

“Статистик програмчлалын R хэл”

сургалтын хөтөлбөр

R програмын товч танилцуулга R бол статистик тооцоолол ба визуалчлалд зориулагдсан нээлттэй эх бүхий програм хангамж ба програмчлалын хэлний цогц бөгөөд орчин үеийн статистикчдын эзэмшсэн байх үндсэн мэдлэг чадваруудын нэг болохын хамтаар дэлхий дахинд түүний хэрэглээ цаашид ч өсөн нэмэгдэх хандлагатай байна. Өнөөгийн байдлаар дэлхийн тэргүүлэх их сургуулиуд статистикийн сургалтын хөтөлбөр болон эрдэм шинжилгээ, судалгааны ажилдаа R програмыг өргөнөөр хэрэглэж байна. Тухайлбал дэлхийн тэргүүлэх нэр бүхий 20 их сургууль дахь статистикийн хөтөлбөрүүд бүгд уг програмыг хичээлийн төлөвлөгөөндөө тусгажээ.

R нь програмчлалын хэл учраас орчин үеийн тулгамдсан асуудал болох big data analysis, data mining зэрэгт нэн тохиромжтой програм хангамжуудын нэг болжээ. Мөн уг програм нь GNU General Public лицензийн дор үнэ төлбөргүйгээр түгээгддэг, голлох үйлдлийн системүүдийг (Windows, Linux, Mac, BSD болон бусад) бүрэн дэмждэг, сүүлийн үеийн статистик болон математикийн арга техникүүд тусгагдах нь хурдан (Matlab програмтай өрсөлдөхүйц) зэрэг олон сайн талуудтай юм.

Сургалтын хөтөлбөрийн онцлог Энэхүү хөтөлбөрийн гол онцлог нь R-ыг програмчлалын хэл талаас нь авч үзсэн явдал юм. Хөтөлбөр нь нийт $7 = 1 + 4 + 2$ сэдвээс бүрдэх бөгөөд эхний сэдвээр R програм, түүний давуу болон сул тал, R хэлийг онцлох шалтгаан зэргийг танилцуулах бол дараагийн дөрвөн сэдвээр R хэлийг практикт хэрэглэхэд нэн тэргүүнд шаардагдах мэдлэгийг өгнө. Харин сүүлийн хоёр сэдвээр R програмын зарим практик хэрэглээг тухайлан авч үзнэ.

Сургалтын гол хэрэглэгдэхүүн болох үзүүлэн слайдуудыг L^AT_EX програмын Beamer классыг ашиглаж чанарын өндөр төвшинд бэлдснээс гадна агуулгыг мэргэжлийн төвшинд боловсруулж өгсөн. Үүнээс гадна слайданд орсон R хэл дээрх кодыг бүрэн эхээр нь интернэтэд <https://goo.gl/RNCFHR> хаягаар байршуулсан. Мөн <https://goo.gl/zUs7Ns> хаягаар орж код болон нэр бүхий 3 сэдвийн үзүүлэн слайдыг татан авч ашгийн бус зорилгоор ашиглах боломжтой.

Сэдвийн жагсаалт ба агуулгын товчоо

0 Танилцуулга

- Статистикийн програмууд, тэдгээрийн зарим үзүүлэлтээрх харьцуулалт:
 - хэрэглэгчдийн тоо
 - ажлын байран дээрх эрэлт
 - статистик шинжилгээнүүдийг тусгасан байдал
 - дэмжих үйлдлийн систем
 - лицензийн төрөл ба төлбөрийн хэмжээ
 - мэргэжилтнүүдийн байр суурь
- Програмчлалын R хэлний тухай ерөнхий ойлголт
- R програмыг суулгах заавар
- R програм дээр ажиллах байдал

- R хэлний багц болон функц зэргийн баримтжуулалтыг үзэх тухай

1. Өгөгдөл оруулах ба гаргах, өгөгдлийн сантай холбох

- R хэлэн дэх өгөгдөлд холбогдох объектийн төрлүүдийн заримтай танилцана.
- Өгөгдлийг файлаас оруулахад ашиглагддаг `scan()`, `read.table()` функцүүд мөн өгөгдлийг файлд хадгалахад ашиглагддаг `write.table()` функцийг товч авч үзнэ.
- Өгөгдлийн сантай ажиллах суурь ойлголтыг MySQL өгөгдлийн санд тулгуурлан өгнө.

