Korrelációs együtthatók
Ábrázolás és csomagok telepítése az R-ben

Változók típusai

Független változó: A minta jellemzésére szolgáló mérőszám, gyakran a minta összeállításának szempontja is.
Független változók: kategoriális változók (nem, anyanyelv), ordinális változók (kor: fiatal, közép, éltes), metrikus változók (évek száma, kockán látható pöttyök).

Függő változó: A kísérlet hipotéziseinek tesztelésére szolgáló mérőszám.
Függő változók: számszerű változó, statisztikai teszt alapja, ordinális (természetességi ítéletek) vagy metrikus (alapfrekvencia, gyakoriságok).

Változók összefüggései

Ha két legalább ordinális függő változónk van, vagy egy legalább ordinális független és egy legalább ordinális függő változónk, a kettő összefüggései kifejezhetőek korrelációs együtthatókkal. Lehetséges összefüggések a független és b függő változók között:

- Ahogy a növekszik, úgy nő b is.
- Ahogy a növekszik, úgy csökken b.
- A két változó között nem látszik összefüggés.

Változók összefüggései

Ha két legalább ordinális függő változónk van, vagy egy legalább ordinális független és egy legalább ordinális függő változónk, a kettő összefüggései kifejezhetőek korrelációs együtthatókkal. Lehetséges összefüggések *a* független és *b* függő változók között:

- Ahogy *a* növekszik, úgy nő *b* is.
- Ahogy *a* növekszik, úgy csökken *b*.
- A két változó között nem látszik összefüggés.

Erős korreláció lehetséges értelmezései:

1. *b* függő változó *a* független változó következménye. Klasszikus példa: iskolapadban eltöltött évek száma és teljesítmény egy műveltségi teszten.
2. A két változó egy harmadik, figyelmen kívül hagyott vagy még ismeretlen változó függvénye. Híres példa: gólyák száma és születések száma (mindkettő csökkenő tendenciát mutat).

Korreláció mértéke

Tartomány: -1 és $+1$ között.

- Ha a két változó szorosan összefügg, a korrelációs együttható $+1$-hez közelít.
- Ha a két változó között szoros fordított összefüggés áll fent, a korrelációs együttható -1-hez közelít.
- Ha a két változó között nincs összefüggés, az együttható értéke 0 körül mozog.

Együtthatók: Kendall-féle τ vagy Spearman-féle ρ ordinális adatokra, Pearson-féle r metrikus adatokra.

Korrelációs együtthatók: Kendall-féle τ (tau)

Ordinális adatok mérőszáma, -1 és $+1$ közötti tartományra esik.

- Előny: kis elemszám (n)esetén megbízhatóbb, mint Spearman-féle ρ.
- Hátrány: négyzete nem fogható fel determinációs együtthatóként (lásd ρ és r).

Eljárás: elemek sorrendjének „jósága”, osztva a lehetséges párok számával.

Proverzió (P): y vektor elemeinek száma, amelyek a várt sorrendbe illeszkednek.

Inverzió (I): a várt sorrendtől való eltérés.

Két halmaz, a halmaz sorrendjében az összetartozó értékek:
$a = [11, 23, 36, 44], b = [25, 12, 33, 49]$
proverziók és inverziók száma: $P_{b1} = 2, I_{b1} = 1, P_{b2} = 2, P_{b3} = 1$

lehetséges összehasonlítások száma: $\frac{n(n-1)}{2}$

$$\tau = \frac{P-I}{\frac{n(n-1)}{2}} = \frac{5-1}{6} = 0,66666666667$$

Spearman-féle *Rho*

Ha az egyik vagy mindkét változó ordinális: Spearman-féle ρ.

Eljárás: mindkét változó értékeit sorrendbe állítjuk, az eltérések különbségét (d) négyzetre emeljük és összegezzük, majd az alábbi képlet szerint számítjuk ki:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

A $\sum$ a görög S-nek megfelelő szigma, az összegzés szimbólumaként használják. Az n az elemszámot jelöli.

Példa

IQ és a tévénézéssel töltött órák száma.

