Utilizarea programului R pentru analiză statistică

Daniel-Corneliu Leucuța, MD, PhD, lector
Departamentul de Informatică Medicală și Biostatistică
Universitatea de Medicină și Farmacie "Iuliu Hațieganu", Cluj-Napoca
dleucuta@umfcluj.ro

Conținut

R utilizare, popularitate

Disponibilitate, instalare

Interfață grafică cu exemple în R Commander

- Lucrul cu datele (importare, gestiunea setului de date și a variabilelor)
- Statistică de bază (statiscă descriptivă, Teste de normalitate, proporții, medii, variațe, neparametrice, corelații)
- Statistică avansată (modele, diagnostica prezumpțiilor),
- Grafică,
- Afișarea rezultatelor

R/ R Commander pentru a preda statistica

Utilizarea R în linie de comandă

Pachete utile

Găsirea documentației, găsirea de ajutor cu funcții și pachete

Ce este R

R este o implementare a limbajului S

Oferă un număr imens de tehnici de analiză statistică și reprezentare grafică

(11,000 packages (July 2017) disponibile pe Comprehensive R Archive Network (CRAN), Bioconductor, Omegahat, GitHub, ...)

Organizat, flexibil, extensibil, coerent

Programare orientată obiect

Limbaj interpretat (o parte din pachete scrise în C/C++ pentru creșterea vitezei)

Gratuit

Avantaje și dezavantaje

Avantaje

- **Gratuit**
- **Zeci de mii de pachete cu funcții statistice**
- **Ușor de extins**
- **Capabilități grafice**
- **În trend cu machine learning**

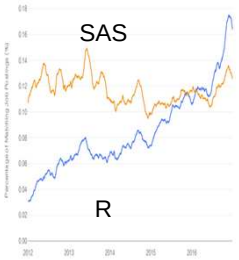
Dezavantaje

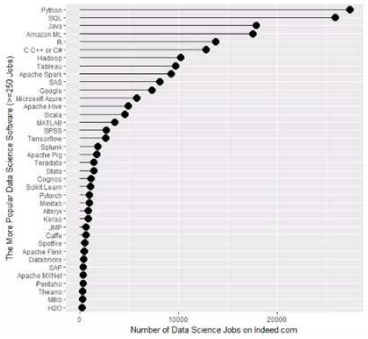
- **Securitate deficitară**
- **Gestiunea memoriei deficitară**

Deși în linie de comandă (majoritar), nu e dificil de folosit de persoane fără experiență în programare

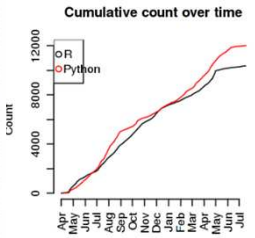
Popularitate

Joburi în data science

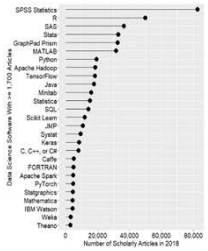




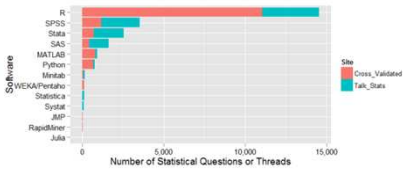
Kaggle data science contests



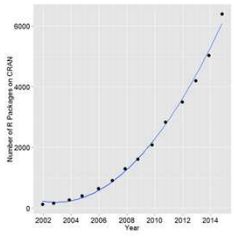
Articole științifice



Întrebări pe site-uri



Pachete în R



http://r4stats.com/articles/popularity/ **The Popularity of Data Science Software.** *by Robert A. Muenchen*

Disponibilitate

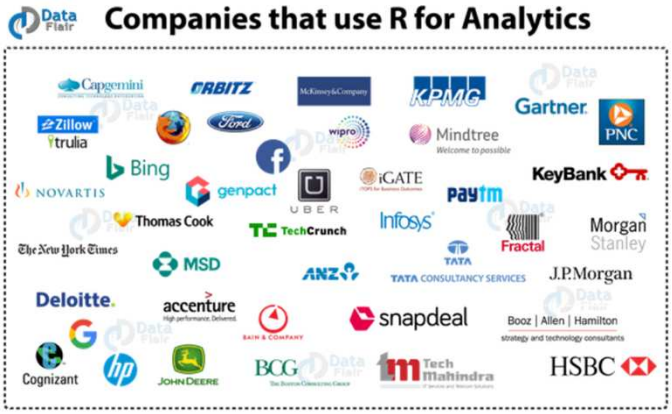
Disponibil pe numeroase platforme:

- **Windows**
- **Linux**
- **OS X**
- **Android**

Ce firme cunoscute folosesc R

Microsoft Google facebook Tweeter

IBM

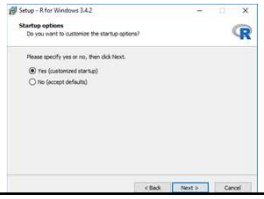


Instalare

Pe Windows – instalare ușoară

http://www.r-project.org/

- **pe prima pagină - link download R**
- **În pagina deschisă (CRAN Mirrors) selectat un server**
- **În noua pagină se selectează link-ul pentru Windows.**
- **În nouă pagină, R for Windows, unde se urmează legătura numită base (binaries for base distribution).**
- **În nouă pagină, intitulată R-*numărul versiunii* for Windows (32/64 bit build) *(ex. R-3.6.1 for Windows)*. Se urmează link-ul din susul paginii Download R *numărul versiunii* for Windows.**
- **Instalare, cu Coustom – selectare SDI, în rest next, next, finish.**



Instalare

Pe Windows

- Există și versiunea de la Microsoft – Microsoft R Open

Pe Linux

http://www.jason-french.com/blog/2013/03/11/installing-r-in-linux/

Pe Mac - OSx

https://cran.r-project.org/bin/macosx/RMacOSX-FAQ.html

Pe Android

https://www.r-bloggers.com/install-r-in-android-via-gnuroot-no-root-required/

Interfețe grafice

Covărșitor funcțiile din R sunt accesibile prin linie de comandă

Există câteva interfețe grafice:

- Mediu de dezvoltare grafic integrat: Rstudio Desktop (gratuit) (desktop/server – cu plată – suport/cloud)
- Editoare R
  - Notepadd ++ (Windows)
  - Tinn-R (Windows)
  - Eclipse (numeroase platforme)
  - Geany (Linux)
- Interfețe Point and click (Windows, Linux, Mac OS):
  - R Commander – cea mai cuprinzătoare și logic structurată
  - RKWard – ca un SPSS, dar relativ puține funcții
  - Rattle GUI - R Analytical Tool To Learn Easily – pentru data mining (cluster analysis, association analysis, decision trees, random forests, boosting, support vector machines, model performance evaluation, exploring data)

R Commander

Interfață point and click

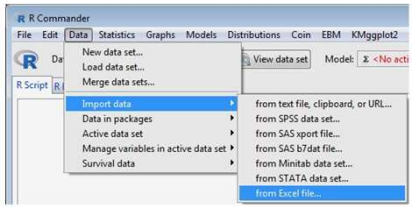
Frumos structurată

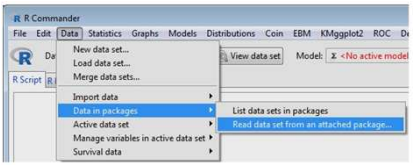
Probabil cea mai cuprinzătoare (extinsă prin plug-in-uri >40, relativ ușor de construit)

