Статистик программчлалын R хэл

Танилцуулга, суулгах заавар, анхны код

Г.Махгал

2016/9/15

Статистикийн өргөн хэрэглэгддэг програмуудаас

- SPSS
- SAS
- R
- STATA
- Matlab

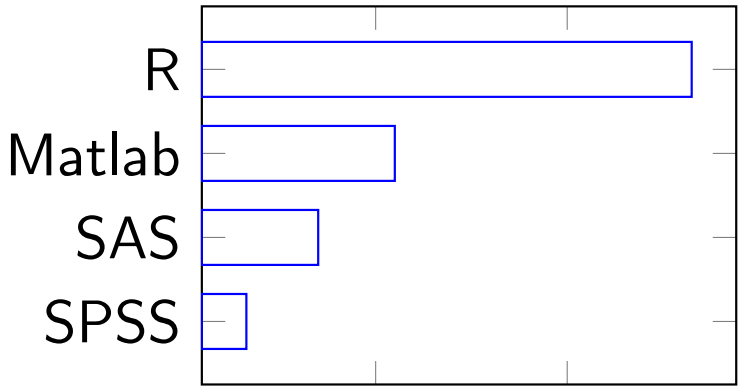
Манай улсын их дээд сургуулиуд сургалтандаа SPSS, EViews болон Matlab зэрэг програмуудыг голчлон ашиглаж байна.

Харьцуулалт №1
Үнэ төлбөр ба Ажиллах үйлдлийн систем

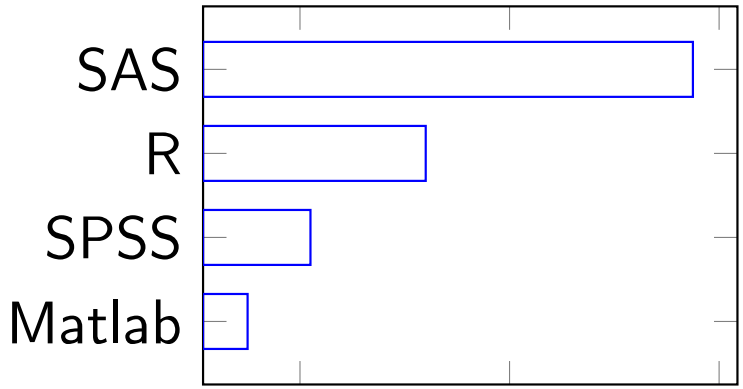
Програм	Үнэ төлбөр	Үйлдлийн систем
SPSS	төлбөртэй	Windows, Linux, Mac
SAS	төлбөртэй	Windows, Linux
R	үнэгүй	Windows, Linux, Mac
STATA	төлбөртэй	Windows, Linux, Mac
Matlab	төлбөртэй	Windows, Linux, Mac

Дэлгэрэнгүй харьцуулалтыг `https://en.wikipedia.org/wiki/Comparison_of_statistical_packages`

Харьцуулалт №2
Хэрэглэгчдийн тоо ба Хөдөлмөрийн зах зээл дээрх эрэлт



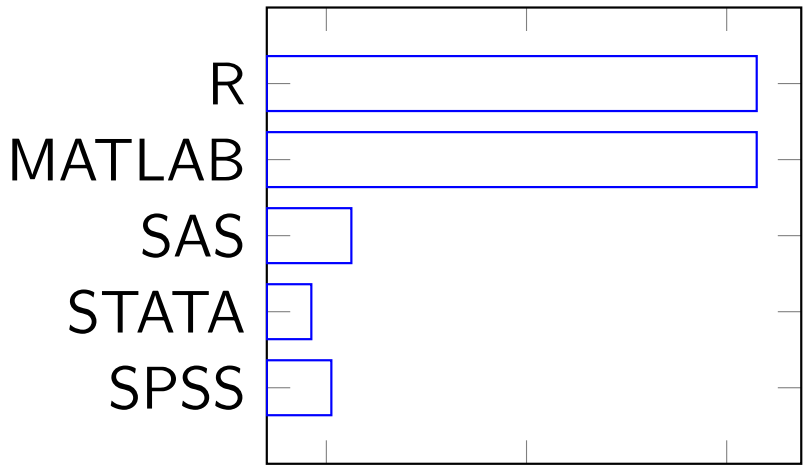
(a) Хэрэглэгчдийн тоо,
Kaggle.com сайтын мэдээгээр