2. Оператор, функц ба хүрээлэл

- R хэлний үндсэн операторууд, тэдгээрийн хэрэглээтэй танилцана. Операторууд нь аливаа үйлдлийг хийж гүйцэтгэх үндэс суурь учраас зайлшгүй эзэмших шаардлагатай юм.
- Шинээр оператор тодорхойлоход шаардагдах мэдэгдэхүүнийг өгнө.
- Функц, түүнийг зарлах ба ашиглах болон холбогдох бусад зүйлсийг авч үзсэнээр R хэлний функцүүдийг илүү үр дүнтэйгээр ашиглахаас эхлэн улмаар өөрийн функцийг зохиож сурна.
- Хүрээллийн тухай ойлголтыг өгнө. Хүрээллийн тухай ойлголтыг авснаар объектуудыг (хувьсагч ба функц) алдаа мадаггүй ашиглаж сурна.

3. Урсгалыг удирдах тухай

- `if` оператор болон `ifelse`, `switch` функцүүдийг авч үзнэ. Эдгээр нь тухайн нөхцөл байдлаас хамааруулан өөр өөр үйлдлүүдийг хийж гүйцэтгэх боломжийг олгодог.
- Давталтын операторууд (`for`, `while`, `repeat`) болон функцүүдээс (`apply`, `tapply`) үзнэ. Үүнийг эзэмшсэнээр олон дахин давтагдах ижил эсвэл төстэй үйлдлүүдийг хялбар програмчлах боломж нээгдэнэ.

4. Хэрэгцээт зарим чухал ойлголт болон алдааны эргэн тойронд

- R програмын ажиллах зарчмыг схемээр авч үзнэ. Үүний дүнд R програмын талаар илүү тодорхой төсөөлөлтэй болно.
- Нэрийн огторгуйн тухай ойлголт болон объектийн нэршил, нөөцлөгдсөн үгс, чухал тогтмолуудтай танилцана.
- Нэр бүхий объектийн оршин буй эсэхийг шалгах болон түүнийг устгах аргыг үзнэ.
- Код дахь алдааг олох, алдаанаас сэргийлэхтэй холбогдох зөвлөмжийг өгнө.
- Зарим нийтлэг алдаа, тэдгээрийн шалтгааныг авч үзнэ.

5. Өгөгдөлтэй ажиллах

- Статистик өгөгдлийг агуулах үндсэн объект болох датафрэйм дээр дараах зүйлсийг авч үзнэ. Үүнд:
 - 1 хувьсагч зохиох
 - 2 хувьсагчийн нэрийг өөрчлөх
 - 3 өгөгдөл устгах
 - 4 өгөгдөл ялгаж авах
 - 5 өгөгдөл нэгтгэх
 - 6 төрөл хувиргах
 - 7 эрэмбэлэх
 - 8 хөрвүүлэх
 - 9 шошго зүүх
 - 10 кодлох
 - 11 орхигдсон утгатай ажиллах
 - 12 өгөгдлийн бүтэц зохион байгуулалтыг өөрчлөх

6. Статистик дахь зарим хэрэглээ

- Тархалтуудтай холбогдох функцүүдийг авч үзнэ. Ингэснээр зарим тархалтын нягтын функц, тархалтын функц, квантилыг олох санамсаргүй тоонууд гарган авах чадвартай болно.
- Өргөн хэрэглэгддэг тайлбарлагч статистикуудын заримыг хэрхэн тооцоолохыг үзнэ.
- Гистограм, хайрцган диаграм, Q-Q диаграм зэргийг байгуулахад ашиглагдах кодыг тайлбарын хамтаар өгнө.
- Статистик шинжүүрүүдийн төлөөлөл болгон дунджийн тухай таамаглал шалгах t шинжүүрт холбогдох t .test функцийг танилцуулна.
- Хувьсагчдын хамаарлыг шинжлэхтэй холбогдуулан Пирсоны корреляцийн коэффициентийг олох сог болон шугаман регрессийн шинжилгээний lm функцүүдийг холбогдох нэмэлт тайлбарын хамтаар авч үзнэ. Тухайлбал lm функцийн хувьд шугаман загварыг хэрхэн томъёолж өгөхийг дэлгэрэнгүй танилцуулна.

Жич

- Сэдэв тус бүрийн үргэлжлэх хугацаа 90 минут байна.
- R хэл дээрх жишээ код болон холбогдох бусад файлуудыг интернэтэд <https://goo.gl/RNCFHR> хаягаар бас шууд татан авах боломжтойгоор <https://goo.gl/zUs7Ns> хуудсанд байршуулсан.
- Танилцуулга болон 1 ба 4 дүгээр сэдвийн үзүүлэн слайдыг ашгийн бус зорилгоор ашиглах бол харин бусад слайдуудыг зохиогчийн зөвшөөрөлгүйгээр ашиглах, хуулбарлах, түгээхийг хориглоно.

Г.Махгал – МУИС-ийн ХШУИС-ийн ХМТ-ийн багш
цахим шуудангийн хаяг: makhgal@seas.num.edu.mn
цахим хуудасны хаяг: www.galaa.mn