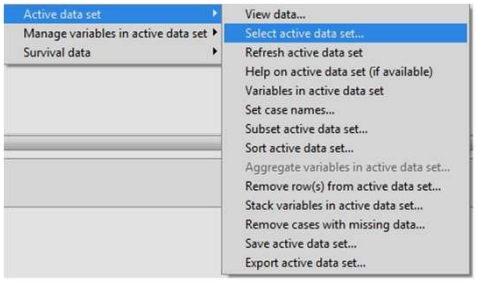
```
[1] "Rcmdr"                    "RcmdrMisc"                "RcmdrPlugin.aRnova"       "RcmdrPlugin.BCA"
 [5] "RcmdrPlugin.BiclustGUI"   "RcmdrPlugin.coin"         "RcmdrPlugin.depthTools"   "RcmdrPlugin.DoE"
 [9] "RcmdrPlugin.doex"         "RcmdrPlugin.EACSPIR"      "RcmdrPlugin.EBM"          "RcmdrPlugin.EcoVirtual"
[13] "RcmdrPlugin.epack"        "RcmdrPlugin.Export"       "RcmdrPlugin.EZR"          "RcmdrPlugin.FactoMineR"
[17] "RcmdrPlugin.FuzzyClust"   "RcmdrPlugin.GWRM"         "RcmdrPlugin.HH"           "RcmdrPlugin.IPSUR"
[21] "RcmdrPlugin.KMggplot2"    "RcmdrPlugin.lfstat"       "RcmdrPlugin.MA"           "RcmdrPlugin.mosaic"
[25] "RcmdrPlugin.MPAStats"     "RcmdrPlugin.NMBU"         "RcmdrPlugin.orloca"       "RcmdrPlugin.PcaRobust"
[29] "RcmdrPlugin.plotByGroup"  "RcmdrPlugin.pointG"       "RcmdrPlugin.qual"         "RcmdrPlugin.RiskDemo"
[33] "RcmdrPlugin.RMTCJags"     "RcmdrPlugin.ROC"          "RcmdrPlugin.sampling"     "RcmdrPlugin.SCDA"
[37] "RcmdrPlugin.SLC"          "RcmdrPlugin.SM"           "RcmdrPlugin.sos"          "RcmdrPlugin.steepness"
[41] "RcmdrPlugin.survival"     "RcmdrPlugin.sutteForecastR" "RcmdrPlugin.TeachingDemos" "RcmdrPlugin.temis"
[45] "RcmdrPlugin.UCA"
```

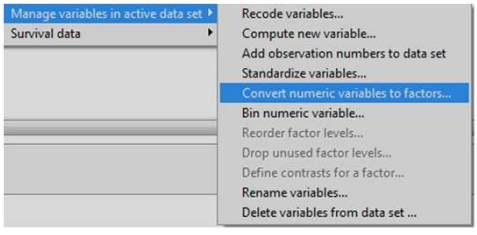
R Commander

- **Lucrul cu datele**
- **Importare**
- **Gestiunea setului de date**
- **Gestiunea variabilelor**



