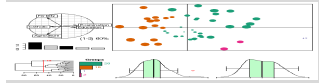


3. rész

Két változó kapcsolatának vizsgálata

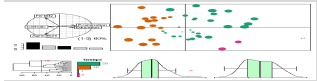
Minden összefügg mindennel!?

Komputerstatisztika kurzus

Barczy Mátyás és Ispány Márton 2010
Informatikai Kar
Debreceni Egyetem

A 3. rész témái

1. A kapcsolat típusai
2. Két diszkrét változó
3. Két folytonos változó
4. Szórásfelbontás
5. Osztályozási feladat
6. Többdimenziós skálázás



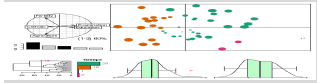
A kapcsolat természete

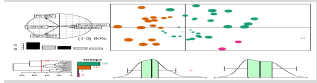
A statisztikai változók (adatbázisok attributumainak) korábban megismert egyenkénti jellemzése leíró statisztikákkal és grafikus eszközökkel többnyire csak egy kezdeti lépés. A statisztikai változók általában nem függetlenek egymástól, az egyes változók értékei befolyásolják más változók értékeit. A kapcsolat természete kétféle lehet:

- determinisztikus (függvényszerű): néhány változó egyértelműen (függvénnyel megadható módon) meghatározza más változó(k) értékét(eit),
- sztochasztikus (véletlenszerű): a fenti meghatározottság csak tendenciaszerű, bizonyos mértékű hiba erejéig érvényes.

A kapcsolat irányultsága is kétféle lehet:

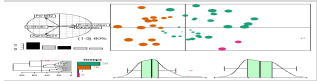
- aszimmetrikus: az egyik változó hat a másikra, pl. a viszony ok–okozati, illetve időben eltérő,
- szimmetrikus: a változók kölcsönösen hatnak egymásra, egyidejűek.





Példa (Determinisztikus kapcsolat)

- A hetente és havonta előállított termékek száma. (Az utóbbi az előbbiek összegzésével adódik). A kapcsolat aszimmetrikus abban az értelemben, hogy a havonta előállított termékek számából nem kapható meg a hetente előállított termékek száma, fordítva viszont igen.
- Az éves és a havi átlaghőmérsékletek.
- Egy termék vagy szolgáltatás ára és az áfa nagysága (20%-os áfa esetén az árat 0.2–del kell szorozni). A kapcsolat szimmetrikus abban az értelemben, hogy az áfa nagyságának ismeretében a termék ára és a kifizetett áfa kölcsönösen egyértelműen meghatározza egymást.



Példa (Sztochasztikus kapcsolat)

- Szemszín és a hajszín mint két diszkrét változó egyidejű kapcsolata (szimmetrikus kapcsolat).
- Testsúly és magasság mint két folytonos változó egyidejű kapcsolata (szimmetrikus kapcsolat).
- A termés nagysága és a különböző termelési módok.
- A gépkocsi sebessége és a fékút hossza (aszimmetrikus kapcsolat).
- Az idő és a számítástechnikai eszközök fejlettsége (tárolókapacitás, számolási sebesség stb.).

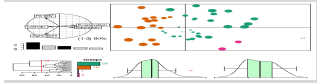
Két változó kapcsolatának leírása

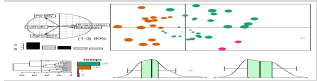
Ez a lehetséges legegyszerűbb kapcsolat, ennek ellenére már itt is eltérő módszerekkel találkozunk aszerint, hogy milyen változókat vizsgálunk, milyen skálán mérjük őket.

Jelöljük a két változót X–szel és Y–nal.

Ha ok–okozati kapcsolat áll fenn, akkor legyen X az ok és Y az okozat. Ekkor X–et magyarázó, Y–t függő változónak nevezzük.

A legalapvetőbb osztályozást akkor kapjuk, ha a változókat a típusuk alapján, mint diszkrét és folytonos, különböztetjük meg.





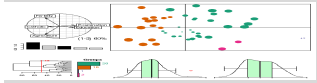
Az alábbi négy esetet vizsgáljuk:

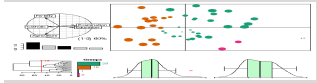
- Mindkét változó diszkrét:
  kontingencia táblák vizsgálata és rangkorreláció.
- Mindkét változó folytonos:
  korreláció– és regresszió analízis.
- X és Y között ok–okozati kapcsolat áll fenn, az X ok diszkrét, az Y okozat folytonos:
  szórásanalízis.
- X és Y között ok–okozati kapcsolat áll fenn, az X ok folytonos, az Y okozat diszkrét:
  osztályozási feladat.

Két diszkrét változó elemzése

Legyenek az X és Y diszkrét változók lehetséges értékei $x_1, \ldots, x_r$ és $y_1, \ldots, y_s$. Tekintsünk az (X, Y) párra vonatkozóan egy olyan $\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$ elemű mintát, ahol n_{ij} azon megfigyelések (rekordok) számát (gyakoriságát) jelöli, amelyeknél $X = x_i$ és $Y = y_j$, $i = 1, \ldots, r$, $j = 1, \ldots, s$.

Az összes elemző módszer ezután az így bevezetett gyakoriságokra épül. Ezek a mintában lévő teljes információt tartalmazzák, használatuk egy nagyon hatékony adattömörítő eszköz.





Elemzési eszköztár:

- statisztikai módszer: kontingencia táblák vizsgálata, rangkorrelációs együttható,
- grafikus eszközök: haladottabb (osztott, csoportosított, halmozott stb.) oszlop–, kör– és fánk–diagramok.

A legfontosabb kérdés a két változó függetlensége. Amennyiben ezt elutasítjuk, a függőségük erősségének mérése.

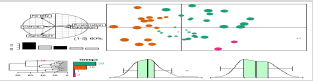
Kontingencia táblázat

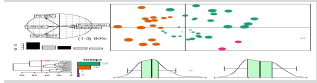
A kontingencia (kereszt) táblázat egy kétdimenziós gyakorisági tábla, amellyel két diszkrét változó együttes gyakoriságait tudjuk megjeleníteni.

X \ Y	y_1	y_2	$\cdots$	y_s	$\sum$
x_1	n_{11}	n_{12}	$\cdots$	n_{1s}	n_{1+}
x_2	n_{21}	n_{22}	$\cdots$	n_{2s}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	n_{r+}
$\sum$	n_{+1}	n_{+2}	$\cdots$	n_{+s}	n_{++}

A táblázatbeli sor és oszlopösszegek az alábbi módon vannak definiálva:

$$n_{i+} := \sum_{j=1}^{s} n_{ij}, \quad n_{+j} := \sum_{i=1}^{r} n_{ij}, \quad n_{++} := \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} =: n.$$





A megjelenítés hasonló a kétdimenziós diszkrét valószínűségeloszlás megadásához. Ekkor a táblázat valószínűségeket tartalmaz, az utolsó sor és oszlop pedig az ún. marginális (perem) eloszlásokat.

