(data$Gender)), col=c(”deepskyblue4”,
“gray70”),labels=c(”Male”, “Female”),
main=”Gender”)
```

The *sum()*calculates the sum of all elements. As the male was defined as *1* and the female as *0*, the sum of all values will return the number of boys in the sample while the complement will bring the total number of girls.

The command to generate the *bar*chart is:

```
barplot(c(length(x1),length(y1)), main=”
Number of students divided by gender”,
col=c(”deepskyblue4”,”gray70”),
names.arg=c(”Male”, “Female”))
```

As in the previous example, the bar graph is generated through the *barplot*, while the categories are named through the *names.arg* (Figure 6).

4. Calculation of Sampling Size

In any study, the margin of error and the level of confidence are closely related to the size of the sample (BUSSAB; MORETTIN, 2010). Thus, when there is no possibility of working with all the population of interest, whether for limited time, high costs, ethical issues, or other limitations, it is necessary to use sampling techniques.

The *calculation of sample size* is an important step of the research. Problems during this step

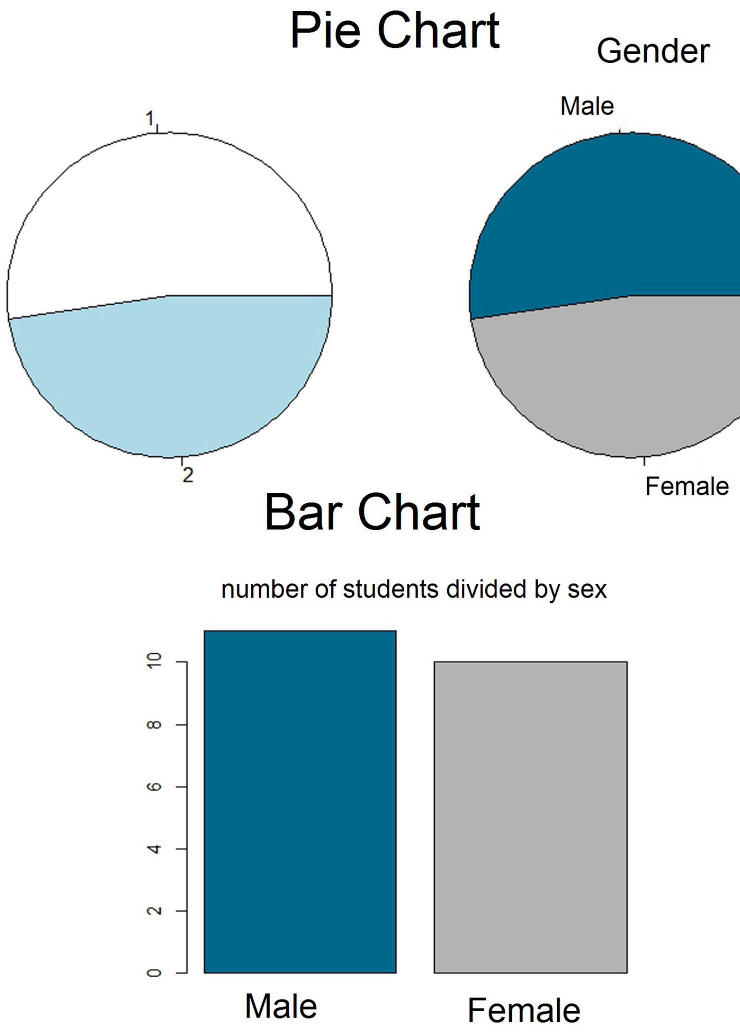


Figure 6. Distribution of the children´s gender.

may compromise the analysis and interpretation of results (MIOT, 2011).

Only the calculation of the sample size of simple random samples will be presented here. First, the *level of significance (α)* and the *margin of error (ε)* must be defined. The *level of significance* is used to construct confidence intervals, showing the percentage of all possible samples that satisfy the *margin of error*, while the *margin of error* reveals how close to the sample is to the population parameter.

When taking as a true parameter the proportion of 4 to 60-month-old children with typical development and attending kindergartens and pre-schools in a city in the interior of São Paulo, Della Barba (2014) defined that the *margin of error* would be ten percent points and that the confidence interval would be 95%. This means that if 80% of the children have typical development in the sample, this group should have between 70% and 90% of children with typical development. Also, if the confidence interval is 95%, it is possible that five out of 100 surveys performed, respecting the same methodological procedures, present a result out of range.

4.1 Sample size for a population

The calculation of the sample size can be performed in four situations, considering the following qualitative variables:

1) Qualitative variable with population less than 10000.

2) Qualitative variable with population greater or equal to 10000.

A routine (*tal*) is shown below from which the sample size for qualitative data will be possible to calculate, that is, when the interest is to estimate the proportion of some characteristic in the population. The function arguments are: *alpha*, which is the level of significance, *E*, which is the assumed error (values between 0 and 1), *N*, which is the population size (if N is greater than 10000, it is not necessary to enter it), and *p*, which is the proportion obtained earlier in a pre-test or previous work. The last two elements should only be inserted when there is this type of I nformation (BUSSAB; BOLFARINE, 2005).