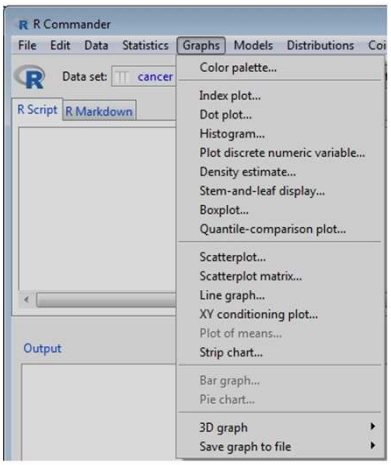






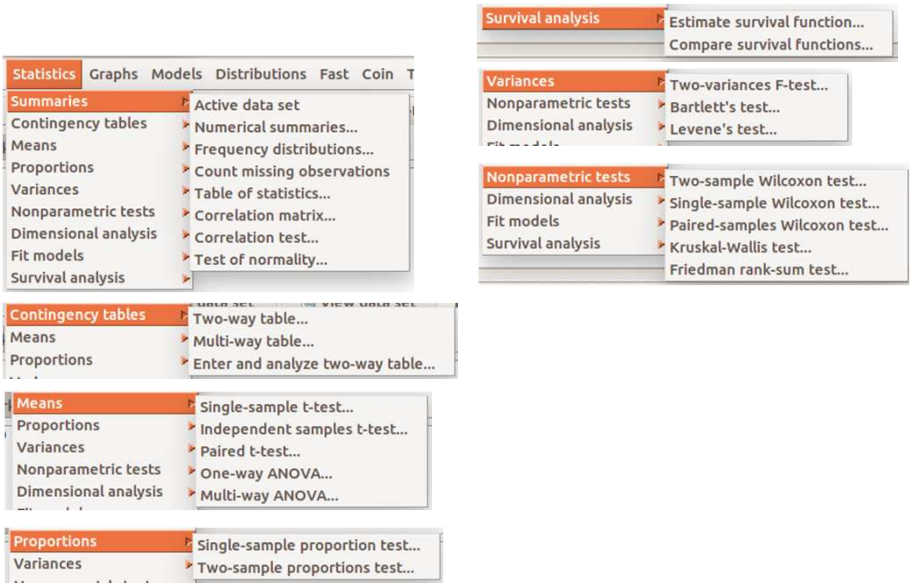
R Commander

- **Grafică**



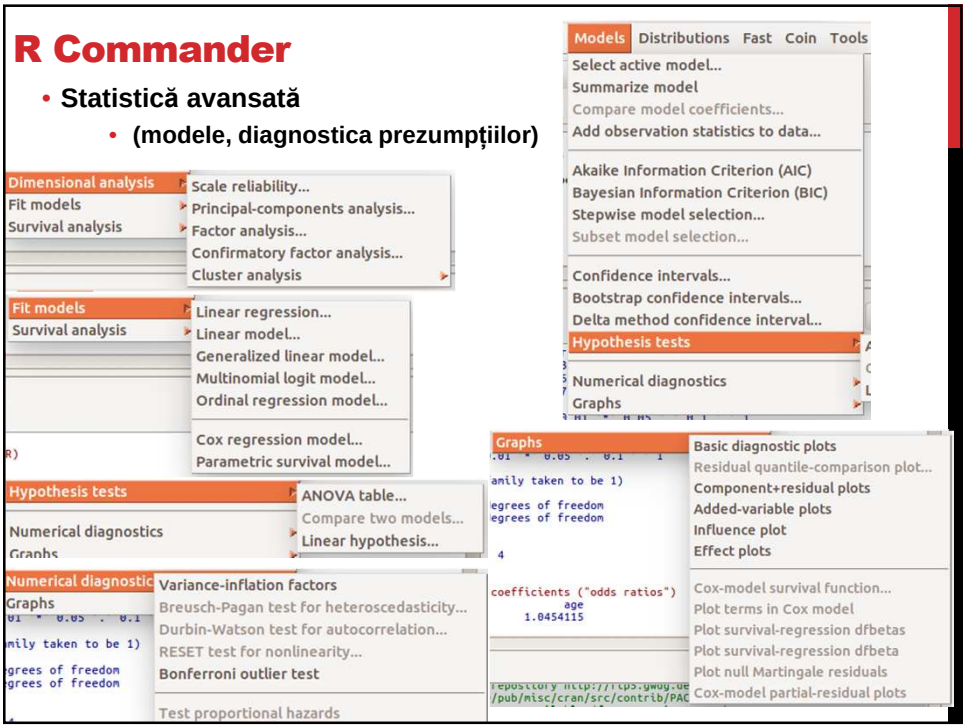
R Commander

- **Statistică de bază**



R Commander

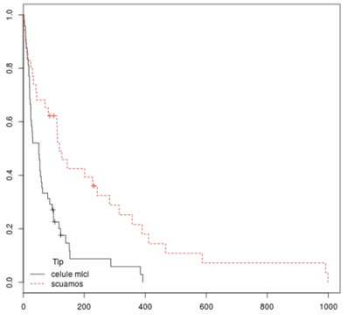
- **Statistică avansată**
  - **(modele, diagnostica prezumpțiilor)**

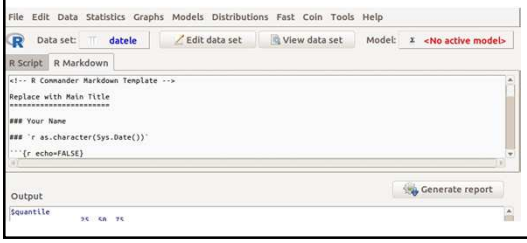


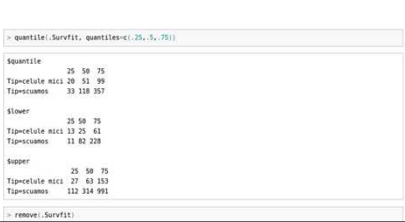
R Commander

Afișarea rezultatelor

- Output (text)
- Fereastra de grafice
- R Markdown
  - Html
  - Doc (dacă e instalat pandoc)
  - PDF (dacă e instalat LaTeX)