(b) Ажлын байр,
Indeed.com сайтын мэдээгээр

Дэлгэрэнгүй харьцуулалтыг
`http://r4stats.com/articles/popularity/`

Харьцуулалт №3
Статистик шинжилгээнүүдийг тусгасан байдал



Зураг: Онол, практикт өргөн тохиолддог статистикийн арга техникүүдийг тусгасан байдал

Дэлгэрэнгүй харьцуулалтыг
`http://stanfordphd.com/Statistical_Software.html`

Харьцуулалт №4: Эрдэмтэн, судлаачдын "нүдээр"

"One reason SPSS and SAS are so prevalent is because many older faculty and established research groups have been using it for years. Prior to R, these were clearly better than having to write your own programs. They provided easily repeatable and easy to verify results.

Flash forward to today, we have R which may require some programming skills; however, many packages are available that minimize this need. R is free, new packages are available as quickly as the theory is published, and it is now being accepted by a wider audience as a valid alternative to the commercial software. While SPSS and SAS are likely not going away, but as budget cuts encourage the use of open source and free-ware, the younger generation of scientists who learn R will encourage its use in subsequent years." Aug 24, 2012

– *Anthony L. Nguy-Robertson*, University of Nebraska at Lincoln

Эх сурвалж
`http://www.researchgate.net/post/Which_is_better_R_or_SPSS`

R

- 1970-аад оноос эхтэй статистикийн S хэл дээр тулгуурласан.
- өргөн хэрэглэгддэг
- програмчлалын хэл
- эрчимтэй хөгжиж байна
- үнэ төлбөргүй (GNU General Public Licence)
- голлох үйлдлийн системүүдийг дэмждэг
- мэргэжилтнүүд сургалтанд ашиглахыг зөвлөдөг

Төслийн веб сайтын хаяг

`https://www.r-project.org/`

R

R бол *interpreted language*[1].

Жишээ

Команд

```
data <- c(3,5,7,9)
mean(data)
```

Үр дүн

```
[1] 6
```

[1] машины хэлэнд урьдчилан хөрвүүлэхгүйгээр ажиллах программчлалын хэл, өөрөөр хэлбэл скрипт буюу кодоо бичээд шууд ажиллуулдаг

R-тай ажиллах GUI[3]-үүд

Санамж

R нь CLI[2]-тэй учраас GUI-ээр дамжуулан хэрэглэхийг зөвлөе.

- RStudio
- RKWard
- Emacs
- Rcmdr
- Vim
- Notepad++
- ...

[2] command line interface
[3] graphical user interface

R хэлний кодыг эхлэн суралцагчдад

R хэлний кодыг бичиж сурахтай холбогдсон онлайн хичээлүүд

- https://www.datacamp.com/
- http://tryr.codeschool.com/[4]
- http://www.cyclismo.org/tutorial/R/
- http://www.statmethods.net/[5]

Санамж

R нь үнэгүй бөгөөд өргөн тархсан учраас анхлан болон гүнзгийрүүлэн суралцагчдад зориулагдсан үнэгүй ном, төлбөргүй веб сайтууд хангалттай олон байдаг.

[4] уг курсыг судалж дуусгахад O'Reilly хэвлэлийн газраас гаргасан R-ын талаарх зарим номыг хямдруулах купон бүхий хуудасны линк гарна: http://www.oreilly.com/data/try-r/congrats.html

[5] зохиогчийн R in Action номыг 38% хямдруулан авах купоныг агуулсан

R хэлийг ашиглан юу хийж болох вэ?

