Az R-rel szebb a statisztika

Pšenák Péter[1], Pšenáková Ildikó[2]

{[1]petkoneo@gmail.com, [2]ildiko.psenakova@gmail.com
[1]Univerzita Komenského Bratislava, [2]Trnavská univerzita v Trnave

Absztrakt. A pedagógiai gyakorlatban sokszor szükséges különböző statisztikai számítások, elemzések használata. Mivel legtöbbször csak néhány adatról van szó a pedagógusok az MS Excel vagy a Calc programot használják. Ha azonban sok adatok kell feldolgozni, más rendszer után kell nézni. Itt jöhet képbe az RStudio statisztikai program, amely nemcsak a sok adat feldolgozásával tud megbirkózni, de hipotézisek tesztelésében, varianciaanalízis számításában és más statisztikai tesztek végrehajtásában is segítségére lehet a pedagógusnak.

Kulcsszavak: matematikai szoftver, RStudio, statisztika, oktatás

1. Bevezetés

A statisztikai számítások nem tartoznak sem a diákok, sem a legtöbb tanár kedvenc időtöltése közé. Ha munkájukban mégsem tudják ezeket elkerülni, leggyakrabban az általuk legjobban ismert szoftvert használják, és azzal igyekeznek minél előbb elvégezni a szükséges számításokat. A leggyakrabban ez az MS Excel, amely szinte már szabvánnyá vált, mert az irodai programcsomag (MS Office) részeként valamelyik verziója szinte bármely munkahelyen megtalálható.

Az iskolarendszerben is az MS Office használata az elfogadott és már az általános iskola alsó tagozatán kezdik oktatni az szövegszerkesztőt (MS Word). Később társul hozzá a táblázatkezelő (MS Excel) és a prezentációkészítő (MS Power Point) is. A középiskolákon folytatódik ez a trend és aprólékosabban oktatják a programcsomag részeinek használatát. Ebből is látható, hogy gyakorlatilag a diákok szinte az iskolai éveik alatt végig ezzel az „irodai csomaggal" ismerkednek, és egyes programjaival dolgoznak. Ebből kifolyólag természetes az is, hogy ha a tanárok ezeket oktatják, akkor ők is ezeket használják.

Mivel az MS Excel tartalmaz statisztikai funkciókat is, egyértelmű, hogy felhasználói elsősorban ezeket fogják szükség esetén használni. De mi tévők legyünk akkor, ha az Excel funkciói nem tartalmazzák a számításhoz megfelelőt, vagy túl sok adatot kell feldolgozni? Ebben az esetben más programot kell használni, amely speciálisan statisztikai számításokra orientált. Ilyen például az R és környezete az RStudio is.

2. Statisztika használata a pedagógiai kutatásban

A pedagógiai kutatás a kutatási tevékenységeknek egy specifikus típusa. Ez egy célzott és szándékos analitikus-szintetikus tevékenység, amely megfelelő módszertani eszközökkel és kutatási technikákkal tanulmányozza az objektív (tárgyilagos) pedagógiai (oktatási) valóságot. Célja, hogy új tudományos ismereteket szerezzen a pedagógiai valóság különféle jelenségeiről, tulajdonságairól, jellemzőiről és más nézőpontjairól. Eredményeként általános törvények, ismeretek vagy tudományos pedagógiai elméletek fogalmazódnak meg, melyek hozzájárulnak a pedagógiai elmélet és gyakorlat fejlődéséhez.

A pedagógiai kutatás *sajátosságai* az oktatási valóság bonyolultságából, sokrétűségéből és a pedagógia és egyéb társadalomtudományok pontatlanságából adódnak. Ezek közé sorolhatjuk a következőket:

- a pedagógiai jelenségek nem mérhetők pontosan (pl. a tanár és a tanuló oktatási tevékenységének jellemzői, eredményei, a tanár és a diák részének aránya az elért eredményekben, ...),
- a pedagógiai jelenségek dinamikusak és megismételhetetlenek – a tulajdonságok, képességek, feltételek fejlődnek és változnak, ezért szinte lehetetlen, hogy egy módszert teljesen azonos feltételekkel tudjunk használni kétszer (pl. ismételt vizsgálatoknál a kutatás csak közel azonos; a válaszadók és az eljárás lehet ugyanaz, de a kutatás során sok megismételhetetlen tényező is fennáll (pl. fizikai és lelki állapot, élmények, tapasztalat, időjárás, ...)
- a pedagógiai kutatás nem valósítható meg, ha bármilyen módon sértene vagy kárt okozna a válaszadónak, vagy ha nem felel meg az erkölcsöknek.