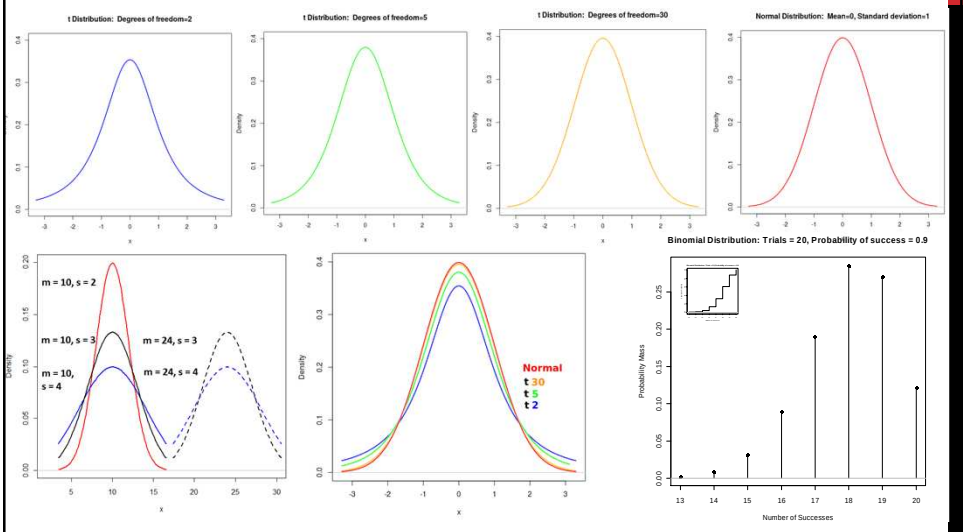


R/R Commander pentru a preda statistica

- **Realizarea de simulări pentru materialul de curs**
  - Distribuții de probabilitate – meniul Distributions ...
  - Plug-in RcmdrPlugin.TeachingDemos
- **Exeplificări de analize**
  - R Commander ...

R/R Commander pentru a preda statistica

Realizarea de simulări pentru materialul de curs



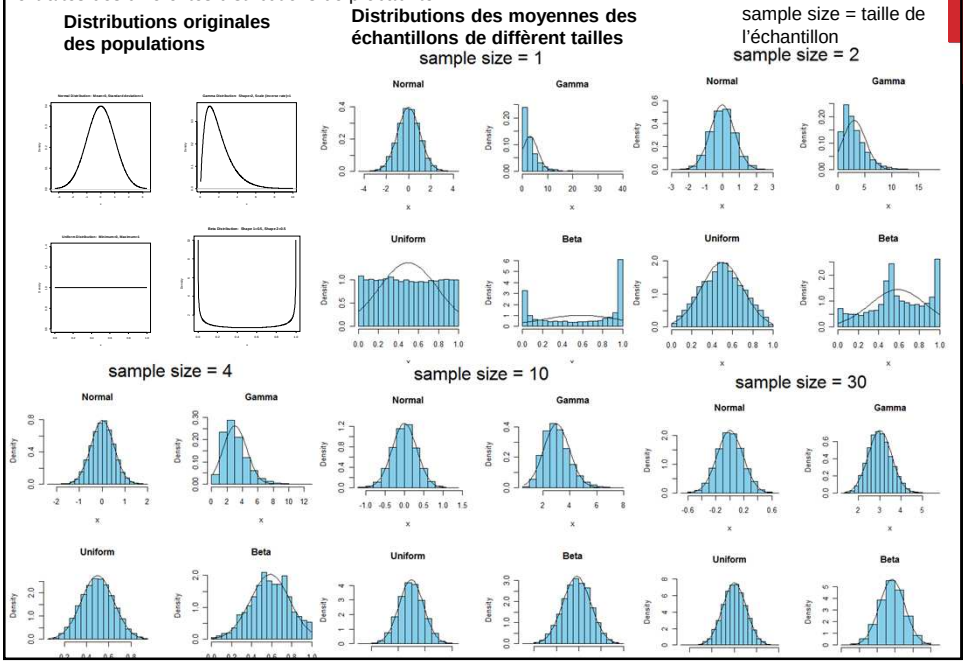
Simulation des moyennes des dizaines des échantillons des diffèrent tailles, extraites des différentes distributions de probabilité:

Théorème Limite-Centrale

sample size = taille de l'échantillon

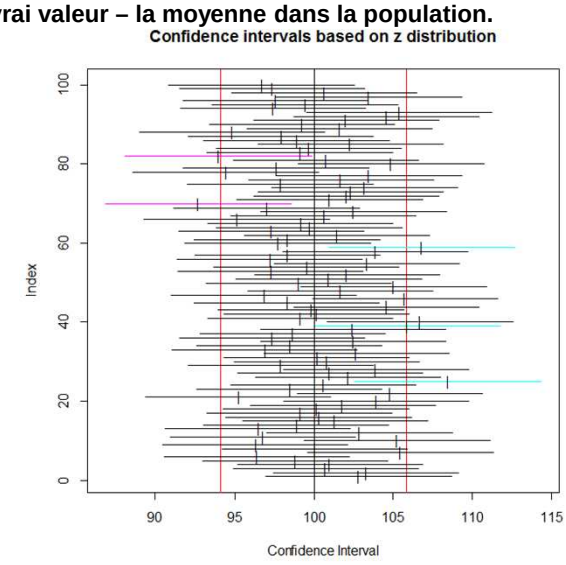
Distributions originales des populations

