

Studiju kursa nosaukums	Lietišķā statistika un datu zinātne
Studiju kursa kods	MateW001
Zinātnes nozare	Matemātika
Zinātnes apakšnozare	Varbūtību teorija un matemātiskā statistika
Kredītpunkti(ECTS)	15
Kopējais kontaktstundu skaits	150
Lekciju stundu skaits	68
Semināru un praktisko darbu stundu skaits	82
Laboratorijas darbu stundu skaits	0
Studenta patstāvīgā darba stundu skaits	225
Kursa īstenošanas valoda	Latviešu un angļu
Kursa apstiprinājuma datums	01.04.2025
Atbildīgā struktūrvienība	Eksakto zinātņu un tehnoloģiju fakultāte

Kursa izstrādātājs(-i)

Dr. math., prof. Jānis Valeinis
Ph.D., doc. Leonora Pahirko

Priekšzināšanas

Kursa apguvei nepieciešamās priekšzināšanas atbilst studiju programmas uzsākšanas nosacījumiem un vispārējām zināšanām, prasmēm un kompetencēm, kas apgūtas iepriekšējā izglītības līmenī.

Studiju kursa anotācija

Gandrīz katrā zinātniskā pētījumā nepieciešams veikt dažādu datu ieguvu, apstrādi un analīzi. Izšķiroša loma veiksmīgai datu analīzei ir korekta un pētāmai problēmai atbilstošu datu ieguve, pamatotu statistisko vai datu zinātnes metožu izmantošana, rezultātu korekta interpretācija un secinājumu formulēšana. Papildu modulī tiks apskatītas tēmas trīs sekojošos virzienos:

- Datu apstrāde un vizualizācija ar programmēšanas valodu R (3 ECTS);
- Lietišķās statistikas un datu zinātnes metodes (6 ECTS);
- Daudzdimensiju statistiskās un mašīnmācīšanās metodes (6 ECTS).

Statistiskā programmēšanas valoda R ir mūsdienās viena no plašāk izmantotajām programmām statistiķu un datu zinātnieku vidū, kurā nepārtraukti tiek izstrādātas jaunas zinātniski pamatotas bibliotēkas un pilnveidotas esošās. Moduļa ietvaros tiks aplūkotas gan dažādas aktuālas datu analīzes metodes, gan analizēti dažādi tipiski datu piemēri, kuri sastopami ekonometrijā, finanšu sfērā, medicīnā kā arī socioloģiskajos pētījumos.

Mērķis

Studiju moduļa mērķis ir sniegt zināšanas par lietišķās statistikas un datu zinātnes, tai skaitā mašīnmācīšanās metodēm, kuras nepieciešamas pētniecisko datu apstrādē un analīzē.

Uzdevumi:

- apgūt R programmēšanas valodas pamatus, programmēšanas pamatprincipus un sintaksi; apskatīt dažādas datu apstrādes un vizualizācijas metodes un pieejas;
- veikt simulācijas un reālu datu statistisko analīzi, izmantojot dažādas statistiskās un

mašīnmācīšanās metodes ar programmu R;

- apgūt plašāk izmantotās klasiskās, neparametriskās un robustās statistikas metodes.
- iepazīties ar pārraudzītajām datu zinātnes metodēm, lai prognozētu vai izskaidrotu novēroto atkarīgo mainīgo (regresija, tuvāko kaimiņu metode, lēmumu koki, neironu tīkli).
- apgūt nepārraudzītās metodes, kuru uzdevums ir atklāt un apraksīt datu struktūru un izmantot to datu kopas reducēšanā (galveno komponentu analīze, atbilstības analīze, klasteranalīze).
- apgūt daudzdimensiju statistikas metodes, dimensiju reducēšanas tehnikas un klasifikāciju uzdevumu risināšanas metodes.