A kontingencia tábla celláiban (abszolút) gyakoriságok helyett ábrázolhatunk relatív gyakoriságokat is. A statisztikai szoftverek emellé még számos egyéb lehetőséget is nyújtanak, pl. oszlop vagy sor százalék, várt érték a függetlenség esetén stb.

A két változó függetlenségének vizsgálatát, illetve a függőségük erősségének mérését az alábbi két esetben tárgyaljuk: mindkét (diszkrét) változót nominális skálán, illetve mindkét változót ordinális skálán mérjük.

Nominális skálán mért diszkrét változók

Feltételezzük a továbbiakban, hogy $r \geq 2, s \geq 2$ és $n_{i+} \geq 1, n_{+j} \geq 1, i = 1, \ldots, s, j = 1, \ldots, r$.

Cramér–mutató:

$$C := \sqrt{\frac{\chi^2}{n_{++} \min\{r-1, s-1\}}},$$

ahol

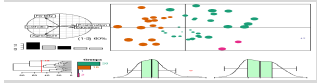
$$\chi^2 := \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}$$

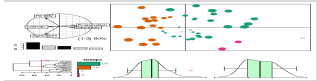
az ún. χ^2–statisztika.

Az $n_{i+}n_{+j}/n$ értékeket (gyakoriságokat) a két változó függetlenségének feltételezése melletti gyakoriságoknak is szokás nevezni, ugyanis

$$\frac{n_{i+}n_{+j}}{n} = n \frac{n_{i+}}{n} \frac{n_{+j}}{n}.$$

©Barczy Mátyás és Ispány Márton 2010





Állítás

A Cramér–mutatóra teljesül, hogy $C \in [0,1]$.

Bizonyítás. Először megmutatjuk, hogy

$$\chi^2 = n_{++} \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i+} n_{+j}} - 1 \right).$$

Valóban,

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i+} n_{+j}}{n} \right)^2}{n_{i+} n_{+j}} = \sum_{i=1}^{r} \sum_{j=1}^{s} \left(\frac{n_{ij}^2}{n_{i+} n_{+j}} - 2 \frac{n_{ij}}{n_{++}} + \frac{n_{i+} n_{+j}}{(n_{++})^2} \right)$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i+} n_{+j}} - \frac{2}{n_{++}} n_{++} + \frac{1}{(n_{++})^2} \sum_{i=1}^{r} n_{i+} \sum_{j=1}^{s} n_{+j}$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i+} n_{+j}} - 1.$$

Így elég azt ellenőriznünk, hogy

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i+}n_{+j}} \leq \min\{r,s\}.$$

Felhasználva, hogy

$$\frac{n_{ij}}{n_{i+}} \leq 1, \qquad i=1,\ldots,r,\ \ j=1,\ldots,s,$$

kapjuk, hogy

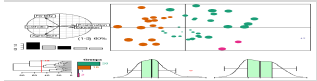
$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i+}n_{+j}} \leq \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{n_{+j}} = \sum_{j=1}^{s}\frac{1}{n_{+j}}\sum_{i=1}^{r}n_{ij}$$

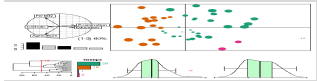
$$= \sum_{j=1}^{s}\frac{1}{n_{+j}}\cdot n_{+j} = s.$$

Hasonlóan belátható, hogy

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i+}n_{+j}} \leq r.$$

□





Megjegyzés.

- $C = 0$ akkor és csak akkor, ha

$$n_{ij} = \frac{n_{i+}n_{+j}}{n}, \qquad i = 1, \dots, r, \quad j = 1, \dots, s,$$

azaz, ha a megfigyelt és várt gyakoriságok megegyeznek. $C = 0$: „függetlenség”.

- Ha $s \leq r$, úgy $C = 1$ akkor és csak akkor, ha minden $i = 1, \dots, r$ esetén

$$n_{ij} \in \{0, n_{i+}\}, \qquad j = 1, \dots, s,$$

azaz minden sorban pontosan egy db n_{ij} nem nulla. $C = 1$: „függőség”.

- Példa arra, hogy $C = 0$. Tekintsük az alábbi kontingencia táblázatot:

X \ Y	2	4	$\sum$
1	2	2	4
2	2	2	4
$\sum$	4	4	8

.

Ekkor $\chi^2 = 0$ és $C = 0$.

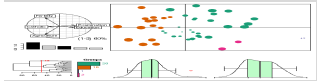
- Példa arra, hogy $C = 1$. Tekintsük az alábbi kontingencia táblázatot:

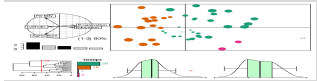
X \ Y	2	4	$\sum$
1	0	3	3
2	1	0	1
3	0	2	2
$\sum$	1	5	6

.

Ekkor $\chi^2 = 6$ és $C = 1$.

©Barczy Mátyás és Ispány Márton 2010





Csuprov–mutató:

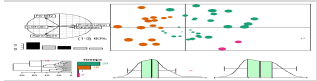
$$T := \sqrt{\frac{\chi^2}{n_{++}\sqrt{(r-1)(s-1)}}}.$$

Nyilván, $T \leq C$ és így $T \in [0,1]$.

Értelmezés ugyanaz, mint a Cramér–mutatónál.

Megjegyzés

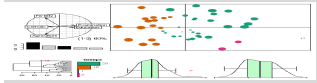
A későbbiekben elemezni fogjuk, hogy a mutatók (pontosabban a χ^2–statisztika) milyen nagy értékei utalnak függőségre, lásd két teljes eseményrendszer függetlenségének vizsgálatára vonatkozó χ^2–próba.



Példa

A szem és a haj színének a kapcsolatát vizsgáltuk 400 embernél. Eredményül az alábbi kontingencia táblát kaptuk:

	Hajszín				
Szemszín	barna	fekete	szőke	vörös	$\sum$
kék	4	4	40	2	50
sötét	120	80	75	25	300
zöld	10	10	15	15	50
$\sum$	134	94	130	42	400



Megoldás.

$r = 3$, $s = 4$ és a χ^2–statisztika értéke 84.32.

A Cramér–mutató értéke:

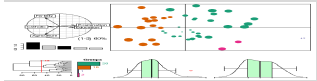
$$C = \sqrt{\frac{84.32}{400 \min\{3-1, 4-1\}}} \approx 0.3246.$$

A Csuprov–mutató értéke:

$$T := \sqrt{\frac{84.32}{400\sqrt{(3-1)(4-1)}}} \approx 0.2933.$$

A Cramér–, ill. Csuprov–mutató értéke alapján (mivel 0–hoz vannak közelebb, mint 1–hez) gyenge függőségre következtethetünk.