```
ta1<-function(alpha,E,N=NA,p=0.5){
z<-qnorm(1-alpha/2,0,1)
if(is.na(N)) { n<-(z*sqrt(p*(1-p))/E)^2 } else {
n<-(N*p*(1-p)*(z^2))/(((N-1)*(E^2))+(p*(1-
p)*(z^2))) }
return(round(n))}
```

When entering the above command, the command is generated to calculate the sample size that will depend on the inclusion of information about *alpha, E, N,* and *p*:

```
ta1(alpha,E,N,p)
```

According to Miot (2011), it is extremely important to perform a pre-test with 30 to 40 individuals and to consider the results as a population estimate for the calculation of sample size. For quantitative variables, the pre-test is necessary and for qualitative variables, it is optional. In this case, the fact that the proportion that will lead to the largest sample size among all proportion estimates is p = 0.5can be used. This result is the most conservative and is commonly applied in the absence of a pre-test. Although a pre-test involves an additional step in the work, it can decrease by more than four times the sample size (in relation to the conservative interval), allowing cost cutting and greater agility in the tabulation and in obtaining the results.

Example 1: Assuming the interest is to find the proportion of children in the risk zone of development.

a) Calculate the sample size needed to estimate this proportion with a significance level of 0.05 (95% confidence) and a margin of error of 0.03.

```
ta1(alpha=0.05,E=0.03)
[1] 1067
```

In this case, a sample of 1067 people would be required to estimate this proportion, considering 95% confidence and a margin of error of 3 percentage points.

b) Assuming that a pre-test was performed and 6% of the children were found to be in the risk zone. What would be the sample size needed under the same assumptions as in the previous example?

```
ta1(0.05,0.03,0.06)
[1] 241
```

It is noticed that the pre-test performed allowed a significant reduction of the sample size.

c) Assuming that the interest is to estimate the proportion of children who are in the risk zone in a small town and assuming the population number is 1000 children, calculate the sample size under the same assumptions.

Without pre-test data:

```
ta1(0.05,0.03,1000)
[1] 516
```

With pre-test data:

```
ta1(0.05,0.03,1000,0.06)
[1] 194
```

Without a pre-test, there would be 516 children required to perform the experiment. With a pre-test indicating 6% of the population in the risk zone, the sample size would be 194 children.

In cases related to quantitative variables, that is, when the interest is to estimate the average of some characteristic in the population, there is the following possibilities:

3) Quantitative variable with population less than 10,000.

4) Quantitative variable with population greater or equal to 10000.

The following routine must be entered in the R before performing the calculations. It uses the *alpha* values, the *level of significance* considered, the *E* that is the margin of error of the estimate, the *sigma*, the *standard deviation* of the variables of interest and the *N* that is the size of the population (BUSSAB; BOLFARINE, 2005).