IQ	óra	x sorszáma	y sorszáma	d	d^2
86	1	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Példa

IQ és a tévénézéssel töltött órák száma.

IQ	óra	x sorszáma	y sorszáma	d	d^2
86	1	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

$$\rho = 1 - \frac{6*194}{10*(100-1)} = 1 - \frac{1164}{990} = -0,175757$$

Kovariancia

Kiindulás: ha két mérőszám (x, y) függ egymástól, akkor ha egy konkrét x_i érték eltér az x halmaz átlagától $(\bar{x})$, akkor a hozzá tartozó y_i is el fog térni az y halmaz átlagától $(\bar{y})$ pozitív vagy negatív irányba.

Eljárás:

1. Kiszámítjuk minden egyes pont átlagtól való eltérését.
2. Összegezzük az eltéréseket.
3. Ha a teljes populáció rendelkezésünkre áll: összeget elosztjuk az összes elem számával, n-nel.
4. Ha populáció helyett mintával dolgozunk, n-*1*-gyel osztunk, mert a minta kovarianciája csak közelítő értéket ad.

$$\sigma_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})$$

Kovariancia alsó értéke 0.

Kovariancia standardizálása

Probléma: az érték függ a minta méretétől és az értékek átlagtól való eltérésétől, vagyis a szórástól.

Koordinátarendszer origója $\leftarrow$ átlag. Kovariancia osztása a két változó szórásával.

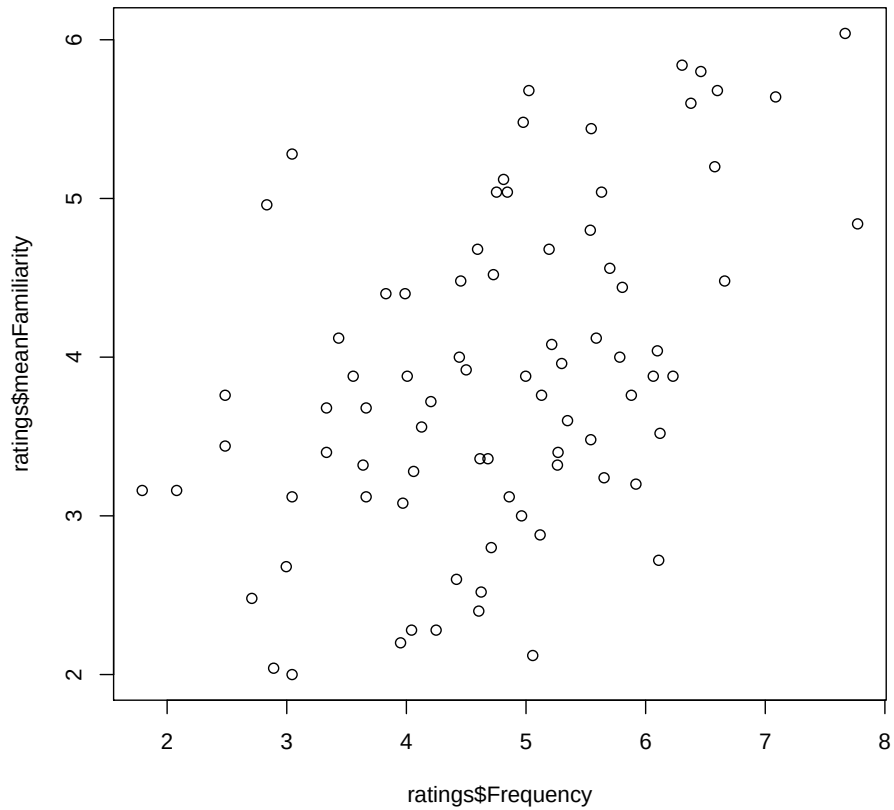
Pearson-féle r:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

A Pearson-féle r normális eloszlású parametrikus adatok korrelációs együtthatója.

Példa

Angol beszélőket megkérdeztek, hogy mennyire jól ismernek bizonyos növényeket és állatokat. Az ismertségi fokot 7-es skálán adták meg. Függ-e az ismertségi fok a szavak (növény- vagy állatnév) gyakoriságától?