Azonban a későbbi vizsgálatok során kiderül majd, hogy a jelen feladatban a Cramér–mutató 0.3246 értéke alapján erős függőségre kell következtetnünk.



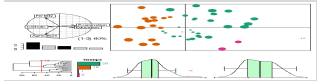
A függőség mértékének a megítélésében a $P(\chi^2_6 > 84.32)$ valószínűség nagysága az irányadó, ugyanis belátható, hogy a két változó függetlensége esetén a χ^2–statisztika aszimptotikusan $\chi^2_{(r-1)(s-1)}$ eloszlású, ahol χ^2_n az n–szabadsági fokú χ^2–eloszlást jelöli.

Esetünkben $(r-1)(s-1) = 6$ és $P(\chi^2_6 > 84.32) \approx 0$, így tényleg erős függőségre következtethetünk.

Megjegyezzük továbbá, hogy érvényes ugyan a

$$\frac{\chi^2_n - n}{\sqrt{2n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \qquad \text{ha } n \to \infty,$$

eloszlásbeli konvergencia, ez azonban elég „lassú”.

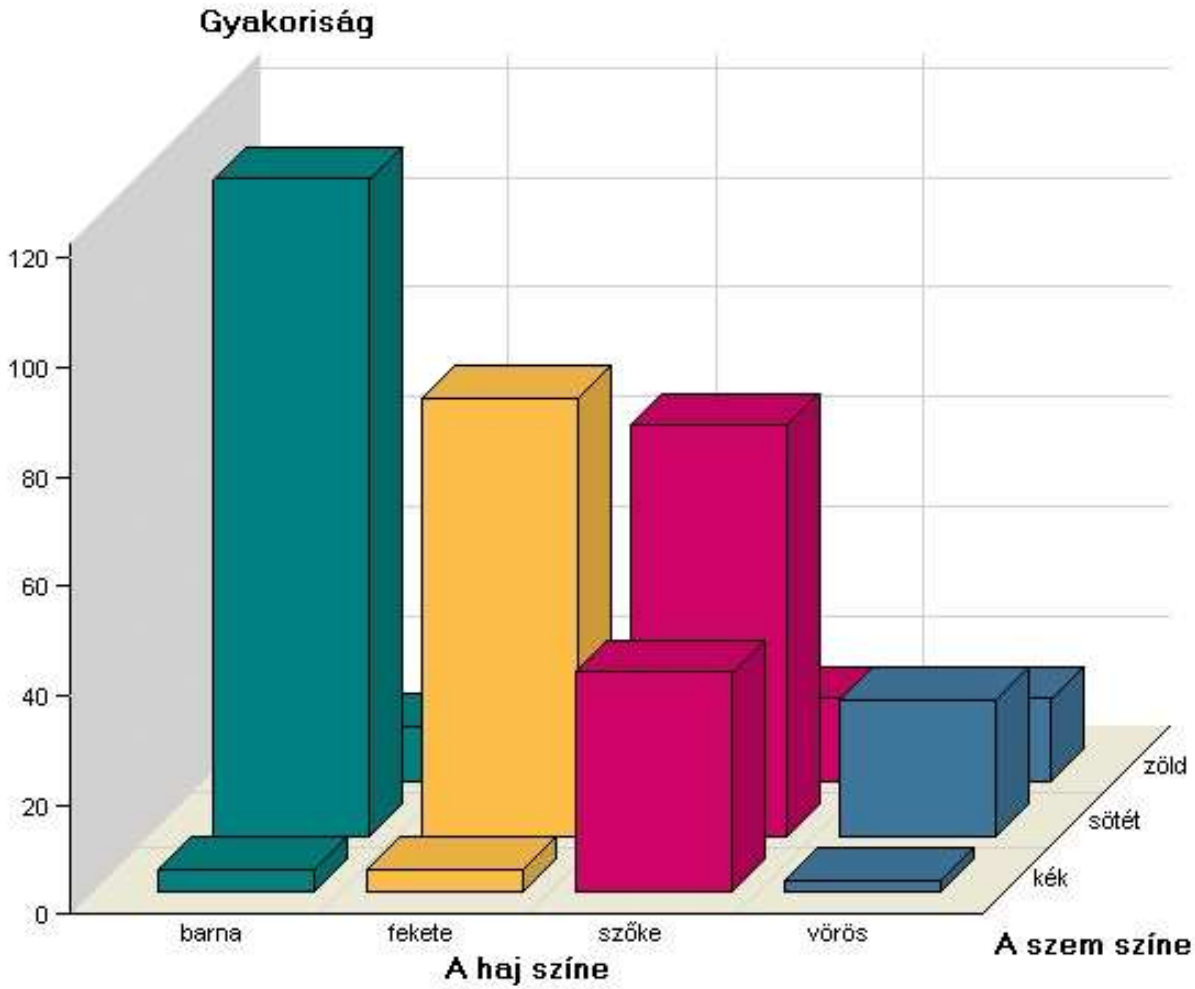


Gondoljunk ugyanis a Berry–Esseen–tételre, ahol is a konvergenciasebesség nagysága az abszolút ferdeségtől (harmadik abszolút centrált momentum osztva a szórás köbével) függ, ami χ^2_1–eloszlás esetén relatíve nagy.

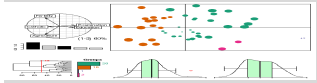
Mivel a gyakorló feladatokban a $\chi^2_{(r-1)(s-1)}$ határeloszlás szabadsági foka, azaz $(r-1)(s-1)$, általában kicsi, így a határeloszlásnak $\mathcal{N}((r-1)(s-1), 2(r-1)(s-1))$ eloszlással való közelítését nem ajánljuk. □

A következő lapokon a példabeli kontingencia táblát ábrázoltuk különböző módokon.