Kursa plāns Pilna laika: klātiene

- 1.Programmēšanas valodas R pamati, pamatprincipi un sintakse 2L 4Pd
 - 2.Datu tipi un datu ielasīšana 2L 4Pd
 - 3.Darbs ar datu tabulām ar bibliotēku dplyr 2L 4Pd
 - 4.Datu vizualizācija ar bibliotēku ggplot2 2L 4Pd
 - 5.Datu apstrāde ar bibliotēku tidyr 2L 4Pd
 - 6.Varbūtību teorijas pamatjēdzieni, gadījuma lielumi, to sadalījumi un raksturlielumi 2L 2Pd
 - 7.Statistiskā populācija un gadījuma izlasei. Lielo skaitļu likums un centrālā robežteorēma 2L 2Pd
 - 8.Aprakstošā statistika 4L 4Pd
 - 9.Parametru novērtētāji un to atrašanas metodes 2L 2Pd
 - 10.Statistiskās hipotēžu pārbaudes pamatprincipi. Efektu un izlašu apjomu noteikšana 4L 4Pd
 - 11.T-tests vienas un divu izlašu gadījumā. Ticamības intervāli 2L 2Pd
 - 12.Datu normalitātes pārbaude 2L 2Pd
 - 13.ANOVA metode 2L 2Pd
 - 14.Kategoriālu datu analīze (šķērstabulas, Hī-kvadrātā testi) 2L 2Pd
 - 15.Neparametriskas un robustas statistiskās metodes 2L 2Pd
 - 16.Klasiskā vienkāršā un daudzfaktoru lineārā regresija 4L 4Pd
 - 17.Loģistiskā regresija 2L 2Pd
 - 18.Daudzdimensiju sadalījumi. Kopējais, marginālais, nosacītais sadalījums 4L 4Pd
 - 19.Sodītās regresijas metodes. Modeļu novērtēšanas metodes un optimizācija 4L 4Pd
 - 20.Galveno komponentu analīze. Faktoru un atbilstības analīze 4L 4Pd
 - 21.Lēmumu koki un gadījuma meži 4L 4Pd
 - 22.Klasifikatori (atbalsta vektoru mašīnas, lineārā diskriminantu analīze, naivais Beijesa algoritms) 4L 4Pd
 - 23.Klasteranalīze (tuvāko kaimiņu metode, hierarhiskās metodes, DBSCAN, u.c.) 4L 4Pd
 - 24.Neironu tīkli 4L 4Pd
 - 25.Studentu patstāvīgo darbu prezentācijas 4S
- Kopā 68L 4S 78Pd

Studējošo patstāvīgo darbu organizācijas un uzdevumu raksturojums

Studentiem patstāvīgi jāveic:

1. Lekciju laikā sniegto teorētisko zināšanu nostiprināšana, regulāri pārskatot lekciju materiālus un zinātnisko literatūru;
2. Praktisko iemaņu attīstīšana, atkārtojot un pārskatot praktisko nodarbību laikā veiktos uzdevumus;
3. Papildus literatūras avotu pētīšana pēc pasniedzēja norādījumiem.

Studiju rezultāti

Zināšanas:

1. Zina dažādus pētījumu dizainus un raksturo datu analīzes procesa gaitu un veicamos

soļus.

2. Pārzina R programmēšanas valodas sintaksi, funkcijas un svarīgākās bibliotēkas.
3. Definē moduļa ietvaros aplūkotās klasiskās, neparametriskās un robustās statistiskās un datu zinātnes metodes.
4. Nosauc daudzdimensiju sakarību modelēšanas un dimensiju reducēšanas metodes.
5. Atšķir pārraudzītās un nepārraudzītās mācīšanās metodes.
6. Pārzina kursā aplūkotās mašīnmācīšanās metodes un algoritmus.

Prasmes:

7. Prot ievākt un korekti sagatavot pētījuma datu kopu statistiskajai analīzei programmā R.
8. Aprēķina nepieciešamo paraugkopas apjomu dotam pētījuma dizainam un plāno atbilstošus datu apstrādes un analīzes soļus.
9. Realizē teorētiski apgūtās metodes datu analīzei programmā R un veido vizualizācijas.
10. Prot aprakstīt nepieciešamos un īstenotos datu apstrādes soļus un iegūtos datu analīzes rezultātus.
11. Risina uzdevumus un ar datiem saistītas problēmas, izmantojot dažādas statistiskās un mašīnmācīšanās metodes.

Kompetence:

12. Veido un pielāgo kodus sava pētījuma vajadzībām programmēšanas valodā R.
13. Izstrādā sava pētījuma dizainu, datu apstrādes un analīzes plānu.
14. Izvēlās pētījuma problēmas nostādnei atbilstošas datu apstrādes un analīzes metodes.
15. Spēj izprast un izvērtēt citu pētnieku veiktu un zinātniskajā literatūrā aprakstītu datu statistisko analīzi.
16. Orientējas statistiskās un mašīnmācīšanās metodēs dažādu datu analīzes kontekstā.