Ezért a pedagógia számos jelensége csak a hipotézisek szintjén marad, feltételezhetők, de nincs lehetőség pontos ellenőrzésre és egzakt megerősítésre.

A hipotézisvizsgálatokra ezekben az esetekben jól használható a következtetési (inferenciális) statisztika, melyre jól megfelel az R szoftver.

3. R

Az R egy szabad, nyílt forráskódú, ingyenesen használható, professzionális statisztikai szoftvercsomag, amelyben már kidolgozott eljárásokat tartalmazó függvények és munkakörnyezetek állnak rendelkezésre. [1]

Az R nyelv a https://www.r-project.org/ honlapról letölthető és telepíthető. Használata elsajátítása alap programozási készségekkel rendelkező felhasználóknak nem okoz gondot. Segítségként ajánlhatjuk a https://www.statmethods.net/, https://www.r-bloggers.com/, weboldalakon található tutoriálokat. A tanuláshoz további segítségül szolgálhatnak a https://www.udemy.com/ weboldalon található videók.

Az R lehetőséget ad tetszőleges célú további számolási algoritmusok kifejlesztésére, programozáshoz és alkalmas a kidolgozott algoritmusok tesztelésre is. Tehát az R egy programozási nyelv, de egyúttal programkörnyezet is. [2]

Ebből ered az R egyik legnagyszerűbb tulajdonsága, az aktív közösség, amely a szabadon írható és közzétehető forráskódjának köszönhető. A számtalan algoritmus szabadon elérhető csomagokban a webről letölthető (https://cran.r-project.org/) és ismételten felhasználható különböző adathalmazokkal. Gyakran előfordul, hogy elegendő csak a szükséges referenciákat átállítani más adathalmazra, és már meg is kaptuk a keresett eredményeket. A csomagok témakörönként vannak a honlapon feltűntetve, amelyek közül nagyon sok használható a pedagógiai gyakorlatban is, például adatbázis kezelés, idősorok számítása, grafikonok stb.

Néhány ok, miért érdemes R-t használni:

- több adatot képes kezelni, mint az MS Excel, mivel az R nem dolgozik grafikus környezetben, így nem kell az összes adatott megjeleníteni a felhasználó számítógép képernyőjén, ezért több adattal képes számolni,
- platform-független, ezért a program telepíthető Windows, Linux és Mac OS operációs rendszert használó számítógépekre is, sőt ma már szerver-oldali megoldások is elérhetőek,
- gyors és stabil, a parancsok írása viszonylag egyszerű és azonnal végrehajtódnak, a grafikonok ábrázolása is nagyon gyors. A számítások végeredményeit általunk létrehozott változókban menthetjük el, amiket bármikor megnézhetünk vagy használhatunk további számításokhoz.
- használható mátrixok és vektorok számolására, egyenletek megoldására stb.
- szabadon használható oktatásban, vállalati környezetben és otthon is,

- ingyenesen letölthető az RStudio is, amely megkönnyíti az R használatát, mivel tartalmaz egy kódszerkesztőt valamint hibakeresési és megjelenítési eszközöket. [3]