- Бэлэн багцуудыг[6] нь ашиглан олон төрлийн статистик шинжилгээ хийх
- Өөрийн скрипт, функц цаашилбал багцыг үүсгэх замаар ямар ч статистик шинжилгээг хийх
- CLI-ээр дамжуулан дурын програмчлалын хэлтэй холбох

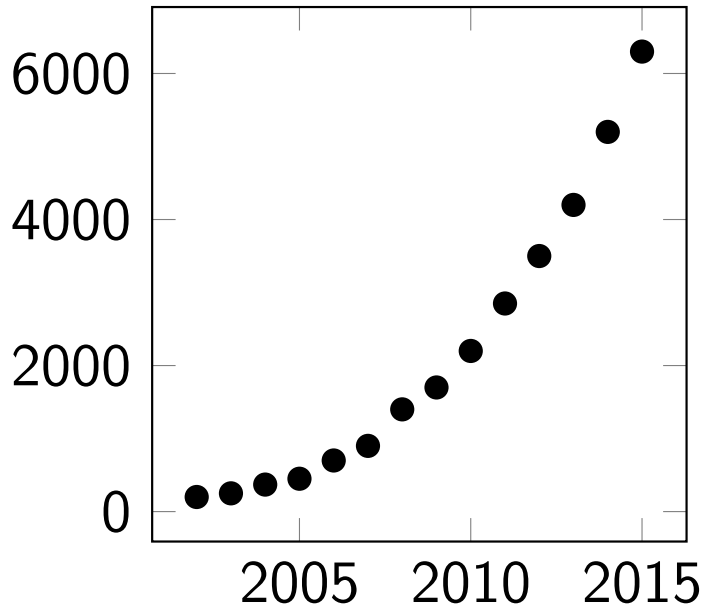
Жишээ

PHP хэлний `shel_exec()` зэрэг функцийг ашиглан статистикийн боловсруулалт хийдэг веб аппликэшн бичиж болно.

- Ердийн тооны машины (calculator програм) оронд ашиглах

[6] package

R хэлний багцын тоо, түүний өсөлт



Зураг: CRAN дээрх R багцын тоо

2015 оны 9 сарын 1-ний байдлаар 7119 багц байна.

Шаардагдах програмчлалын технологиуд

- R нь C, Fortran болон R хэлүүд дээр бичигдсэн.
- R програмын зарим багцууд Java технологид тулгуурласан байдаг. Иймд Java[7] нэмж суулгах шаардлага гардаг.

Санамж

Зарим тохиолдолд

1. Java-г R програмд зориулан дахин тохируулах
2. rJava багцыг ахин суулгах
3. R-ийг дахин ачаалах

болдог.

[7] Java Runtime Environment (JRE) эсвэл Java Development Kit (JDK)

Эхлэн суралцагчдад заавал хэлэх үг

"Using R is a bit akin to smoking.

The beginning is difficult, one may get headaches and even gag the first few times.

But in the long run, it becomes pleasurable and even addictive.

Yet, deep down, for those willing to be honest, there is something not fully healthy in it."

– *Francois Pinard*

Эх сурвалж
http://www.johndcook.com/blog/2012/02/29/comparing-r-to-smoking/

Жишээ: Шугаман загвар: $Z = \beta_1 + \beta_2 Y + \beta_3 XY$

```
X <- c(1,2,3); Y <- c(2,4,1); Z <- c(5,3,2);
lm(Z ~ (X + Y) ^ 2 - X)
```

R програмын талаарх зарим дүгнэлт

"The best thing about R is that it was developed by statisticians.
The worst thing about R is that ... it was developed by statisticians."
– *Bo Cowgill* from Google

Эх сурвалж

Machine Learning for Hackers[8] by Drew Conway and John Myles White, 2012

Санамж: Дутагдалтай тал

R бол чадлын цараа[9] багатай: Multithreading[10] технологийг бүрэн дэмждэггүй.

[8] O'Reilly хэвлэлийн компаниас хэвлэн гаргасан ном

[9] scalability – Ачааллыг нэмэгдүүлэхэд тохирох нөөцөөр чадал буурахгүй ажиллах чадвар. Сайн чадлын цараатай програм нь 10 дахин их нөөц өгөхөд 10 дахин их ачаалал даах эсвэл 10 дахин чадалтай ажиллах, муу цараатай програм 2 дахин их ачаалалд 10 дахин их нөөц шаардах. (www.bolor-toli.com дээрх тайлбар)

[10] https://en.wikipedia.org/wiki/Multithreading_(computer_architecture)

Програмчлалын хэл болон Статистикийн зүгээс харахад

- R бол domain-specific language[11] юм.
  Гэвч general-purpose language маягаар ашиглах боломжтой.
- R хэлийг C++ ба Python зэргээс илүүтэйгээр SAS-тай адилтгах нь зүйтэй.
- R бол статистикчдад л илүү "зохино", тодруулбал статистикчдын хийдэг зүйл програмистууд эсвэл бусад математикчдынхаас өөр юм.

