R estatistikarako programazio-lengoaia

Josemari Sarasola

2017

1 Zer da R?

R bereziki estatistika lantzeko programazio-lengoaia eta softwarea da. Software askea da, hots, doan da eta nahieran moldatu eta lantzeko aukerak eskaintzen ditu. R softwarearen beste berezitasuna interfazea da: eskuz idatzi beharreko aginduen bitartez garatzen baita, menuen bitartez funtzionatzen duten egungo programa gehienek ez bezala.

R urte batzuk lehenago, 1980ko hamarkadaren erdian, garatu zen S eta S-PLUS programa komertzialetatik eratortzen da. 1994 urtean Ross Ihaka eta Robert Gentleman S bezalako programa bat garatu zuten eta R izena eman zioten. Hasieratik, software aske moduan jarri zuten jendearen eskura eta geroztik zabaltzen joan da, hainbat modulo berrirekin.

2 Oinarrizko aginduak

Estatistikan maiz egiten dira kalkuluak. R kalkulagailu moduan erabil daiteke. Adibidez, honako lehenengo agindu hau idatziz bigarren lerroko emaitza

suertatuko da, lehenengo aginduaren ondoren "Enter"sakatuz (agindu bat exekutatzeko, beti sakatu behar da "Enter"tekla):

```
> (45+67+92)/7
> 29.14285714
```

Azken emaitza hau gordeta geratzen da ".Last.value"izenarekin. Izena oso konplikatua denez, beste izen bat jarriko diogu:

```
> emaitza=.Last.value
> emaitza
> 29.14285714
```

Ikusten dugunez, lehenengo aginduak ez du ezer egiten; izan ere, izen aldaketa bat besterik ez du egiten eta programaren memorian gorde. Izena horrekin deia egiten denean (bigarren lerroan), izen horrekin gordetako balio zehatza emango digu (hirugarren lerroan).

Emaitza zehaztasun gutxiagorekin azaltzea nahi bada:

```
> print(emaitza,digits=3)
> 29.1
```

Aukeran, honela ere egin daiteke:

```
> borobil=round(emaitza,digits=1)
> borobil
> 29.1
```

Datuak Rtik bertatik sartzen dira era sinple eta errazenean. Adibidez, ikasle batzuen notak modu honetan sar ditzakegu:

```
> notak=c(6,7,5,5,3,8,6,7)
```

"Enter"tekla sakatzen bada, ez du ezer egiten, datu batzuk sartu eta horiei izena eman besterik ez baitugu egin. Izena memorian gordeta geratzen da. notak izenerako deia egiten bada, datu zerrenda izango da emaitza

```
> notak
> [1] 6 7 5 5 3 8 6 7
```

Daturen bat ahaztu bazaigu, ez du axola, aise gaineratzen baitira. Adibidez, 3 nota gaineratu nahi bada zerrendaren hasieran:

```
> notak=c(3,notak)
> notak
> [1] 3 6 7 5 5 3 8 6 7
```

Datuak kendu ere egin daitezke. Adibidez, bigarren datua kentzeko hau egin behar da:

```
> notak=notak[-2]
```

Balio jakin bat ezabatu daiteke. 5 balioa duten datuak ezabatzeko:

```
> notak=notak[notak!=5]
```

Izena nahi den eran alda daiteke. Adibidez, notak ordez, lehenzatikonotak izena nahi bada jarri. Beste alde batetik agindu baten ondoren ikurraren ondoren idazten denak ez du inongo eraginik eta iruzkinak idazteko erabiltzen da

```
> lehenzatikonotak=notak #Hau izen aldaketa besterik ez da
> lehenzatikonotak
> [1] 6 7 5 5 3 8 6 7
```

Sar ditzagun bigarren zatiko notak:

```
> bizatikonotak=c(8,6,4,6,4,9,7,5)
```

Bi noten batuketa aise egiten da:

```
> batuketa=lehenzatikonotak+bizatikonotak
> batuketa
```

Bi noten batezbestekoa (bbko izena emango diegu) honela kalkulatu behar da:

```
> bbko=batuketa/2
> bbko
```

Datu-multzoak ordenatu ere egin daitezke:

```
> sort(bbko)
```

Aldagai kualitatiboak komatxoen artean sartzen dira. Adibidez, pertsona zenbaiti telefono mugikorraren marka galdetu zaie (s: Samsung, n:Nokia, a:Apple, l:LG):

```
> tele=c("s","n","a","n","s","l")
> tele
[1] "s" "n" "a" "n" "s" "l"
```

Idatzitako agindu luze bat idaztean akatsa egin eta zuzendu nahi badugu (datu bat sartzea ahaztu zaigunean, adibidez) ezin dugu atzera egin, baina ez da agindu osoa berriz ere idatzi behar: gorako gezia sakatzen bada, sartutako azken agindua agertzen da pantailan eta bertan zuzendu daiteke akatsa.