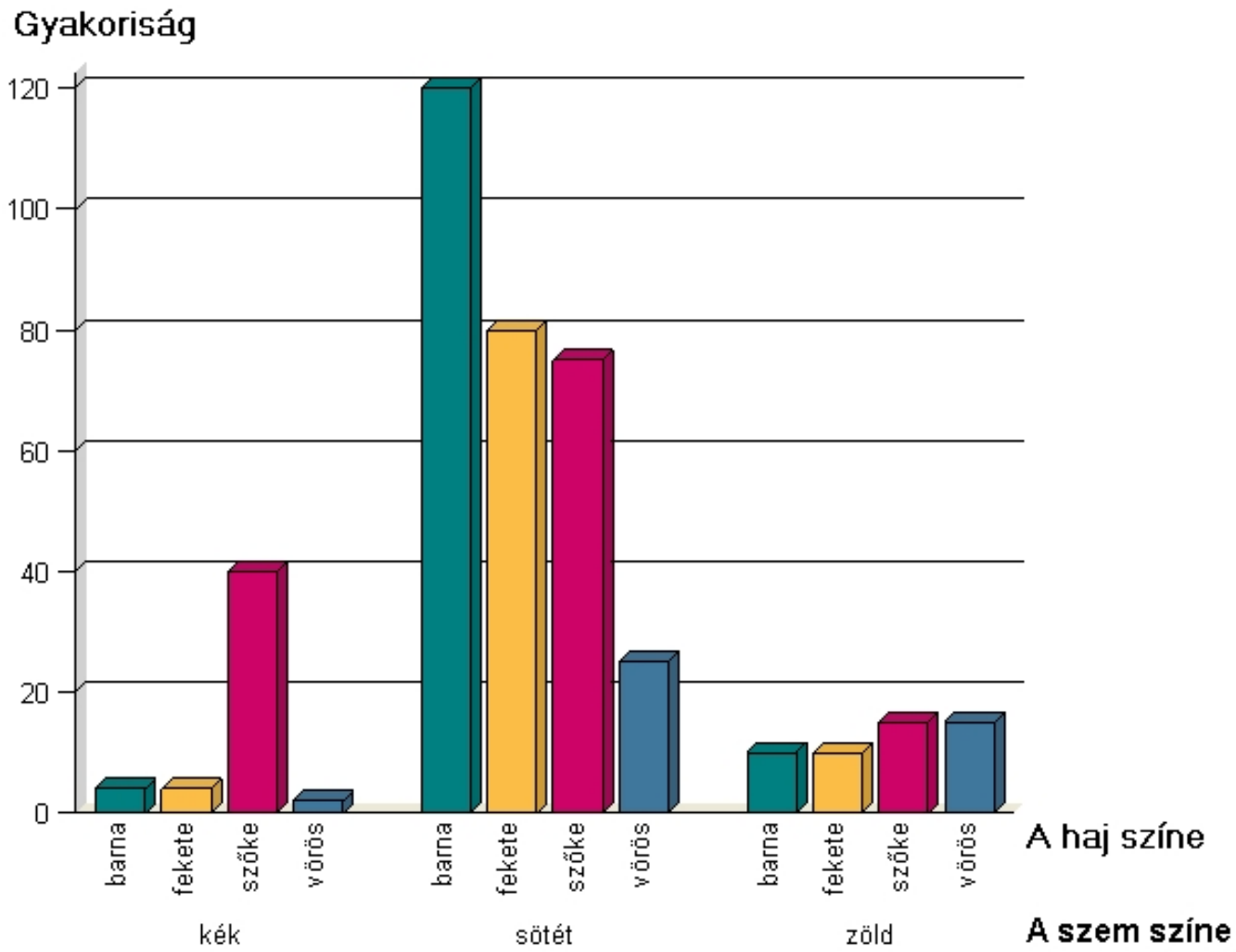
Térbeli oszlopdiagram



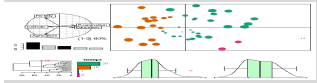
Előállította a SAS rendszer



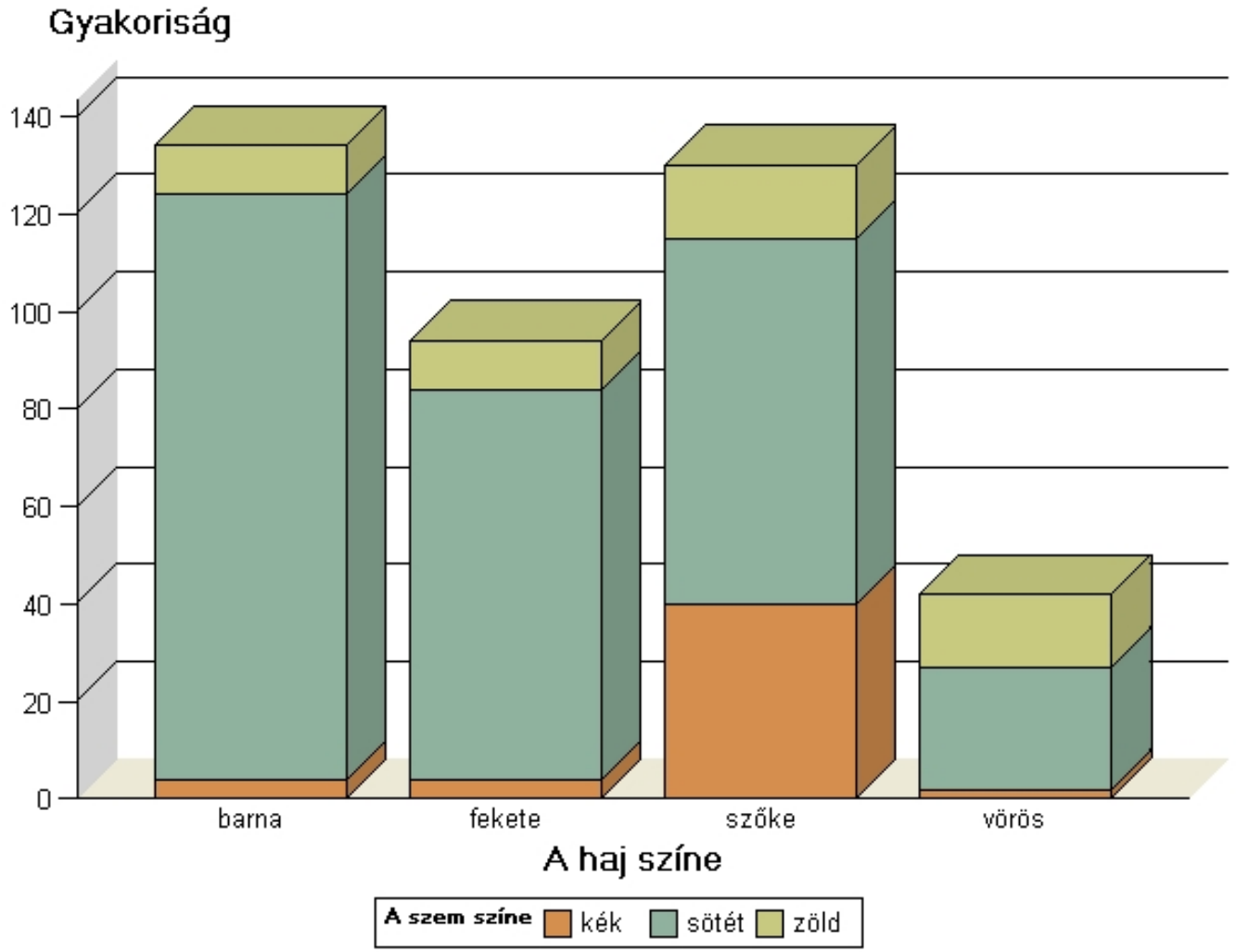
Csoportosított oszlopdiagram



Előállította a SAS rendszer

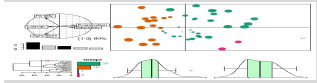


Halmozott oszlopdiagram



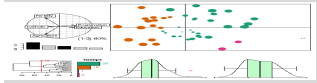
Előállította a SAS rendszer

©Barczy Mátyás és Ispány Márton 2010

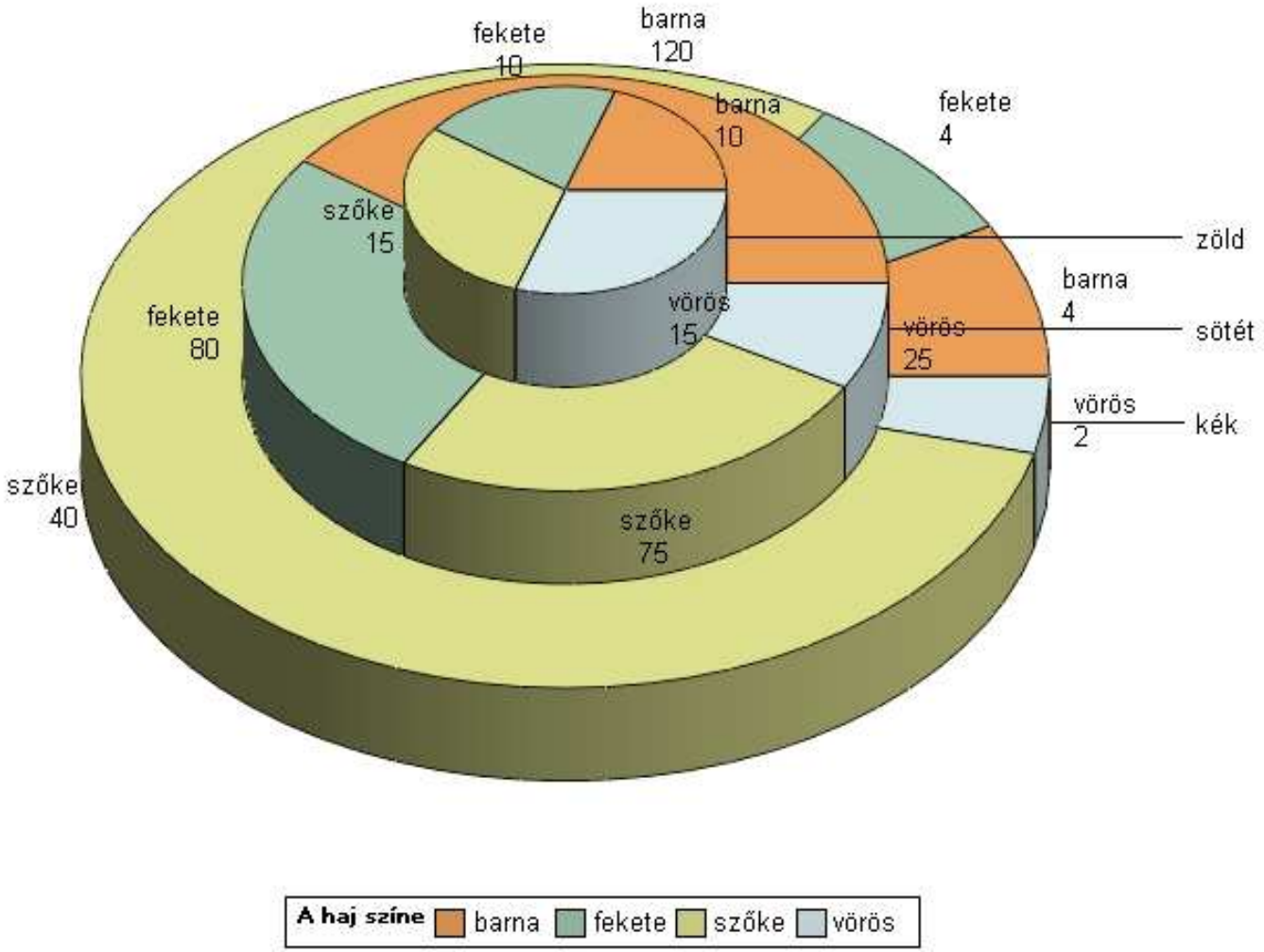


Csoportosított kördiagram

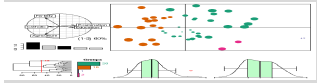
Előállította a SAS rendszer



Halmozott kördiagram

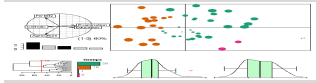


Előállította a SAS rendszer



Fánkdiagram

Előállította a SAS rendszer



Ordinális skálán mért diszkrét változók

Legyen a két ordinális skálán mért, diszkrét változóra megfigyelt minta a következő:

$$X : x_1, x_2, \ldots, x_n$$
$$Y : y_1, y_2, \ldots, y_n.$$

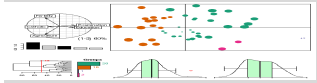
Megjegyezzük, hogy valójában mintapárokat, úm. (x_i, y_i), figyelünk meg, amelynek elemei összetartoznak.

Rang:= a rendezett mintában hányadik az illető mintaelem.

Jelöljük az X, ill. Y változó szerinti rangokat R_X, illetve R_Y módon.

