R szoftver: gyakorló feladatok

Mády Katalin

MTA Nyelvtudományi Intézet

1. feladat

Töltsük le a trans.RData fájlt innen:
clara.nytud.hu/~mady/courses/statistics/materials/trans.RData
Betöltés `load("konyvtar/trans.RData")` függvénnyel, NEM `read.table()`.

A mátrixban angol, ill. portugál, kb. 1500 szavas szövegek hossza van megadva, majd a másik nyelre való lefordítás utáni hosszuk.

Ellenőrizzük a normális eloszlást és a varianciahomogenitás meglétét.

```
z = trans$language == "English"
shapiro.test(trans$length[z])/(trans$length[!z])
leveneTest(trans$length~trans$language)
```

Ha nem teljesülnek a *t*-próba feltételei, Mann-Whitney-próbát alkalmazunk:
`wilcox.test(trans$length~trans$language)`

2. feladat

`ratings` adatmátrix a languageR könyvtárból:

1. Igaz-e, hogy az állatnevek gyakorisága alacsonyabb mértékű, mint a növényeké? És az ismertségük?

2. Különbözik-e a növény- és állatnevek egyes és többes számú gyakoriségi indexe?

Teszteljük minden esetben, hogy az adatok normális eloszlásúak-e, és hogy a varianciák homogének-e.

3. feladat

Regressziós egyenes ábrázolása a trans objektum angol és portugál szószámaira:
`lm(függőváltozó~függetlenváltozó)`
kimenet: *a* és *b* regressziós együtthatók

Érdemes az eredményt eltárolni egy változóban, mert így hozzáférünk a számított értékekhez:
`lmcoef = lm(ratings$Frequency~ratings$Length)`

`coef(lmcoef)` vagy `lmcoef$coefficients`: vektor a két együtthatóval.

`fitted(lmcoef)`: az egyeneshez igazított (hipotetikus) *y* értékek.

`resid(lmcoef)`: reziduumok, a hipotetikus *y* értékektől való eltérések.

Egyéb elérhető adatok listázása:
`str(lmcoef)`
`abline(intercept,slope)`, vagyis
`abline(coef(lmcoef))`
Az `abline()` függvény mindig egy már meglévő grafikonba rajzol

Hasznos függvények az ábrázoláshoz

Mindkettő már létrehozott grafikonhoz ad hozzá további információt.
Grafikon koordinátái „ismertek” és felhasználhatók az elhelyezésben.

`text(x,y,"my text")`: szöveg elhelyezése a grafikonban megadott pozícióban, pl.:
`text(9,6,"y(i)-mean(y)")`
Alapbeállítás: szöveg **középpontja** esik a megadott koordinátákra.

`legend()`: jelmagyarázat
Számos opció, kötelező argumentumok: pozíció
(”center”,”topleft”,”bottom” stb.), magyarázatok:
legend=c(”növény”,”állat”), szín vagy satírozás: col=c(”red”,”blue”), ha lwd (vonalvastagság) definiálva van, akkor vonal kerül elé, és az színes, stb.

4. feladat

`ratings` külön a növényekre és az állatokra eltérő színnel: x tengely: név hosszúsága, y-tengely: gyakoriság. Eljárás: először a növények adatpontjait ábrázoljuk, utána

```
par(new=T)
```

majd az állatok adatpontjait hozzáadjuk ehhez az ábrához. Figyelem: a tengelyek terjedelmét meg kell adni, mert az R automatikusan számolja ki az optimális szélső értékeket, és ezek eltérhetnek.

Készítsünk jelmagyarázatot (legend) a színek jelentéséről.

5. feladat

oz.csv: három különböző területen mérték újszülött hím és nőstény őzgidák tömegét és testhosszát. Ábrázoljuk a hím és nőstény őzek testtömegét a három mérési hely szerint dobozdiagrammal. A hím őzek tömegének doboza legyen világkoskék, a nőstényeké rózsaszín. Lássuk el az ábrát magyar nyelvű feliratokkal. Készítsünk jelmagyarázatot (legend) a színek jelentéséről.

Ellenőrizzük a két nemen belül a normális eloszlás és a varianciahomogenitás feltételét a tömeg és a testhossz adataira tesztekkel. A két csoportra jellemző eloszlást ábrázoljuk egyazon sűrűségfüggvényben. Az átlag értékét jelöljük vízszintes egyenessel az x-tengelyen.

Teszteljük a megfelelő próbával, hogy az őzgidák neme és születési helye befolyásolja-e méretüket. Ha a mintaszámok azonosak, a varianciaanalízis függvénye `aov()`, ha nem, akkor `lm()`. Post-hoc teszt: `TukeyHSD()`. A Kruskal-Wallis próba függvénye `kruskal.test()`, a Mann-Whitney-é `wilcox.test()`.