Az R alapjában egy interaktív statisztikai/adatelemző környezet, ahol a felhasználók utasításokat adnak ki az R konzolnak a parancsok végrehajtására és az eredmények is itt jelennek meg. Például (1. ábra) a „summary" parancs alap leíró statisztikákat számol a megadott adathalmazból (datacomp). Elég egyetlen parancs és egyszerre megjelenik több eredmény, ellentétben például az MS Excellel, ahol leggyakrabban minden értéket külön funkcióval kell kiszámítani.

```
> summary(datacomp)
     Female           Male      
 Min.   : 3333   Min.   : 3508  
 1st Qu.: 5153   1st Qu.: 4965  
 Median : 9252   Median : 8953  
 Mean   :12808   Mean   : 9856  
 3rd Qu.:22525   3rd Qu.:14714  
 Max.   :26383   Max.   :17489  
```

1. ábra: Példa leíró statisztikai eredményre

Az alap statisztikák számolására az R-ben elég néhány egyszerű parancsot használni és ezeket a konzol segítségével megadni. De ha egy összetett kísérlet eredményeit szeretnénk statisztikailag kiértékelni már ez nem elegendő és R szkript-et kell használni. A szkript fájlok kiterjesztésükről (*.R*) ismerhetők fel. Több egymást követő parancsot tartalmaznak, melyeket egyszerre lehet indítani és a szoftver egymás után hajtja végre őket. (2. ábra)

A megírt kódokat (szkripteket) külön mappában is tárolhatjuk, az R szoftver csak a futtatásukhoz szükséges. Ugyanazt a szkriptet többször is használhatjuk, például az ismételt kísérletnél, az adathalmaz változásánál vagy akár teljesen más adatokkal.

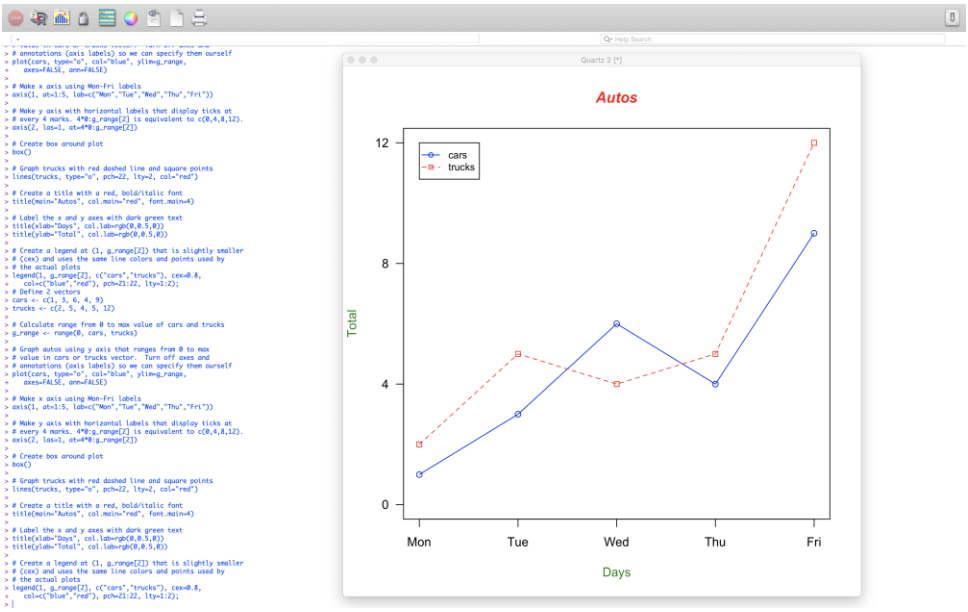
```
# Importing Dataset
dataset = read.csv(file.choose(),sep=";")
# Changing to date
dataset$Rok <- as.Date(dataset$Rok, format="%d/%m/%Y")
# Renaming Columns
colnames(dataset) <- c('Year', 'Female', 'Male')
# Removing Date - or kkeping it
datacomp <- dataset[2:3]
datacomp <- ts(datacomp, start=c(1989), end=c(2017), frequency = 1)
autoplot(datacomp)
datafemale <- dataset[2]
datafemale <- ts(datafemale, start=c(1989), end=c(2017), frequency = 1)
autoplot(datafemale)
lines(lowess(time(datafemale), datafemale), lwd=3, col=2)
datamale <- dataset[3]
datamale <- ts(datamale, start=c(1989), end=c(2017), frequency = 1)
plot(datamale)
lines(lowess(time(datamale), datamale), lwd=3, col=2)

# Trend is statistically significant?
library(tseries)
maletrend = factor(ifelse( datamale >= median(datamale), 1, 0))
femaletrend = factor(ifelse( datafemale >= median(datafemale), 1, 0))
runs.test(maletrend, a = 'less')
# trend significant Standard Normal = -5.1056, p-value = 1.649e-07
runs.test(femaletrend, a = 'less')
# trend significant Standard Normal = -5.1056, p-value = 1.649e-07


# Finding Trend statistically changes
# install.packages('changepoint')
require(changepoint)
plot(datamale)
#detect change in trend at different points of time
cm=cpt.mean(datamale) #abrupt change in mean
print(cm)
plot(cm)
autoplot(cpt.meanvar(datamale))
```