```
#Function for quantitative data
ta2<-function(alpha,E,sigma,N=NA){
z<-qnorm(1-alpha/2,0,1)
if(is.na(N)) { n<-(z*sigma/(100*E))^2 } else {
n<-(N*(sigma^2)*(z^2))/
(((N-1)*((100*E)^2))+(sigma^2*(z^2))) }
return(round(n))}
```

The function to be called in R will then be:

```
ta2(alpha,E,sigma,N=NA)
```

Unlike the previous case, there is no conservative choice for *sigma* and such a result needs to be obtained through a pre-test or related research. The *N* value must be entered for small populations (N<10000). If it is not inserted, the program will consider it complementary (N≥ 10000).

Example 2: If the objective is to describe the fine motor coordination score of children 16-27 months old in a city, the data presented in Table 1 can be used as a pre-test. In this case, the *sigma is 11.77*. Assuming a significance level of 0.05 and an error of 0.02:

a) Calculate the required sample size, assuming that the population of children in this age group is 4000.

```
ta2(0.05,0.02,11.77,4000)
[1] 129
```

There were 129 children required to find an estimate of the population score with the level of significance, standard deviation, and margin of error assumed. It is important to note that as the variability of the data set (*sigma*) increases, the larger the sample size needed to find the population estimate.

b) Calculate the required sample size assuming that the child population is 30,000.

```
ta2(0.05,0.02, 11.77,30000)
[1] 132
```

In this case, as N> 10000, the last item should not be filled.

```
ta2(0.05,0.02, 11.77)
[1] 133
```

By inserting or not the information, similar results are obtained. However, the second case is the correct one. The methods presented here for sample size calculation are used for basic purposes of estimating a ratio or a mean. For other purposes, such as comparing proportions or means of several groups, overlapping patterns over time, among others, it is important to consult a statistician because each method of analysis has different formulas for calculating sample sizes.

It is also important to note that sample size calculations assume that the selection process of individuals is random, which in practice is not always the case. Therefore, it is always important to discuss and weigh each case.

5 Hypotheses Testing in Two Populations

Hypothesis testing is a statistical inference method used to evaluate unknown population parameters, through the evidence that a sample provides (MIGON; GAMERMAN; LOUZADA, 2014).

This method is widely used to validate clinical research when the main interest is, for example, to verify if there was a difference between a standard treatment and an alternative treatment. This method also allows answering several questions about the parameters of interest, such as mean, proportion, and variance and etc. Thus, two hypotheses are assumed, the null *(*H_0*)*and the alternative H_A and make a decision to accept or reject the *null hypothesis*. Although it is more common to define the *alternative hypothesis* as the one proposed by the researcher and the *null hypothesis* as the complement of the alternative hypothesis, in the *R* it is not necessary to define the *null* and *alternative hypotheses*. The program itself already does this, it is only necessary to verify what were the assumptions made by *R*.

5.1 Hypothesis of normality

Several statistical tests use the assumption that the data comes from a normal distribution, giving great tools for hypothesis testing.

There are a large number of tests in the literature to verify the assumption of normality. When comparing several tests Yap and Sim (2011) concluded that, in general, the *Shapiro-Wilk* test is more sensitive for detection or not of normality. It can be done through the following command:

```
shapiro.test()
```

The hypotheses to be tested will be:

$$\begin{cases} H_0 \text{ The sample comes from a normal distribution} \\ H_a \text{ The sample does not come from a normal distribution} \end{cases}$$

When making a decision regarding a hypothesis, the following errors can be made: Type I error is the rejecting H_0 when in fact a H_0 is true. On the other hand, the type II error can be made, which is not to reject H_0 when in fact the H_0 is false. The probability of occurrence of the type I error is denoted by α, known as the level of significance. Such a value is usually determined by the researcher before data collection. In several applications, the level of significance assumed is 0.05. A simple rule to follow for decision making is to consider the p-value. The p-value can be seen as the probability of obtaining a equal or more extreme test statistic than that observed, by means of a population sample considering H_0. as true. Thus, after defining the level of significance, $H_0$is rejected if the p-value is less than α or $H_0$is not rejected if the p-value is greater than α.

Thus, for the data set *x3*, by applying the *Shapiro.test(x3)* function, there are the following results:

```
>shapiro.test(x3)
Shapiro-Wilk normality test
data: x3
W = 0.905, p-value = 0.2125
```

This command outputs the test statistic (W) and its associated p-value (p-value). As defined $\alpha = 0.05$, the decision rule will be if p-value <α, we reject H_0, that is, the data does not come from a normal distribution. If p-value < α, H_0, is rejected, that is, the data comes from a normal distribution. Then, as 0.2125>0.05 with a significance level of 0.05, there is no evidence that the data does not follow this distribution. It is also necessary to verify the normality assumption for the other variable *y3* to be compared.

```
>shapiro.test(y3)
Shapiro-Wilk normality test
data: y3
W = 0.85874, p-value = 0.07374
```

Analogous to *x3*, as 0.07374>0.05, then with a significance level of 0.05, the data comes from a normal distribution.

5.2 Hypothesis of accepted normality

Once the normality assumptions for both variables were verified, the t-test was performed to compare the means between the two groups (*x3* and *y3*), enabling to state (at a pre-defined level of significance) whether the means differ statistically or not. However, the tests will be different if the group variances are the same or different. Therefore, before performing such a procedure, the equality or difference of the variances must be verified. For this, another hypothesis test known as *F test* is used. In this case, there are the following assumptions:

$$\begin{cases} H_0 \text{ The samples have equal variances} \\ H_a \text{ The samples have different variances} \end{cases}$$

From there, the same decision rule adopted in the previous section is used. Assuming a significance level of $\alpha = 0.05$, the decision rule will be if p-value < α reject H_0, that is, the samples have different variances. If p-value> α does not reject H_0, that is, the samples have equal variances. The command for *F test*is:

```
>var.test(x3,y3,conf.level = 0.95)
F test to compare two variances
data: x3 and y3
F = 0.9472, numdf = 10, denomdf = 9, p-value =
0.926
alternative hypothesis: true ratio of variances is not
equal to 1
```

As *p-value*> 0.05, H_0 is not reject, that is, the hypothesis test presents evidence that there is no significant difference between the variances.

Having verified the assumption of equality or difference between the variances of *x3* and *y3*, the *t-test* for the comparison of means is considered:

```
t.test(x, y, paired = FALSE, var.equal=FALSE,
conf.level = 0.95)
```

If *x* is the first group and*y*is the second group, there is *var.equal* if the variances are equal or not (TRUE or FALSE) and *conf.level* the confidence interval (1-α). If the significance level is 0.05, then the *conf.level*=1-0.05=0.95. If no information is entered, *R*will default to 0.95. The paired argument refers to the case where the data is paired. A paired sample means that each observation of the first sample is related to the respective observation of the second sample.An example would be to consider a measure that is observed pre and post-test in the same individual. If the data is paired, it should be set to *paired* = TRUE.

> **Example 1:** Perform a hypothesis test to see if there is a difference between the mean of boys and girls 16-27 months old for the variable fine motor coordination at a significance level of 0.05.

```
t.test(x3,y3, paired = FALSE, var.equal = TRUE,
conf.level = 0.95)
Two Sample t-test
data: x3 and y3
t = -1.2236, df = 18.696, p-value = 0.2363
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:-16.890851 4.436305
```

The hypothesis to be tested here is *equality versus difference* between the groups, where the *null hypothesis* is *equality*. Therefore, using the decision rule already defined as 0.236>0.050, with a significance level of 0.05, H_0 is not rejected, that is, there is no difference between the means. Although the mean number of boys is *42.27* and the number of girls is *48.5*, when performing the hypothesis test, it cannot be said that both differ at a significance level of 0.05.

Example 2: Assuming that the interest is to perform a hypothesis test to verify if there is a difference between the mean score of the boys and girlsfor the personal/social variable at a significance level of 0.05.