Példa

A szógyakoriság parametrikus. A 7-es skálát a gyakorlatban parametrikusnak szokás tekinteni, ha az értékek ekvidisztánsak, azaz ha bármely két szomszédos érték távolsága egyenlőnek tekinthető.

Az ábrán látszik, hogy minél nagyobb x, annál nagyobb $y \rightarrow$ pozitív korreláció:

$r = 0,48$

$\rightarrow$ erős pozitív korreláció.

Korreláció erőssége

r értelmezése

$+0{,}70$ és 1 között: nagyon erős pozitív összefüggés
$+0{,}40$ és $0{,}69$ között: erős pozitív összefüggés
$+0{,}30$ és $0{,}39$ között: közepesen erős pozitív összefüggés
$+0{,}20$ és $0{,}29$ között: gyenge pozitív összefüggés
$+0{,}19$ és $-0,19$ között: nincs vagy elhanyagolható összefüggés
$-0,20$ és $-0,29$ között: gyenge negatív összefüggés
$-0,30$ és $-0,39$ között: közepesen erős negatív összefüggés
$-0,40$és $-0,69$ között: erős negatív összefüggés
$-0,70$ és -1 között: nagyon erős negatív összefüggés

Determinációs együttható

d**:** a mért y értékek varianciájából mekkora részt (hány százalékot) magyaráz meg a regressziós becslések varianciája, vagyis a regressziós becslés milyen mértékben mutatja meg a b változó viselkedését.

Ha az összefüggés lineáris, az együttható: r^2.

Előbbi példa:

$d = r^2 = 0,48^2 = 0,2304$

Azaz: az összefüggés 23%-ban magyarázza meg y változó viselkedését x függvényében.

R

Adatok beolvasása az R-be

Az adatokat többnyire más szoftverrel állítottuk elő (E-Prime, Praat, manuális lejegyzés stb). Ezek beolvasása:
`read.table()`, alapértelmezett beállításai:
`read.table(file, header = FALSE, sep = "", dec = ".")`
`header`: ha első sor eggyel kevesebb elemet tartalmaz, automatikusan TRUE lesz az érték.
`sep`: szóköz vagy tab, problémás lehet, ha vannak üres cellák, a pontosvessző megbízhatóbb.
`dec`: ha közép-európai kódolású szoftvert használunk, a decimális vessző, tehát `dec = ","`

Feladat: töltsünk le egy adatfájlt innen, majd olvassuk be az R-be `soc` változóként:
http://clara.nytud.hu/~mady/courses/statistics/materials/soc.dat
Érdemes a felhasználói név alatt az R számára létrehozott könyvtárba menteni.

Néhány hasznos függvény

`ls()`: R munkamemóriában tárolt objektumok (változók).

`names(soc)`: oszlopban tárolt változók neve.

`head(soc)`: első hat adatsor, `tail(soc)`: utolsó hat.

data.frame változóira (oszlopaira) hivatkozás: soc$valtozo, ahol valtozo az oszlop nevével azonos.

Feladatok

Helynevek gyakorisága:
`barplot(table(soc$loc))`

Százalékos gyakoriság:
Eljárás: table() adatait osztjuk az összes adat számával, megszorozzuk 100-zal.
`barplot(table(soc$loc)/length(soc$loc)*100)`

Összes adat megadása: több alternatív eljárás

- `length(soc$loc)`: egydimenzionális vektorokra érvényes: vagy egy `a=c("Budapest","Szeged","Kiskunlacháza"...)` típusú vektor, vagy egy adatmátrix oszlopa, pl. `soc$loc`.
- `nrow(soc)`: kétdimenziós objektum (mátrix vagy adatmátrix) sorainak száma. Itt: megfelel kísérleti személyek számának.
- `dim(soc)[1]`: kétdimenziós objektum sorainak [1] és oszlopainak [2] száma.

Feladat

Kor eloszlása

- kördiagrammal: `pie(table(soc$age))`.
- oszlopdiagrammal: abszolút: `barplot(table(soc$age))`, százalékos `barplot(table(soc$age)/length(soc$age)*100)`.
- hisztogrammal: `hist(soc$age)`.