2. ábra: Példa egy szkript részletre a benne elhelyezett megjegyzésekkel.

Az R komplex adat vizualizáció lehetőséget is szolgáltat. A grafikus és ábrázolási képessége vitathatóan felülmúlja más programokét. Például a *dplyr* és a *ggplot2* ingyenesen letölthető csomagok nagy mennyiségű lehetőséget adnak az adatok manipulálásához és ábrázolásához, amellyel könnyítik megoldani a vizualizációt az analízis elvégzése közben. A grafikonokat külön ablakban nyitja meg (3. ábra) és külön fájlokba is elmenthetők. [4]



3. ábra: Példa egy grafikon megjelenésére.

Az R az adatokat és az elemzést külön mutatja be, ami a felhasználó számára lehetővé teszi az adatok pontosabb ellenőrzését, a felmerülő hibák javítását vagy az adatok megtekintését különböző elemzési pontokban. [5]

Persze, mint minden programnak, az R-nek is vannak negatívumai. Mivel az R alapja az 1960-as években használt programozási nyelvekből volt átvéve, egy viszonylag régi technológiának mondható. Jellemző rá, hogy az adatokat az operációs memóriában tárolja. Szerencsére a mai számítógépek memória kapacitása sokkal nagyobb, mint régebben, így ez a tulajdonsága már kevésbé (vagy egyáltalán nem) jelent problémát. [4]

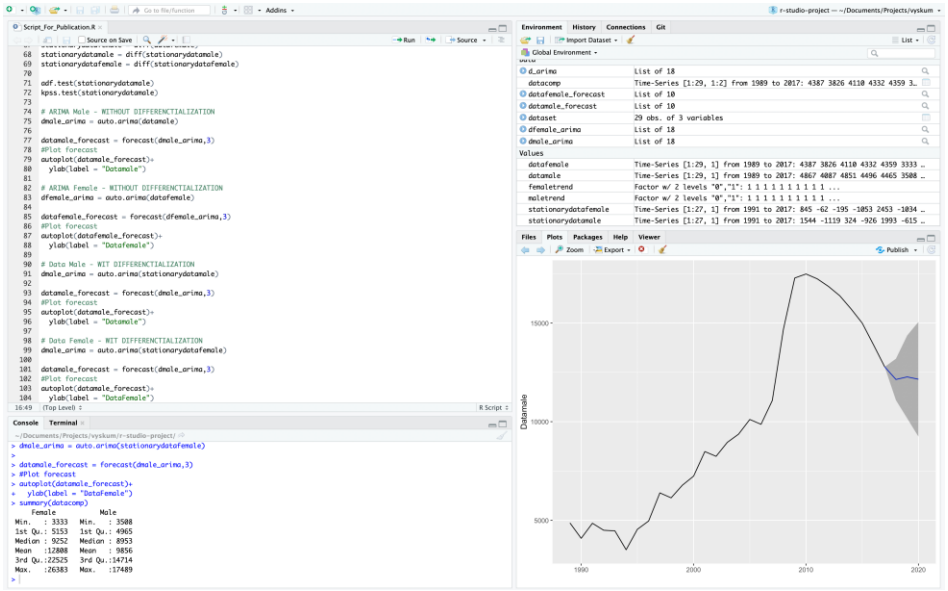
Az R nem biztonságos rendszer és nem lehet böngészőkbe vagy weboldalakba beágyazni. Az R-t nem lehet önmagában back-end szolgáltatásként sem használni. Ahhoz, hogy például az R számításokat végezzen el szükséges, hogy egy szerveren futó programozási nyelvből legyen meghívva. Ilyen nyelv lehet a PHP, Python, JavaScript stb.