```
t.test(x5,y5, paired = FALSE, var.equal = TRUE,
conf.level = 0.95)
Two Sample t-test
data: x5 and y5
t = -2.2536, df = 14.66, p-value = 0.03999
alternative hypothesis: true difference in means is
not equal to 0
```

In this case, using the same decision rule, there is*0.03999<0.05*, with a significance level of 0.05, then H_0 is rejected, that is, there is a difference between the means. However, such results are not valid since the normality of *y5* was not tested here.

```
shapiro.test(y5)
Shapiro-Wilk normality test
data: y5
W = 0.82495, p-value = 0.02909
```

As *0.02909 <0.05*, then with a significance level of 0.05, it cannot be said that the data come from a normal distribution (remember that one can accept normality hypothesis if *p-value*>0.05). In this case, another type of statistical test should be used, which will be discussed next.

5.3 Hypothesis of rejected normality

When the data does not have a normal distribution, the previous tests should not be applied. An alternative is to consider the *Wilcoxon-Mann-Whitney* test (WMW) to compare the distributions of the two groups. In this case, the comparison is not done directly through the mean between groups, but through the distribution of the data.

Unlike the previous procedure, the WMW test does not require the assumption of equality or difference between the variances.

Command for the *WMW* test:

```
wilcox.test(x, y, paired = FALSE, conf.level =
0.95)
```

Example 3: Consider the previous example (where *y5* does not come from a normal distribution) and verify if there is a difference in the mean scores of the boys and girls for the personal/social variable at a significance level of 0.05.

```
wilcox.test(x5,y5,paired = FALSE, exact=FALSE,
conf.level = 0.95)
Wilcoxon rank sum test with continuity correction
data: x5 and y5
W = 28, p-value = 0.05497
alternative hypothesis: true location shift is not
equal to 0
```

The hypothesis to be tested here is *equality versus difference* between groups where the *null hypothesis* is *equality*. Therefore, using the decision rule already defined *p-value*>0.05, with a significance level of 0.05, H_0 is not rejected, that is, there is no difference in the mean scores of boys and girls for the personal/social variable, at a significance level of 0.05.

It is important to emphasize that when the *t-test* is used erroneously, it leads to rejection of the equality hypothesis, whereas the WMW test did not lead to rejection of the equality hypothesis. In this way, the assumptions of normality and equality of variance should always be checked.

6 Linear Correlation

In several researches, there is an interest in establishing relationships between variables, but more than that, they must be relevant and not mere coincidence. The website available in Vigen (2019) has some strange correlations (Figure 7).

In this way, it is concluded that one solution to lower the divorce rate in Maine would be to eliminate per capita margarine consumption. It is known that there is no relationship between these variables and this high correlation is just a mere coincidence. Therefore, during the construction of the hypotheses to be verified, the researcher should use common sense.

6.1 Calculation of Pearson's linear correlation coefficient

Thus, there is an *explanatory variable (x)* and a *response variable (y)* to be found. The *explanatory variable* seeks to explain the response variable, while the *response variable* is the answer itself. To know

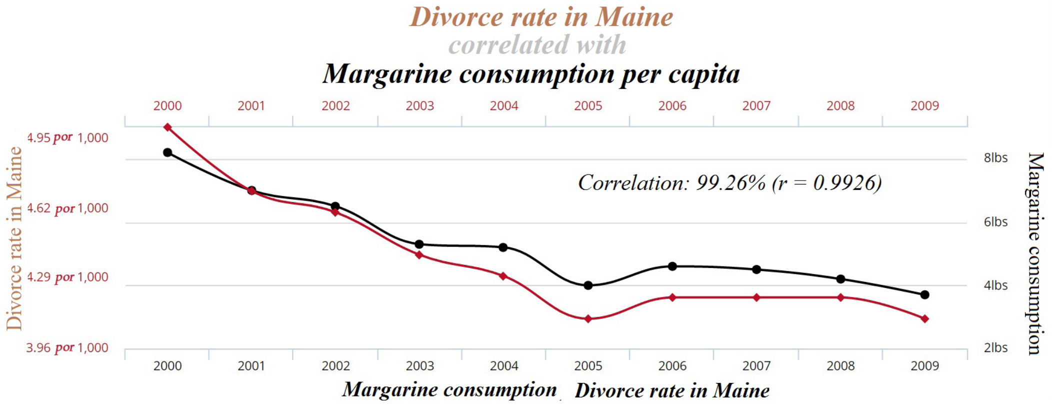


Figure 7. Correlation between the divorce rate in Maine and margarine consumption per capita. Source: Vigen (2019).

how they are related, *Pearson's linear regression coefficient* will be used here.

The linear regression coefficient (cor) will always be between -1 and 1. Thus, cor<0 means negative linear correlation,cor> 0 means positive linear correlation and cor=0 means no linear correlation (BUSSAB; MORETTIN, 2010). In R, it is given by the expression:

```
cor(x,y)
```

Example 1: Check that there is a linear relationship between the communication and motor coordination variables, as well as between communication and fine motor coordination for boys, assuming a 95% confidence level.

```
#Relationship between communication and motor
coordination variables for boys
>cor(x1,x2)
[1] 0.462047
#Relationship in the variables communication and
fine motor coordination for boys
>cor(x1,x3)
[1] 0.6153194
```

In the first case, the correlation is *0.46*, which indicates a *weak linear correlation*. While in the second case, the correlation is *0.61*, which indicates a *positive linear correlation*. However, to know if the correlation is, in fact, significant, a hypothesis test is performed.

6.2 Hypothesis testing

Command to generate the level of significance of the correlation:

```
cor.test(x,y,conf.level = 0.95)
```

Where *x* is the first group, *y* is the second group and conf.level is the confidence interval. If no information is entered, the program defaults to the 95% confidence level (related to 5% significance).

```
cor.test(x2,x3)
Pearson's product-moment correlation
data: x1 and x2
t = 1.563, df = 9, p-value = 0.1525
alternative hypothesis: true correlation is not equal
to 0
sample estimates: cor 0.462047
```

The hypothesis to be tested here is $H_0 : r = 0$ versus $H_0 : r \neq 0$, assuming a significance level of 0.05. H_0 is not rejected if the p-value>0.05, which indicates a lack of a linear relationship between the variables.

Considering the previous example, although the correlation obtained is *0.46* when performing a hypothesis test, 0.1525>0.05 is obtained, which means that it is not significant. Thus, it cannot be said that there is a linear correlation between the variables. On the other hand, there are the following results for other variables:

```
cor.test(x1,x3)
Pearson's product-moment correlation
data: x1 and x3
t = 2.3418, df = 9, p-value = 0.0439
alternative hypothesis: true correlation is not equal
to 0
Sample estimates: cor 0.615319
```

It has been found that 0.0439<0.05. Therefore, with a significance level of 0.05, H_0 is rejected, that is, the linear correlation between the variables is different from zero. Depending on the research area and sample size, a correlation of for example 0.5 may be different than zero, but not be considered a high linear correlation.

7 Final Considerations

The current global movement for free software that has been established in Brazil since the 1990sis evident. This movement with representatives of the different social spheres emerges as a way to achieve the necessary freedom to gear the national technological development, decentralizing the accumulated knowledge and potentializing the construction of new possibilities.

In line with this movement, this article aimed to provide researchers with an opportunity to perform basic statistical procedures in a more flexible and independent way, from R, a free and open source software whose popularity has been increasing considerably in the academic area.

The "free" attribute, associated with minors or at no cost is seen by us as essential in Brazilian universities since much research is carried out based on extremely modest financing that needs to be apportioned and often compensated with own resources. Thinking about this, R shows the possibility of not only fueling this movement but also of giving the opportunity of the use of a differentiated tool and the re-allocation of funds for other purposes.

Although R offers so many other statistical possibilities, we seek to demonstrate, in a more practical way, only those analyzes that have been perceived to us in the scope of a group of researchers in occupational therapy who, many times, they are faced with the need to analyze their data, with limited resources and time. It is worth mentioning that in many cases, it will be necessary to apply more complex methods that will require the support of a professional with expertise in statistics.

References

BUSSAB, W. O.; BOLFARINE, H. *Elementos de amostragem*. São Paulo: Ed. Edgard Blucher, 2005.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística básica*. São Paulo: Saraiva, 2010.

COLUMBIA UNIVERSITY. DEPARTMENT OF STATISTICS. *Colors in R*. 2019. Disponível em: <www.stat.columbia.edu/~tzheng/files/Rcolor.pdf>. Acesso em: 23 maio 2018.

COSTA, G. Governo economiza R$ 370 milhões com sistemas operacionais de computador. *Serpro Sede,* Brasília, 5 abr. 2009. Disponível em: <http://www.serpro.gov.br/menu/noticias/noticias-antigas/governo-economiza-r-370-milhoes-com-sistemas-operacionais-de-computador>. Acesso em: 23 maio 2018.

DELLA BARBA, P. C. S. *O empoderamento de pais para o conhecimento sobre o desenvolvimento de seus filhos*: o Ages and Stages Questionnaire - ASQ-BR. 2014. Relatório (Pós-Doutorado em Estudos da Criança) – Universidade do Minho, Braga, 2014.

FUNDAÇÃO OSWALDO CRUZ – FIOCRUZ. COMPREHENSIVE R ARCHIVE NETWORK – CRAN. *Apresenta o link para download do software R*. Vienna, 2019. Disponível em: <https://cran.fiocruz.br/>. Acesso em: 1 fev. 2019.

KIELHOFNER, G. *Research in occupational therapy:* Methods of inquiry for enhancing practice. Philadelphia: FA Davis, 2006.

MIGON, H. S.; GAMERMAN, D.; LOUZADA, F. *Statistical inference:* na integrated approach. Boca Raton: Taylor & Francis, 2014.

MIOT, H. A. Tamanho da amostra em estudos clínicos e experimentais. *Jornal Vascular Brasileiro*, Botucatu, v. 10, n. 4, p. 275-278, 2011.

OTTENBACHER, K.; PETERSEN, P. Quantitative trends in occupational therapy research: Implications for practice and education. *American Journal of Occupational Therapy*, Bethesda, v. 39, n. 4, p. 240-246, 1985.

R CORE TEAM. R. *A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2018.

SAMPAIO, R. F.; MANCINI, M. C.; FONSECA, S. T. Produção científica e atuação profissional: aspectos que limitam essa integração na fisioterapia e na terapia ocupacional. *Revista Brasileira de Fisioterapia,* São Carlos, v. 6, n. 3, p. 113-118, 2002.

VIGEN, T. *Spurious Correlations*. New York: Hachette Books, 2019. Disponível em: <http://www.tylervigen.com/spurious-correlations>. Acesso em: 01 fev. 2019.

YAP, B. W.; SIM, C. H. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation,* Blacksburg, v. 81, n. 12, p. 2141-2155, 2011.

Author's Contributions

Maysa Marinho Antunes Ramos – Text conception; organization of the sources and/or analyzes, text writing. Pedro Luiz Ramos - Text conception; organization of the sources and/or analyzes, text writing. Francisco Louzada Neto – Text conception; Text review. Patrícia Carla de Souza Della Barba – Text conception; Text review. All the authors approved the final version of the text.