The Fi Pioyeci to- Siaiic;ic<ii Comput ng Stzitsiicai Tarnius iicirg U Plata zwaiysis with Fi Fioetsvap in we ond

Introduction to R — a computing software for

statistical analysis

Krzysztof Podgérski

Department of Mathematics and Statistics

University of Limerick

September 8, 2009

Statslical Table; licirg Tl Fate ialysis mth Fl Rootstrap vi +0 and

The highlights

The package is available for all popular operating systems:

windows

Mac os m

Linux

It is free!

Everyone (knowledgeable enough) can contribute to the software by

writing a package

Packages are available for download through a convienient facility

It is fairly well documented and the documentation is available either

from the program help menu or from the website

It is the top choice of statistical software among academic statisticians

but also very popular in industry specially among biostatistlclans and

medical researchers (mostly due to the huge package called

Bioconductor that is built on the top of R)

It is a powerful tool not only for doing statistics but also all kind of

scientific programming

Stat stical Tables usirg I? Data m walysia with Fi Floctstrap in ‘lwc and

Fl is a language and environment for statistical computing and graphics.

Fi provides a wide variety of statistical and graphical techniques, and is highly extensible. Among its tools

one can tind implemented

linear and nonlinear modelling,

classical statistical tests,

timerseries analysis,

classltication,

clustering,

One ot Rs strengths is the ease with which welldesigned publicatlonrguality plots can be produced.

including mathematical symbols and tormulae where needed.

Fl is an integrated suite ot software tacilities for data manipulation. calculation and graphical display. It

includes

an eftective data handling and storage facility,

a suite ol operators lor calculations on arrays, in particular matrices,

a large, coherent. integrated collection ot intermediate tools tor data analysis,

graphical tacilities tor data analysis and display either onrscreen or on hardcopy, and

a well-developed, simple and elfective programming language which includes conditionals, loops,

user-defined recursive functions and rnput and output facilities.

# Reproduction of Tables

The Fl Plolccl fo· Slallsrical Conrztll ng Viale ; ialysls mth Vi Roctsllap vi +0 mid

There is more than meets the eye in the table

It is not only the table values that can be explored for the

standard normal distribution using R. Recall that the normal

distribution is defined by the density

f(z) : Le‘Z2/2.

x/ 2rr

The density represents distribution of probability for a

random variable associated with it.

The area under the density represents the probability so

the that the total area under it is equal to one.

The area accumulated up to certain value z represents

probability that a corresponding random variable takes

value smaller than z and this probability defines the

cumulative distribution function F(z) which is tabularized.

THQ R Project $01 S\a\\5U0a\ C0¤\pu\ Hg Uma Q·Wa\ys\$ w.'\\}\ U B00ls\’z1p — W We and

All this can be seen in R

The mllcwmg code explores various aspecis ci the standard n0rma\ distnbuticn

#PL0tt1n:; the density functum of the standard ncrmei varlabie

z:seq{\*3,3,l:y:0.U1W

plat {2, dn¤:m(z} , type? L' , <:ol;"rEd", lwd;é}

#Pl0LL4r1g Lhe cdmdlaxgve d;sL;;buL;0n [uncLi¤n (\.haL one [tow Lbe Mable}

pl ct 47,p\*10r·n(7) ,\2ype;' I ' , ccl:"red", Wwdzé)

per (mEr¤w:<‘(2, 1))

plat {Z, dn¤1m(z} ,type—' L' , <:0L—"rsd", lwd—é}

plot 4z,p:10x:n(z> ,\*:ypc:' L' , c0L¤"rcd", Lwd:4}

KS1de by sxde

par (mf10w:C('l, 2))

plot {2, dn¤1\*n(z} ,type—' L' , <:ol—"xEd", lwd—é}

plat {z,pr1¤x·n(z> , ;ypc:' l' , c<>L¤"rcd", lwdié}

Data Analysis

Tm ti Proiecl t¤· Slalisvcal <:¤npni ng Stnisiieei Tables hwg R Bootsvap in ‘lwo end