Az említett szabadon elérhető R csomagjai között hibásak is találhatóak, melyek nem működnek tökéletesen, de ebben az esetben, sajnos, nincs kihez panasszal fordulni, hogy javítsa a hibákat.

Az R negatívumként említhető az a tulajdonsága is, ha egy szkriptnek csak pár sorát kell lefuttatni, akkor külön ablakot kell megnyitni és ezért az ablakok között gyakran kell váltogatni. Ezt a problémát küszöböli ki a már említett RStudio.

RStudio Integrált fejlesztői környezet, amely intuitíven használható. A képernyő felülete ablakrészekre van osztva, melyek mérete változtatható (4. ábra):

baloldalon felül: szövegszerkesztő, fájlba elmentendő, szkriptek (kódok) számára,
baloldalon alul: konzol a parancssoros módhoz,
jobb oldalon felül: fejlesztői környezett (workspace), előzmények (history),
jobb oldalon alul: segítség (help), csomagok (packages), grafikonok az ablakok.



4. ábra: RStudio felülete

4. Az R használata kutatásban

Az R lehetséges használatát egy kutatás eredményein keresztül mutatjuk be. A kutatás azt akarta felmérni milyen a munkakereső egyetemi végzetségének hatása a munkaadó választására.

A kutatáshoz szükséges adatokat a profesia.sk weboldal szolgáltatta, amelyen azok a személyek helyezik el életrajzukat, akik munkahelyet keresnek. Az életrajzok alapján a munkaadók válogatnak munkaerőt. A kutatáshoz 5 269 életrajz adatai voltak felhasználva. [6]

Szlovákiában jelenleg 35 felsőoktatási intézmény működik. A vizsgálat az informatika szakterületére koncentrált, vagyis olyan munkavállalók elhelyezkedésének eredményességét figyelte, akik a diplomájukat különböző egyetemeken, de valamelyik informatika ágazatában szerezték.

A kutatás statisztikai klasszifikáció segítségével igyekezet felmérni és bizonyítani vagy cáfolni többek között azt a hipotézist is, hogy a munkaadók előnyben részesítik azt a munkavállalót, aki egy jobb minősítéssel rendelkező egyetemen szerzett végzettséget.

Az adatok alapján a következő leíró statisztikákat lehet megjeleníteni (5. ábra). A táblázatok az R-Studio eredményei alapján készültek. A leíró statisztikák a már említett „summary” funkcióval készültek el. A táblázat első része az egyes egyetemeken végzett személyek a weboldalon közzétett életrajzainak számát tartalmazza, amit az R a „count” funkció segítségével készített el, mivel azok szövegként voltak reprezentálva az adathalmazban. Hogy ezeket hány munkaadó tekintette meg, vagyis a megtekintések értékelésénél automatikusan elvégezte a leíró statisztikákat. A táblázatban megjelenik a minimum, maximum, és az adatokból látható, hogy a maximális érték

és a harmadik kvartilis nagyon messze vannak egymástól, ezért feltételezhető, hogy az adathalmaz extrém értékeket is tartalmazhat.

Egyetem	*Mennyiség*
Ekonomická univerzita - Ökonómiai Egyetem (EU)	439
Többi (11 egyetem)	895
Slovenská technická univerzita - Szlovák Műszaki Egyetem Pozsony	1864
Technická univerzita v Košiciach - Kassai Műszaki Egyetem (TUKE)	964
Univerzita Komenského - Comenius Egyetem Pozsony (UK)	396
Žilinská univerzita - Zsolnai Egyetem (ZU)	711

Megtekintések száma	
Min:	1.000
1. kvartilis:	2.000
Medián:	4.000
Átlag:	6.553
3. kvartilis:	9.000
Max:	64.000

5. ábra: Leíró statisztika eredmények

A lineáris regresszió számításánál az R automatikusan kiszámolja a maradványértékeket, koefficienseket és más adatokat, mint például az R^2 értékét, az F statisztikát, p értékét, stb. (6. ábra).