Ha egy mintaelem többször is előfordul, akkor ezekhez azon rangszámok súlyozatlan számtani átlagát rendeljük rangszámként, melyet akkor kapnánk, ha az adott mintaelemek páronként különbözőek lennének. Ezek az ún. kapcsolt rangok.

©Barczy Mátyás és Ispány Márton 2010



Spearman–féle rangkorrelációs együttható:

az R_X és R_Y rangszámok tapasztalati (lineáris) korrelációs együtthatója:

$$\varrho_S(X,Y) := \varrho(R_X, R_Y),$$

ahol

$$\varrho(X,Y) := \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2 \sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

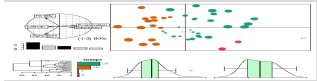
Azaz

$$\varrho_S(X,Y) = \frac{\sum_{i=1}^{n}(R_{x_i}-\overline{R_X})(R_{y_i}-\overline{R_Y})}{\sqrt{\sum_{i=1}^{n}(R_{x_i}-\overline{R_X})^2 \sum_{i=1}^{n}(R_{y_i}-\overline{R_Y})^2}},$$

ahol

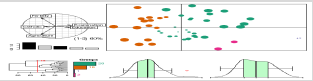
$$\overline{R_X} := \frac{1}{n}\sum_{i=1}^{n} R_{x_i}, \qquad \overline{R_Y} := \frac{1}{n}\sum_{i=1}^{n} R_{y_i}.$$

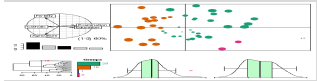
Megj.: A tapasztalati (lineáris) korrelációs együtthatóval két folytonos változó kapcsolatának elemzésekor majd részletesen foglalkozunk.