Prasības kredītpunktu iegūšanai

Starppārbaudījumi:

1. Mājasdarbi par programmu R – 10%
2. Mājasdarbi par lietišķo statistiku un datu zinātnei - 10%
3. Mājasdarbi par datu zinātnes problemātiku - 10%
4. Gala darbs: atskaite (10%) un prezentācija par semestra sākumā izvēlētu datu kopu (15%)

Noslēguma pārbaudījumi:

5. Rakstisks eksāmens 1. daļa par R programmu – 15%
6. Rakstisks eksāmens 2. daļa par matemātiskās statistikas jēdzieniem – 15%
7. Rakstisks eksāmens 3. daļa par datu zinātnes metodēm un mašīnmācīšanās algoritmiem – 15%

Studiju rezultātu vērtēšanas kritēriji

Studiju kursa apguve tā noslēgumā tiek vērtēta saskaņā ar Latvijas Republikas normatīvajiem aktiem, vadoties pēc šādiem kritērijiem: iegūto zināšanu apjoms un kvalitāte, iegūtās prasmes, iegūtā kompetence atbilstīgi plānotajiem studiju rezultātiem.

Pārbaudījumu veidi	Studiju rezultāti															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Mājasdarbi par programmu R	+	+	-	-	-	-	+	-	+	+	-	+	-	-	-	-
2. Mājasdarbi par lietišķo statistiku un datu zinātnei	+	+	+	-	-	-	+	+	+	+	+	+	+	+	-	+

3. Mājasdarbi par datu zinātnes un mašīnmācīšanās metodēm	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+
4. Gala darbs: atskaite un prezentācija	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5. Eksāmens: 1. daļa par programmu R	+	+	+	+	+	-	+	+	+	+	-	+	-	-	-	-
6. Eksāmens: 2. daļa par matemātiskās statistikas jēdzieniem	-	-	+	-	-	-	-	-	-	-	+	-	-	+	+	+
7. Eksāmens: 3. daļa par datu zinātnes metodēm un mašīnmācīšanās algoritmiem	-	-	-	+	+	+	-	-	-	-	+	-	-	+	+	+

Obligāti izmantojamie informācijas avoti

1. Agresti, A., Franklin, C. A. (2013). Statistics: The Art and Science of Learning from Data (3rd ed.), Pearson Education.
2. Holmes, S. H., & Huber, W. (2018). Modern statistics for modern biology. Cambridge university press.
3. Thulin, M. (2024). Modern Statistics with R: From wrangling and exploring data to inference and predictive modelling. CRC Press.
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer.
5. Chan, B. K. (2015). Biostatistics for epidemiology and public health using R. Springer Publishing Company.
6. Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.". Pieejams tiešsaistē: <https://r4ds.had.co.nz/>

Papildus informācijas avoti

1. Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media.
2. Field, A. (2012). Discovering Statistics Using R.
3. Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). R for data science. O'Reilly Media.
4. Chang W. (2013). R Graphics cookbook: practical recipes for visualizing data, O'Reilly Media.
5. Xie, Y., Allaire, J. J., & Grolemund, G. (2018). R markdown: The definitive guide. CRC Press
6. Izenman, A. J. (2008). Modern multivariate statistical techniques. Regression, classification and manifold learning. Springer.
7. Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis. Upper Saddle River, NJ: Prentice hall.
8. Koch, I. (2013). Analysis of multivariate and high-dimensional data. Cambridge University Press.
10. Bain, L. J. and Engelhardt, M. Introduction to probability and mathematical statistics. Duxbury Classic, 2nd edition, 2000.
11. Wassermann, L. All of nonparametric statistics, Springer, 2006.

Periodika un citi informācijas avoti

1. R mājas lapa: <https://cloud.r-project.org/>
2. RStudio mājas lapa: <https://rstudio.com/>
3. Kuhn, M., Silge, S., Tidy Modeling with R, <https://www.tmw.org/>
4. Tidymodels (RStudio), <https://www.tidymodels.org/packages/>
5. Book on multivariate analysis. <https://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/>

Kursa ietvaros nav pieļaujams plaģiātisms un citi akadēmiskā godīguma pārkāpumi lūdzu iepazīties ar LU Noteikumiem par akadēmisko godīgumu. Šī kursa ietvaros ģeneratīvā mākslīgā intelekta (MI) rīku izmantošana ir atļauta izņēmuma gadījumos, ja to rakstiski ir norādījis un atļāvis šī kursa docētājs. Visos citos gadījumos nav pieļaujama mākslīgā intelekta ģenerētu materiālu (teksta, attēlu, audio, video u.c.) iesniegšana patstāvīgajos un grupu uzdevumos, pārbaudes darbā, eksāmenā vai jebkādā citā pārbaudes darbā, šāda veida materiālu iesniegšana tiks uzskatīta par neatļautu palīgīdzekļu izmantošanu