Kurtsorea aginduaren hasierara eraman nahi baduzu, CTRL+a sakatu behar da,. Aginduaren amaierara joan nahi baduzu, berriz, aski da CTRL+e sakatzea.

Aurretik idatzitako agindu batzuk badira berriz ere begiratu nahi dituzunak history agindua erabili behar da. Adibidez, azken 4 aginduak bistaratzeko:

```
> history(4)
```

Telefono marken kode hauek esanguratsuak ez direnez, itzulpena egiteko agindua badago:

```
> telefonoak=factor(tele,levels=c("s","n","a","l"),
labels=c("Samsung","Nokia","Apple","LG"))
> telefonoak
[1] Samsung Nokia  Apple  Nokia  Samsung LG
Levels: Samsung Nokia Apple LG
```

Aurretik bektorea zenbakizko balioekin osatu eta zenbakizko balio horiek itzuli ere egin daitezke, aldagai kualitatiboa eskuratzeko.

Aldagai kuantitatiboa kualitatibo bihur daiteke, zenbakizko tarteei kategoriak esleituz. Adibidez, zenbakizko kalifikazioen multzo bat kalifikazio kualitatibo bihurtu nahi bada:

```
>kalif=c(2,7,5,3,9,4,6,7,9,5,4,5,6,7,8,4,10,3)
>kalifkual=cut(kalif,breaks=c(0,5,7,9,10),
labels=c("ez gainditu","aprobatu","ongi","bikain"))
>kalifkual
 [1] ez gainditu aprobatu    ez gainditu
 ez gainditu ongi         ez gainditu
 [7] aprobatu    aprobatu    ongi
 ez gainditu ez gainditu ez gainditu
[13] aprobatu    aprobatu    ongi
 ez gainditu bikain       ez gainditu
Levels: ez gainditu aprobatu ongi bikain
```

Aldagai kualitatibo nahiz kuantitatiboetan elementu jakin batzuk isolatzerik badago. Adibidez, hurrengo aginduak notak datu-multzotik aprobatuak bereizi eta aprobatunotak izena ematen dio azpimultzoari:

```
> notak=c(8,6,4,6,4,9,7,5,3,2,6,7,2,3,4,7,8,9,6,7)
> aprobatunotak=notak[notak>=5]
> aprobatunotak
[1] 8 6 6 9 7 5 6 7 7 8 9 6 7
```

Berdintzaren kasuan bi berdintza ikur jarri behar dira:

```
> bostekoak=notak[notak==5]
> bostekoak
[1] 5
```

Datu-multzoak bateratu ere egin daitezke, euren jatorria galdu gabe:

```
> anotak=c(2,5,6,7)
> bnotak=c(3,2,1,8)
> notak=stack(list(a_gela=anotak,b_gela=bnotak))
> notak
   values     ind
1       2 a_gela
2       5 a_gela
```

```
3       6 a_gela
4       7 a_gela
5       3 b_gela
6       2 b_gela
7       1 b_gela
8       8 b_gela
```

Estatistikan aldagaien arteko erlazioak bilatzen dira askotan (aurreko adibidean bezala, non notak gelaren arabera azter daitezkeen). Aurreko agindua erabil daiteke horretarako, baina nahiko desegokia da kasu gehienetan. Asko ere errazagoa da data.frame agindua erabiltzea:

```
> notak=c(2,5,6,7,3,2,1,8)
> gela=c("a","a","a","a","b","b","b","b")
> dena=data.frame(notak,gela)
> dena
```

Askotan datu-multzo anizkoitz hauetan aldagai bakar batekin lan egin nahiko dugu. Aldagaiak banaka aukeratzeko, data multzoaren izena eta "$"ikurraren ondoren aldagaiaren izena jartzen da. Horrela, aldagaia isolatu egiten da:

```
> dena$notak
> dena$gela
```

Aldagai bati buruzko kalkuluak beste aldagaiak hartzen dituen balioen arabera burutu nahi badira, tapply agindua erabiltzen da. Adibidez, batez besteko (ingelesez, mean) nota kalkulatu nahi bada gelaren arabera:

```
> tapply(notak,gela,mean)
```