A rangkorrelációs együttható tulajdonságai:

- $|\varrho_S(X,Y)| \leq 1$,
- Ha $\varrho_S(X,Y) = 1$, akkor az R_X és R_Y rangszámsorozat egybeesik.
- Ha $\varrho_S(X,Y) = -1$, akkor az R_X és R_Y rangszámsorozat egymás „fordítottjai” abban az értelemben, hogy $R_{y_i} = -R_{x_i} + n + 1$, $i = 1, \ldots, n$.
- Ha $\varrho_S(X,Y) > 0$, akkor pozitív irányú függésről beszélünk: az X-re vonatkozó mintában a nagyobb rang együttjár az Y-re vonatkozó mintában a nagyobb ranggal (illetve fordítva is).
- Ha $\varrho_S(X,Y) < 0$, akkor negatív irányú függésről beszélünk: az X-re vonatkozó mintában a nagyobb rang együttjár az Y-re vonatkozó mintában a kisebb ranggal (illetve fordítva is).
- Ha $\varrho_S(X,Y) = 0$, akkor a két rangsor között nincs kapcsolat.





Megjegyzés

Ha nincsenek kapcsolt rangok, akkor

$$\varrho_S(X,Y) = 1 - \frac{6\sum_{i=1}^{n}(R_{x_i} - R_{y_i})^2}{n(n^2-1)}.$$

A gyakorlatban akkor is használatos a fenti formula, ha vannak kapcsolt rangok, de nem sok.

Az SPSS algoritmusa a rangkorrelációs együttható számolására

$$\varrho_S(X,Y) = \frac{T_X + T_Y - \sum_{i=1}^{n}(R_{x_i} - R_{y_i})^2}{2\sqrt{T_X T_Y}},$$

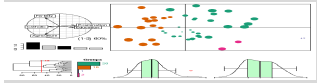
ahol

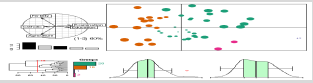
$$T_X := \frac{n^3 - n - ST_X}{12}, \qquad T_Y := \frac{n^3 - n - ST_Y}{12},$$

$$ST_X := \sum_{\{t>1\,:\,\text{t db mintaelem egybeesik az } x_1,\ldots,x_n \text{ mintában}\}} (t^3 - t),$$

$$ST_Y := \sum_{\{t>1\,:\,\text{t db mintaelem egybeesik az } y_1,\ldots,y_n \text{ mintában}\}} (t^3 - t).$$

Ha $T_X = 0$ vagy $T_Y = 0$, akkor $\varrho_S(X,Y)$ nem kerül kiszámolásra.
Valóban ellenőrizhető, hogy a fenti algoritmussal számolt rangkorrelációs együttható megegyezik az általunk bevezetettel.





Két folytonos változó elemzése

Legyen a két folytonos változóra megfigyelt minta a következő:

$$X : x_1, x_2, \ldots, x_n$$
$$Y : y_1, y_2, \ldots, y_n$$

Újra felhívjuk a figyelmet, hogy valójában mintapárokat, úm. (x_i, y_i), figyelünk meg, amelynek elemei összetartoznak. Így bármilyen művelet (pl. rendezés) a mintán csak a párokon hajtható végre, változóként pedig nem.

Statisztikai eszközök:

- szimmetrikus eset: korreláció analízis,
- ok–okozati eset: regresszió analízis.

Grafikus eszköz: pontdiagram, melyben a Descartes koordinátarendszerben ábrázoljuk az (X, Y) párra kapott megfigyeléseket mint pontokat. A koordinátatengelyek kijelölése gyakran feltételez ok–okozati viszonyt a két változó között.

Pearson–féle korrelációs együttható

Két valószínűségi változó közötti függőséget a kovariancia és korrelációs együttható mennyiségekkel mérhetjük.

A ξ és η valószínűségi változók (elméleti) kovarianciája:

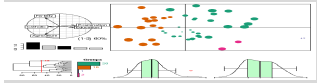
$$\mathrm{Cov}(\xi,\eta) := \mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta),$$

és korrelációs együtthatója:

$$\mathrm{Corr}(\xi,\eta) := \frac{\mathrm{Cov}(\xi,\eta)}{\sqrt{\mathbb{D}^2\xi\,\mathbb{D}^2\xi}}.$$

A korrelációs együtthatót akkor értelmezzük, ha
$0 < \mathbb{D}^2\xi < \infty$ és $0 < \mathbb{D}^2\eta < \infty$.

Ezek tapasztalati megfelelőit vezetjük be az alábbiakban.



Definíció (Tapasztalati kovariancia és korreláció)

Az X és Y közötti tapasztalati kovariancia:

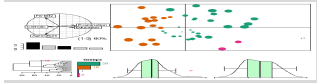
$$c(X,Y) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

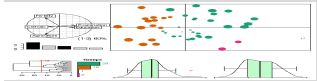
Pearson-féle tapasztalati (lineáris) korrelációs együttható:

$$\varrho(X,Y) := \frac{c(X,Y)}{s(X)s(Y)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

feltéve, hogy $s(X) \neq 0$ és $s(Y) \neq 0$, ahol $s(X)$, illetve $s(Y)$ az $x_1, \ldots, x_n$, ill. $y_1, \ldots, y_n$ minta korrigálatlan tapasztalati szórását jelöli.

Megj.: Korábban az $x_1, \ldots, x_n$ minta korrigálatlan tapasztalati szórását s_n módon jelöltük. A fentiekben a minta elemszámát nem tüntettük fel a jelölésben, azt viszont igen, hogy melyik mintára vonatkozik.





Megjegyzés

- $$c(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y} =: \overline{xy} - \overline{x}\cdot\overline{y}.$$
- X és Y között nem feltétlen van ok–okozati kapcsolat.
- Bevezetve az alábbi jelöléseket:

$$\widetilde{x}_i := \frac{x_i - \overline{x}}{\sqrt{n}s(X)}, \qquad \widetilde{y}_i := \frac{y_i - \overline{y}}{\sqrt{n}s(Y)}, \quad i = 1,\dots,n,$$

az $\widetilde{X} := (\widetilde{x}_1,\dots,\widetilde{x}_n)$ és $\widetilde{Y} := (\widetilde{y}_1,\dots,\widetilde{y}_n)$ mintákra teljesül, hogy

$$\varrho(X,Y) = \sum_{i=1}^{n} \widetilde{x}_i \widetilde{y}_i = \langle \widetilde{X}, \widetilde{Y} \rangle,$$

ahol $\langle\cdot,\cdot\rangle$ az euklideszi belső szorzatot jelöli.