A különböző koefficiensek t- és p- értékei a táblázatban külön jelennek meg, így könnyedén kiszűrhetőek azok, amelyek statisztikailag szignifikánsak.

A modellben szignifikáns értéket érnek el a TUKE, ZU és a Többi (11) egyetem. Annak ellenére, hogy az R^2 értéke alacsony, ezek az egyetemek hatással vannak a megtekintések számára, mivel a referencia egyetemhez (a táblázatban a korrekciós tényezőbe van jelen) viszonyítva mindegyikük alacsonyabb mennyiségű megtekintést ért el.

A 6. ábrán feltüntetett statisztikai eredmények alapján, bizonyítható a felvetett hipotézis, hogy a munkavállaló egyetemi végzetsége hatással van a munkában való elhelyezkedésre, mivel a munkaadók előnyben részesítik azt a munkavállalót, aki egy jobb minősítéssel rendelkező egyetemen szerzett végzettséget. [6]

A lineáris regresszió eredményei

Maradványértékek:	*Min*	*1. kvartilis*	*Medián*	*3. kvartilis*	*Max*
	-6.86	-4.30	-2.05	2.07	58.07

Koefficiensek:	*Becslés*	*Std. hiba*	*t-érték*	*p-érték*
Korrekciós tényező*	7.304	0.151	48.25	$<2e^{-16}$***
EU	-0.427	0.347	-1.23	0.22
Többi	-1.984	0.266	-7.47	$9.6e^{-14}$***
TUKE	-1.371	0.259	-5.29	$1.3e^{-7}$***
UK	0.552	0.362	1.53	0.13
ZU	-1.253	0.288	-4.35	$1.4e^{-5}$***

Standard hiba	6.45, 4 209 szabadságfoknál
R^2	0.0161
Beállított R^2	0.0152
F statisztika	17.2, ami 5 és 4 209 szabadságfoknál
p érték	$<2e^{-16}$
Szignifikancia kódolása:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

6. ábra: A lineáris regresszió

5. Befejezés

Lehet, hogy az R felhasználói környezete kevésbé barátságos, mint az MS Excel-é, de minden tulajdonságát és funkcióit összegezve az R „statisztikai tudása” messze meghaladja az Excel funkcióit. Nagyon könnyen és jól használható program és használatát gyorsan el lehet sajátítani.

A pedagógiai gyakorlatban nagyon hasznosnak bizonyulhat, ha gyorsan szeretnénk sok adatot elemezni és statisztikailag kiértékelni.

Irodalom

1. *GNU R: szoftver, programozási nyelv, közösség* (2018) http://r-projekt.hu/mi-az-r/ (utoljára megtekintve: 2018.10.30.)
2. Tóthmérész Béla: *Bevezetés az R használatába* http://biodiversity.unideb.hu/files/oktatas/Tothmeresz_Bevezetes-az-R-hasznalataba.pdf (utoljára megtekintve: 2018.10.13.)
3. *RStudio* (2018) https://www.rstudio.com/ (utoljára megtekintve: 2018.10.3.)
4. Krill Paul: *Why R? The pros and cons of the R language* (2015), InfoWorld, https://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html (utoljára megtekintve: 2018.10.13.)
5. Gebhart Kendra: *Understanding R programming over Excel for Data Analysis*, (2016) https://www.gapintelligence.com/blog/2016/understanding-r-programming-over-excel-for-data-analysis (utoljára megtekintve: 2018.10.3.)
6. Pšenák Peter, Káčer Ján: *A munkakereső egyetemi végzetségének hatása a munkaadó választására.* PEME 16. Budapest (Maďarsko): Professzorok az Európai Magyarországért Egyesülete, 2018. ISBN (online) 978-615-5709-03-6, s. 239-247