Distributions des moyennes des échantillons de diffèrent tailles



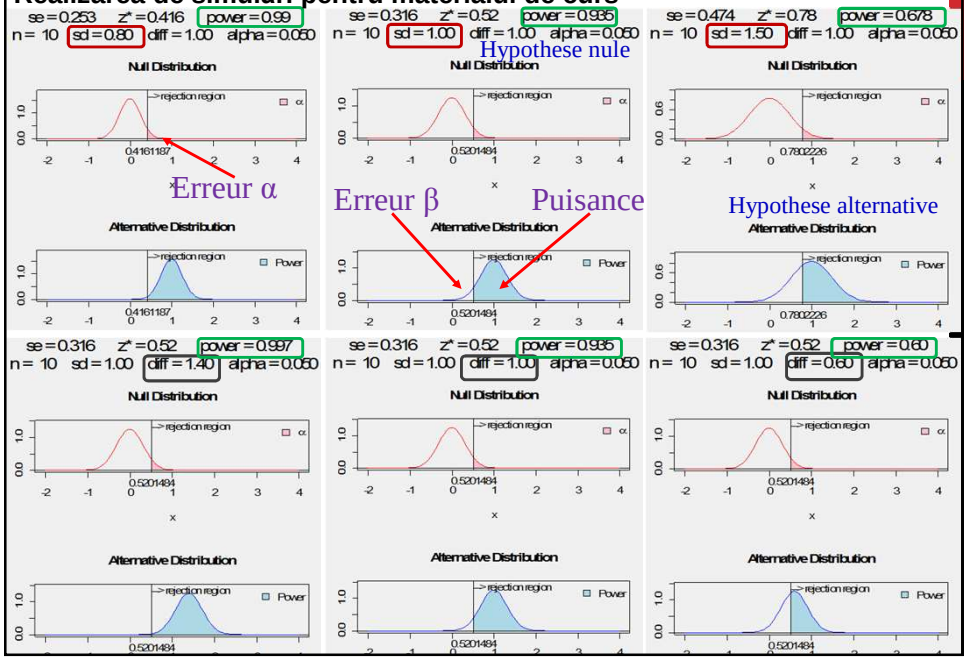
Simulation des intervalles de confiance

Depuis une population avec moyenne 100, déviation standard 15, on a extrait k = 100 échantillons aléatoires de n = 25 sujets. Leurs moyennes et intervalles des confiance 95% calculées sont affichées en bas. 95 intervalles de confiance incluent la vrai valeur – la moyenne dans la population.



R/R Commander pentru a preda statistica

Realizarea de simulări pentru materialul de curs



R/R Commander pentru a preda statistica

Exeplificări de analize

Statistique descriptive:

```
mean       sd  IQR         cv   skewness   kurtosis  0% 25%
50%    75% 100% data:n data:NA

un homme  180.2162 5.432081 8.00 0.03014202 0.49072474 -
0.7899348 172 176 180 184.00  191     37       0

une femme 166.0962 6.065847 8.25 0.03652009 0.08475652 -
0.7696238 154 162 165 170.25  178     52       1
```

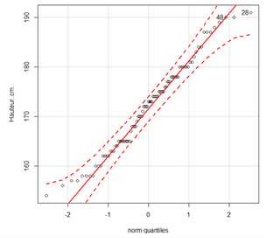
Le test F du Fisher pour comparer les variances:

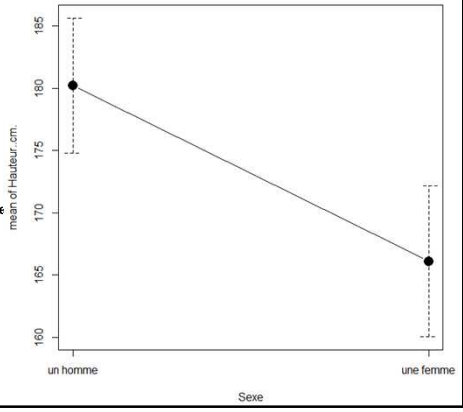
p-value = 0.4903, donc > 0,05 => on peut pas dire que les variances sont inégales => on va utiliser le test t pour échantillons indépendants avec variances égales

Le test t pour échantillons indépendants avec variances égales:

p-value < 0.001, donc < 0,05 => il y a une différence statistiquement significative entre les homes et les femme du point de vue de l'hauteur

Sur le graphique on observe que l'hauteur des hommes est plus grand que l'hauteur des femmes.