- $\varrho(X,Y)^2$: determinációs együttható.

A korrelációs együttható tulajdonságai

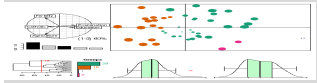
A tapasztalati kovariancia a két változó közös skáláján méri a kapcsolat erősségét, a tapasztalati korrelációs együttható ezzel szemben már normalizált skálán mér, így összehasonlításra jobban használható.

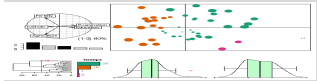
Ha $\varrho(X, Y) > 0$, akkor pozitív irányú függésről beszélünk: az egyik változó értékének növelése a másik változó értékének a növekedését eredményezi.

Ha $\varrho(X, Y) < 0$, akkor negatív irányú függésről beszélünk: az egyik változó értékének növelése a másik változó értékének a csökkenését eredményezi.

Ha $\varrho(X, Y) = 0$, akkor korrelálatlanságról beszélünk. (Nem tévesztendő össze a függetlenséggel, ami egy erősebb dolog.)

©Barczy Mátyás és Ispány Márton 2010

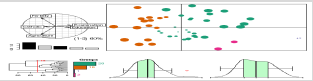




Tétel

Legyen $X = (x_1, \ldots, x_n)$ *és* $Y = (y_1, \ldots, y_n)$ *olyan, hogy* $s(X) \neq 0$ *és* $s(Y) \neq 0$. *Ekkor* $|\varrho(X, Y)| \leq 1$ *és egyenlőség pontosan akkor áll fenn, ha léteznek olyan* $a, b \in \mathbb{R}$ *valós számok, hogy* $Y = aX + b$, *azaz* $y_i = ax_i + b$ *minden* $i = 1, \ldots, n$*–re. Utóbbi esetben lineáris kapcsolatról beszélünk.*

Továbbá, $a > 0$, *illetve* $a < 0$ *aszerint, hogy* $\varrho(X, Y) = 1$, *illetve* $\varrho(X, Y) = -1$.



Bizonyítás. A Cauchy–Schwarz egyenlőtlenség alapján:

$$\left|\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right| \le \left(\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2\right)^{1/2},$$

melyből átrendezéssel adódik az állítás.
Egyenlőség pontosan akkor teljesül, ha

$$(x_1-\overline{x},\dots,x_n-\overline{x}) \quad \text{és} \quad (y_1-\overline{x},\dots,y_n-\overline{x})$$

lineárisan függőek. Mivel $s(X)\neq 0$ és $s(Y)\neq 0$, kapjuk, hogy pontosan akkor teljesül egyenlőség, ha létezik olyan $a\in\mathbb{R}$, hogy $y_i-\overline{y}=a(x_i-\overline{x})$ minden $i=1,\dots,n$ esetén. Innen közvetlenül adódik, hogy $y_i=ax_i+b$, ahol $b:=\overline{y}-a\cdot\overline{x}$.

Ha $Y = aX + b$, ahol $a > 0$, akkor

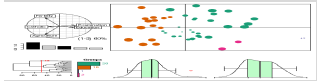
$$\varrho(X, Y) = \frac{c(X, aX + B)}{s(X)s(aX + b)} = \frac{a \cdot c(X, X)}{a \cdot s(X)s(X)} = 1.$$

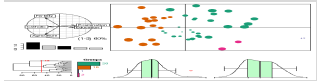
Ha $Y = aX + b$, ahol $a < 0$, akkor

$$\varrho(X, Y) = \frac{c(X, aX + B)}{s(X)s(aX + b)} = \frac{a \cdot c(X, X)}{|a| \cdot s(X)s(X)} = -1.$$

□

Összefoglalva: a korrelációs együttható a $[-1, 1]$ intervallumon (skálán) méri két folytonos változó kapcsolatának az erősségét. A skála két végpontja esetén a kapcsolat lineáris ($Y = aX + b$), a -1 végpont esetén $a < 0$, a $+1$ végpont esetén pedig $a > 0$.





Példa (olyan determinisztikus kapcsolatra, amelynél a változók korrelálatlanok)

Tekintsük az alábbi mintát:

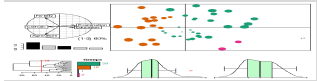
$$(x_1, y_1) = (1, 1), \qquad (x_2, y_2) = (-1, 1),$$
$$(x_3, y_3) = (2, 8), \qquad (x_4, y_4) = (-2, 8).$$

Ekkor $\overline{x} = 0$,

$$\overline{y} = \frac{1 + 1 + 8 + 8}{4} = 4.5,$$
$$\overline{xy} = \frac{1 - 1 + 16 - 16}{4} = 0,$$

így $c(X, Y) = 0$ és $\varrho(X, Y) = 0$.

X és Y között determinisztikus kapcsolat van: $y_i = x_i^3$, $i = 1, 2, 3, 4$.



Példa (A tapasztalati korrelációs együttható számolása)

A minta:

X:	1,	2,	-1,	2,	1
Y:	2,	4,	0,	8,	1

Mivel $\overline{x} = 1$ és $\overline{y} = 3$ a centralizált minta:

X$-\overline{x}$:	0,	1,	-2,	1,	0
Y$-\overline{y}$:	-1,	1,	-3,	5,	-2

$$c(X,Y) = \frac{0\cdot(-1)+1\cdot 1+(-2)\cdot(-3)+1\cdot 5+0\cdot(-2)}{5} = 2.4,$$

$$s^2(X) = \frac{1}{5}(0^2+1^2+(-2)^2+1^2+0^2) = 1.2,$$

$$s^2(Y) = \frac{1}{5}((-1)^2+1^2+(-3)^2+5^2+(-2)^2) = 8.$$

Ezért: $\varrho(X,Y) = 2.4/\sqrt{1.2\cdot 8} = 0.7746.$

Lineáris regresszió

Láttuk, hogy maximális (azaz 1) abszolút értékű korrelációs együttható lineáris kapcsolatot jelent a változók között. Hogyan lehet ennek a kapcsolatnak az együtthatóit meghatározni a minta alapján?
Általánosabban, az Y függő változót szeretnénk az X magyarázó változó lineáris függvényével közelíteni:

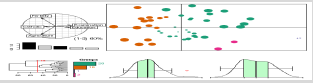
$$Y \approx aX + b.$$

Milyen a és b valós együtthatókat válasszunk az (X, Y)-ra vonatkozó n elemű minta ismeretében?