Data from Table 1.1 of the textbook

Table 1.1 Flandorn and systematic errors

Student Results (ml) Comment

A Precise biased

B 9.88 10.14 10.02 9.80 10.21 Im reclse unbiased

C 10.19 9.79 9.69 10.05 9.78 Im recise, biased

D 10.04 9.98 10.02 9.97 10.04 Precise, unbiased `

This is also given in the text file Tab1e1\_1.txt contents of

which is given below

A 10.08 10.11 10.09 10.10

B 9.88 10.14 10.02 9.80

C 10.19 9.79 9.69 10.05

D 10.04 9.98 10.02 9.97

mw ¤ V’ro;oc\fo· S\a\1sv1oa\Compu\ng sm¤5m¤ mm wg n Bootsvap 7 we and

Reading data from a filo to R

#Reading the data from

Titra;raad.tab1e("Tab1a1\_1.txt", row.namas ; D

Titra

# V2 V3 V4 v5

#1-x 10.08 10.11 10.09 10.10

#8 9.88 10.14 10.02 9.80

#C 10.19 9.79 9.69 10.05

#0 10.04 9.98 10.02 9.97

#Listing the first row

Titra[l,]

#and the last column

T1tra[,4]

The R Propet im 5\a\1si1ca\C¤mpm~ng Stat<s\1ca\Tab\eausmg R Bcotstrap 7 in cha and

I\/leans and standard deviations

~=·=·¤2-M \_

lr. · r the mean and standard deviation 0f A's results.

l I {'

V \_ · - ><. 0.\*0 (wif

i V l v 10.08 —0.02 0.0004

G 10.11 0.01 0.0001

` 10.09 -0.01 0.0001

10.10 0.00 0.0000

. ‘ ` 10.12 0.02 0.0004

Totals 50.50 ¤ 0.0010

· ll y:&:@:10.1m1

, C rr 5

‘

'\_ S = q Z(h 7 Ez/(YI — 1) = V0.001/4 = 0.0158 ml

. ¤ Q e that Z(x,— Q is always equal to 0.

The i3 Protect fen Statisncai Comput ng Statsticai Tabies using R Bectsvap — in We and

I\/leans and standard deviations much faster and better

#C0muting means

1rc>wMeans (Titra)

tt A B c D

#10.0950 9.9600 9.9300 10.0025

#and standard deviation

apply (Tiara, 1, sd)

# A B C D

#0.01290994 0.15055453 0.23036203 0.03304038

The F1 Pr01cc\f0· S\a\wsvwca1C0w1p14\ng Statstwcd Tnbwa nawg R Uma mmysns mm F1

Table 2/I Results uf 50 deherminatiuns uf mcrate non nuncervtrauun, an pg m|’”

0.51 0.51 0.51 0.50 0.51 0.49 0.52 0.53 0.50 0.47

0.51 0.52 0.53 0.48 0.49 0.50 0.52 0.49 0.49 0.50

0.49 0.48 0.46 0.49 0.49 0.48 0.49 0.49 0.51 0.47

0.51 0.51 0.51 0.48 0.50 0,47 0.50 0.51 0.49 0.48

0.51 0.50 0.50 0.53 0.52 0.51 0.50 0.50 0.51 0.51

Also wn lhehle 1‘;b1e?\_\ .L><t

0.51 0.51 0.51 0.50 0.51 0.45 0.52 0.53 0.50 0.47

0.49 0.48 0.46 0.49 0.49 0.48 0.49 0.49 0.51 0.47

0.51 0.51 0.51 0.48 0.50 0.4/ 0.50 U.51 0.49 0.48

The R V’r0]ect €0· Stattsvtca] Comput ng Statsttcd `[ztbwe uatrg R [lata c.-wmysts mm Ft

The mean concentration

#Get;t;ing data in a vector

x:scan(’ Tab1e2\_1 .t;xt;')

mean (X)

#[1] 0.4998

Sd(x)

#[1] 0.01647385

Tito l'l Piolocl ’o· Slalisncqil Conptil ng Stzllslicai Tables liowg if Bala ialysis with Fl

Pulling ourselves by bootstraps

If we would repeat our experiment of collecting 50 samples of

nitrate concentrations many times we would see the range of

BUOY.

But it would be a waste of resources and not a viable method.

Instead we resample ‘new‘ data from our data and use so

obtained new samples for assessment of the error.

The following R code does the job.

mzmeanlxl

bt;ctst¤ap:vec:o1(' nuner,<;' , BDU)

for (1 11 1:50Cl

S,. ..S... pl.i;....t. ..p.

l

nist (boostrapl

We can safely say that the nitrate concentration is 49.99 1 0.005.