Interkvartilis félterjedelem megállapítása a boxplot létrehozásához kiszámolt értékek mentésével:
`h = boxplot(soc$age)`
Objektum típusa `list`, vektorok hossza eltérő lehet $\leftrightarrow$ mátrix, dataframe, ahol minden oszlop egyforma hosszú. Vektor lekérdezése: objektum$vektornév, itt: `h$stats`.

`ik = (h$stats[4]-h$stats[2])/2`
azaz: 4. kvartilis és 2. kvartilis különbségének a fele.

Feladat

Két dobókockával való 10, 100, 1000 dobálás összege: módusz, medián, átlag, ezek eloszlásának ábrázolása hisztogrammal és dobozdiagrammal.

Feladat

Két dobókockával való 10, 100, 1000 dobálás átlaga.

Hipotézis: leggyakoribb szám 7 lesz. Miért?

Feladat

Két dobókockával való 10, 100, 1000 dobálás átlaga.

Hipotézis: leggyakoribb szám 7 lesz. Miért?

Legtöbb számkombinációval elérhető szám (1+6, 2+5, 3+4, 4+3, 5+2, 6+1).

```
dobas1 = sample(1:6,10,replace=T)
dobas2 = sample(1:6,10,replace=T)
```

Generáltunk két vektort, amelyek tíz, 1 és 6 közötti, véletlenszerűen kiválasztott elemet tartalmaznak. Összegük:

```
dobas.ossz = dobas1+dobas2
```

Feladat

- módusz:
  `tab = table(dobas.ossz) # táblázat létrehozása`
  `tab.sort = sort(tab) # gyakoriságok sorrendje`
  `tab.leng = length(tab) # összes érték száma`
  `tab.max = tab.sort[tab.leng] # táblázat utolsó (=` `legnagyobb, n-edik) eleme.`
  Ugyanez egy sorban:
  `sort(table(dobas.ossz))[length(table(dobas.ossz))]`
- medián: `median(dobas.ossz)`.
- átlag: `mean(dobas.ossz)`.
- eloszlások: `barplot(table(dobas.ossz))`,
  `hist(dobas.ossz)`, `boxplot(dobas.ossz)`,
  `plot(density(dobas.ossz))`.

density: sűrűségfüggvény, interpolált értékekkel $\Rightarrow$ nem létező értékek (itt: tört számok) előfordulási valószínűségét is megkapjuk.

Programcsomagok telepítése az R-ben

Mivel az R nyílt forráskódú szoftver, bárki fejleszthet hozzá csomagokat. Elérhető csomagok listája az R tükör oldalakon, Packages menüpont alatt.

Ha többen használják a számítógépet, az R-et érdemes rendszergazdaként megnyitni és a csomagokat úgy telepíteni. Így a csomagok minden felhasználó számára elérhetőek.

Telepítés interneten keresztül:
`install.packages("languageR")`

Csomag betöltése

Betöltés (R megnyitása után minden egyes alkalommal):

`library(languageR)`

Ha $>$ jelet kapunk „válaszként”, akkor a csomag betöltődött az R-be.

Ellenőrzés:
`search()`
aktuálisan betöltött csomagok listája.

Elérhető objektumok listája és rövid leírása: languageR telepítésének könyvtárában, az INDEX fájlban.

Kétváltozós összefüggések ábrázolása

languageR könyvtár `ratings` objektuma.
`names(ratings)`: változók (oszlopok) neve.
`head(ratings)`, `tail(ratings)`: első hat és utolsó hat sor (= eset, record).
Hipotézis: A rövidebb állat- és növénynevek gyakoribbak.

Kétváltozós összefüggések ábrázolása

languageR könyvtár `ratings` objektuma.
`names(ratings)`: változók (oszlopok) neve.
`head(ratings), tail(ratings)`: első hat és utolsó hat sor (= eset, record).
Hipotézis: A rövidebb állat- és növénynevek gyakoribbak.