Utilizarea R în linie de comandă

Nu e așa de dificil cum am putea crede, nu trebuie să fim programatori

Esența pentru începători

- Se importă baza de date cu Rcommander
- Variabilele se identifică prin: nume.dataset$nume.variabilă
- Se caută documentația pentru funcția de interes
- Se înlocuiesc/introduc numele variabilelor
- Se înlocuiesc/introduc numele setului de date
- Se precizează opțiunile

Ex: testul student pentru eșantioane independente:

- t.test(dataset$age~dataset$gender, alternative='two.sided', conf.level=.95, var.equal=FALSE)

sau

- t.test(age~gender, alternative='two.sided', conf.level=.95, var.equal=FALSE, data=dataset)

Utilizarea R în linie de comandă

Variabile

```
> x <- 1
> x
[1] 1
> x <- "abc"
> x
[1] "abc"
> x <- seq(from=1,to=5,by=.5)
> x
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> x <- rep("A",5)
> x
[1] "A" "A" "A" "A" "A„
```

- `x <- TRUE`
- `X <- c("A“, "A“, “b”)`

Utilizarea R în linie de comandă

Datele lipsă

```
NA – ideal de înlocuit manual spațiile goale din fișierul Excel cu NA
```

Comentarii – semnul

```
x <- 2x
# x <- 1
x
```

Utilizarea R în linie de comandă

Matrice

```
x = matrix(
   c(1, 2, 3, 4, 5, 6),
   nrow=3,
   ncol=2)
x

y = matrix(
   c(3, 2, 1),
   nrow=3,
   ncol=1)
x <- cbind(x,y) – adauga o coloana
```

Dataframe

Utilizarea R în linie de comandă

Funcții definite de utilizator

```
fncDivideBy2 <- function(x) {
        return(x/2)
}
x <- 2
print(fncDivideBy2(x))
[1] 1

fncAxB <- function(a,b) {
        return(a*b)
}
print(fncAxB(2,3))
print(fncAxB(a=2,b=3))

[1] 6
[1] 6
```

Utilizarea R în linie de comandă

Programare în R – bucla for

```
x <- c(5, 4, 3, 2, 1)
for(i in 1:length(x)) {
        print(x[i]^2)
}
print(i)

[1] 25
[1] 16
[1] 9
[1] 4
[1] 1

> print(i)
[1] 5
```

Utilizarea R în linie de comandă

Programare în R – bucla while

```
i <- 1
while(i<5) {
        print(i^2)
        i <- i+1
}
print(i)
```

Utilizarea R în linie de comandă

Programare în R – execuție condiționată

```
x <- 1
if (x==1) {
        print("x=1")
} else {
        print("x!=1")
}


x <- 1
y <- ifelse(x==1,y=1,y=0)
```

Pachete utile

Sunt nenumărate pachete utile a fi utilizate în R, depinde mult deomeniul în care lucrăm

Există liste de pachete pe site-ul programului:

https://cran.r-project.org/web/views/

Unele pachete utile în statistica medicală ar putea fi:

GrapheR – pentru realizare de grafice cu calitate înaltă penru a publica

LogisticDx – pentru analiza diagnostică a regresiilor logistice

Google ;)

Găsirea documentației, găsirea de ajutor cu funcții și pachete

Google R + întrebare, sau nume pachet R + întrebare, sau nume funcție R + întrebare

Stackoverflow.com – site cu întrebări și răspunsuri (include și R)

În R

help(t.test)

help("t.test")

?t.test

?"t.test„

Utilizarea de ”vignete” – PDF-uri cu tutoriale, materiale didactice, de clarificare

browseVignettes() – pentru a afla toate vignete-le disponibile în pachetele instalate

Utilizarea demonstrațiilor

demo() pentru a afla toate demonstrațiile

Găsirea documentației, găsirea de ajutor cu funcții și pachete

args(nume functie) – tipărește ce e necesar în apelul funcției

str (obiect) – tipărește structura unui obiect

Mai multe sugestii:

https://www.r-project.org/help.html

Documentația de pe siteul r-project.org

https://www.flickr.com/photos/wwworks/4759535970