Жишээ: Стандарт хэвийн тархалттай 1 сая санамсаргүй тоо

```
X <- rnorm(1e+6)
```

[11] тусгай зориулалттай програмчлалын хэл

R програм манай улсад[12]

- **Монголын R хэрэглэгчдийн бүлгэм**, хаалттай
  https://www.facebook.com/groups/227442694110958/
- **Монголын R хэрэглэгчдийн групп**, 1 пост (2014 онд)
  http://mongolianr-users.blogspot.com/
- **Монголын R хэрэглэгчдийн групп**, нээлттэй, 90 гишүүн, 2014 оны 3 сарын 3-нд нээгдсэн
  https://www.facebook.com/groups/238749916309897/
- **МУБИС-ийн цахим хуудас**, Судалгааны үр дүнд статистик боловсруулалт хийх R программын сургалт явагдав
  http://msue.edu.mn/index.php?module=menu&cmd=content&id=2141&menu_id=304
- **Монгольская Ассоциация ПРЯЛ**, Програм хангамж: хэрэглээ ба судалгаанд ашиглах нь
  http://www.monaprial.mn/modules.php?ss=4&id=107

[12] интернэт дэх мэдээ материал, 2016 оны 8 сарын 8-ны байдлаар

R хэлний зарим график интерфэйс

- RStudio
- RKWard

Эдгээр нь зөвхөн R-тай ажиллахад тусгайлан зориулагдсан програмууд юм.

Веб хуудсууд

- RStudio
  `https://www.rstudio.com/`
- RKWard
  `https://rkward.kde.org/`

RStudio програмын онцлог

- R-ийг удирдах боломж (restart, terminate)
- Файл, диаграм болон командын түүх, хувьсагчийг утгын хамтаар харуулах компонентууд
- R-ийн package-ийг удирдах, зохицуулах, хөгжүүлэх илүү өргөн боломж

RKWard програмын онцлог

- Эхлэн суралцагчдад тохиромжтой буюу ашиглахад хялбар
- Өргөн хэрэглэгддэг статистик шинжилгээнүүдийг хийх ба диаграмуудыг байгуулах зэргийг хялбарчилсан
- Скрипт бичих нэмэлт хэсэгтэй.
- Өгөгдлийг импортолж авах боломж (text файл, SPSS, STATA файлууд)
- Командын гүйцэтгэлийг зогсоох
- Өргөн хэрэглэгддэг тархалтуудтай холбогдох функцүүд (утга, квантил, загварчлал)

Суулгах заавар: R болон RStudio

1. **R**
   - **Ubuntu Linux**
     Терминалаас дараах командыг өгнө.

     ```
     sudo apt-get install r-base-core
     ```

   - **Windows**
     `https://cran.r-project.org/bin/windows/base/`
     хуудаснаас татан авч суулгана.
2. **RStudio**
   `https://www.rstudio.com/products/rstudio/download/`
   хуудаснаас үйлдлийн системдээ тохирох суулгацыг татан авч суулгана.

Суулгах заавар: R болон RKWard

- **Ubuntu Linux**
  Терминалаас дараах командыг өгнө.

  ```
  sudo apt-get install rkward
  ```

- **Windows**
  `https://rkward.kde.org/RKWard_on_Windows` хуудаснаас стандарт суулгацийг татан авч суулгана.

Анхны код

Дараах кодыг шивж оруулаад ажиллуулж үзнэ үү.

```
# greetings
print('Hello World')

# assign data
X <- c(1,2,3,4,5,6)

# sample mean
mean(X)
```

Үр дүн

```
"Hello World"
3.5
```

Тусламж

Багц, функц зэргийн баримтжуулалтыг үзэх[13]

```
help(mean)
```

эсвэл

```
?mean
```

Санамж

Дараах командаар `help()` функцийн талаарх дэлгэрэнгүй мэдээллийг авах боломжтой.

```
?help
```

[13] `mean()` функцээр жишээлэхэд