Ábrázolás:
`plot(ratings$Length,ratings$Frequency)`
első érték: x-tengely, második érték: y-tengely.

Kétváltozós összefüggések ábrázolása

languageR könyvtár `ratings` objektuma.
`names(ratings)`: változók (oszlopok) neve.
`head(ratings), tail(ratings)`: első hat és utolsó hat sor (= eset, record).
Hipotézis: A rövidebb állat- és növénynevek gyakoribbak.

Ábrázolás:
`plot(ratings$Length,ratings$Frequency)`
első érték: x-tengely, második érték: y-tengely.

Korrelációs együtthatók:
`cor.test(ratings$Length,ratings$Frequency)`
`method = "pearson" -- default, alternatív:`
`"spearman", "kendall")`

R érzékeny a kis- és nagybetűkre! Elnevezést lehet rövidíteni, ha különbség egyértelmű, itt elég "p", "s", "k".

Grafikus paraméterek

Rengeteg paraméteren lehet állítani. Hogyan lehet ezekről tudni?

- grafikus parancs opcionális argumentumai. Lekérdezés: `?boxplot, ?plot, ?barplot` stb.
- parancsok súgója gyakran utal további hasznos parancsokra (`line(),title(),abline()` stb.
- `par()`: rengeteg paraméter, pl. tengelyek feliratozása (felirat mérete, elhelyezése, egységek mérete), tengelyek aránya stb.

Első lépés a súgó. Felépítés: (1) kötelező és opcionális argumentumok listája, (2) argumentumok rövid magyarázata, (3) részletek: többnyire innen derül ki a releváns infó, ha még nem ismerjük a parancsot, gyakran hivatkozások is, (4) lásd még - hasznos, esetenként hasznosabb, további parancsok, (5) példák - ezek általában túl bonyolultak, ezért nem túl hasznosak.

Néhány hasznos paraméter

- **xlab, ylab:** "x-tengely felirata", "y-tengely felirata".
- **main:** "Ábra címe".
- **xlim, ylim:** Ábrázolt értékek tól-ig. Főleg y-tengelynél fontos, ha összehasonlítható ábrákat akarunk. Pl százalékos ábrázolásnál `ylim = c(0,100)`, azaz 0–100%-ig. Egyenlőségjel előtti szóköz opcionális.
- **col:** színek, vagy névvel, vagy számmal. Pl. `col=2` és `col="red"` azonosak.
- **cex:** `cex.main,cex.axis,cex.names` stb. Default: cex=1, ehhez képest cím, mérőszámok, címkék betűmérete nagyobb (1.3, 1.7) vagy kisebb (0.7).

Ábra mentése

Alapeset: mentés pdf-ként vagy postscript fájléként.
pdf: LaTeX-felhasználóknak hasznos, ha pdflatex-et használnak.
eps: Word-ben és LaTeX-ben egyaránt használható.

`dev.print("célfájl",device=pdf|postscript)`

postscript fájlok alapértelmezése: fektetett, horizontal=T $\Rightarrow$
mentés: `horizontal=F` vagy `dev.copy2eps()`.

Boxplotok

Ábrázolás módja: y-tengely: függő változó interkvartilis eloszlása, x-tengely: csoportok, esetleg további tagolással.

`boxplot(függőváltozó~függetlenváltozó,objektum)`
Ha további csoportosítás:
`boxplot(függőváltozó~csoport*függetlenváltozó,objektum)`

Például:
`boxplot(ratings$Frequency~ratings$Class*ratings$Complex)`
Egyszerű és összetett állat- és növénynevek gyakorisága.

hasznos paraméterek:
`col=c("red","blue")`
`names=c("állat","növény")`

Feladatok

1. feladat: `ratings` fájlban található adatok alapján tetszőleges pontdiagramm készítése és mentése.
`ratings` fájlban szereplő adatok alapján tetszőleges pontdiagramm készítése és mentése.

2. feladat: `ratings` fájlban található adatok alapján tetszőleges boxplot készítése.

Célábra: cím, angol vagy magyar nyelvű tengelyfelirat.