Először az elemi hibákat (veszteségeket) definiáljuk az alábbi módon

$$e_i := y_i - (ax_i + b), \qquad i = 1, \ldots, n.$$

Ezután ezen (elemi) hibákból egy összesített hibát (rizikót) állítunk elő. Végül az így kapott összesített hibát, mint célfüggvényt minimalizáljuk az a és b együtthatók függvényében. Ez egy szélsőértékszámítási (optimalizációs) feladat.



A legkisebb négyzetek módszere

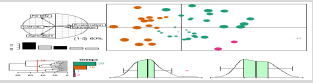
Az együtthatók meghatározására használt legelterjedtebb elv a legkisebb négyzetek módszere. Ekkor az összesített hiba az elemi hibák négyzeteinek összege:

$$E(a,b) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 .$$

Szélsőértékszámítási feladat:

$$\min_{(a,b)\in\mathbb{R}^2} E(a,b).$$

Az E függény előnye, hogy a szélsőértékszámítási feladat megoldásánál támaszkodhatunk a differenciálszámítás eszköztárára, hiszen az $E : \mathbb{R}^2 \to \mathbb{R}$ függvény mindkét változója szerint akárhányszor differenciálható.



Ismert, hogy az $E : \mathbb{R}^2 \to \mathbb{R}$ függvénynek $(\hat{a}, \hat{b})$ lokális minimumhelye, ha

(i) $(\hat{a}, \hat{b})$ stacionárius pont, azaz az elsőrendű parciális deriváltakból álló ún. gradiens vektorra

$$\frac{\partial E}{\partial a}(\hat{a}, \hat{b}) = 0, \qquad \frac{\partial E}{\partial b}(\hat{a}, \hat{b}) = 0;$$

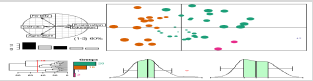
(ii) a másodrendű parciális deriváltakból álló ún. Hesse mátrix pozitív definit az $(\hat{a}, \hat{b})$ helyen, azaz

$$\begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \frac{\partial^2 E}{\partial a^2}(\hat{a}, \hat{b}) & \frac{\partial^2 E}{\partial a \partial b}(\hat{a}, \hat{b}) \\ \frac{\partial^2 E}{\partial a \partial b}(\hat{a}, \hat{b}) & \frac{\partial^2 E}{\partial b^2}(\hat{a}, \hat{b}) \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} > 0$$

minden $(u, v) \neq (0, 0)$ esetén, vagy ekvivalens módon

$$u^2 \frac{\partial^2 E}{\partial a^2}(\hat{a}, \hat{b}) + 2uv \frac{\partial^2 E}{\partial a \partial b}(\hat{a}, \hat{b}) + v^2 \frac{\partial^2 E}{\partial b^2}(\hat{a}, \hat{b}) > 0$$

minden $u, v \in \mathbb{R}$, $u^2 + v^2 > 0$ esetén.



A regressziós együtthatók meghatározása I.

Az E gradiens vektorára kapjuk, hogy

$$\frac{\partial E}{\partial b}(a,b) = 2\sum_{i=1}^{n}(y_i - (ax_i + b))(-1) = 0,$$

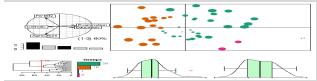
$$\frac{\partial E}{\partial a}(a,b) = 2\sum_{i=1}^{n}(y_i - (ax_i + b))(-x_i) = 0.$$

Ebből átrendezéssel kapjuk az alábbi ún. normál egyenleteket:

$$a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i,$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i.$$

©Barczy Mátyás és Ispány Márton 2010



A regressziós együtthatók meghatározása II.

A normálegyenletek az

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2,\ \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i,$$

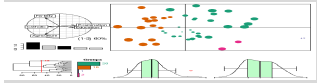
jelölések bevezetésével az alábbi egyszerűbb alakot öltik:

$$a \cdot \overline{x} + b = \overline{y},$$
$$a \cdot \overline{x^2} + b \cdot \overline{x} = \overline{xy},$$

melyeket írhatunk mátrixos alakban is:

$$\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix}.$$

©Barczy Mátyás és Ispány Márton 2010



A regressziós együtthatók meghatározása III.

Az első egyenletből kapjuk, hogy $b = \overline{y} - a \cdot \overline{x}$. Ezt a második egyenletbe behelyettesítve adódik a megoldás:

$$\widehat{a} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}, \qquad \widehat{b} = \overline{y} - \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}\overline{x},$$

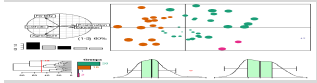
amennyiben $\overline{x^2} \neq \overline{x}^2$. Ez utóbbi akkor és csak akkor teljesül, ha az X–re vett minta nem konstans, ugyanis

$$\overline{x^2} = \overline{x}^2 \qquad \Longleftrightarrow \qquad n\sum_{i=1}^{n} x_i^2 = \left(\sum_{i=1}^{n} x_i\right)^2,$$

illetve a Cauchy–Schwarz–egyenlőtlenség szerint

$$\left(\sum_{i=1}^{n} x_i\right)^2 \leq n\sum_{i=1}^{n} x_i^2$$

és egyenlőség akkor és csak akkor áll fenn, ha $\exists\, d \in \mathbb{R}$, hogy $x_i = d, i = 1, \ldots, n$.



A regressziós együtthatók meghatározása IV.

Mivel a másodrendű parciális deriváltak:

$$\frac{\partial^2 E}{\partial a^2}(a,b) = 2\sum_{i=1}^{n} x_i^2, \qquad \frac{\partial^2 E}{\partial b^2}(a,b) = 2n,$$

$$\frac{\partial^2 E}{\partial a \partial b}(a,b) = 2\sum_{i=1}^{n} x_i,$$

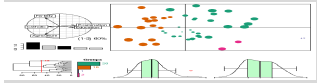
a Hesse mátrix tetszőleges $(a,b) \in \mathbb{R}^2$ helyen

$$\begin{pmatrix} 2\sum_{i=1}^{n} x_i^2 & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2n \end{pmatrix}$$

Ez pozitív definit, hiszen

$$u^2 2\sum_{i=1}^{n} x_i^2 + 2uv2\sum_{i=1}^{n} x_i + v^2 2n = 2\sum_{i=1}^{n} (ux_i + v)^2 > 0$$

ha $u, v \in \mathbb{R}$, $u^2 + v^2 > 0$ és az X–re vett minta nem konstans.



A regressziós együtthatók meghatározása V.

Egyetlen dolgot kell még ellenőriznünk: az E függvény egyetlen lokális minimumhelye egyben globális minimumhely is.

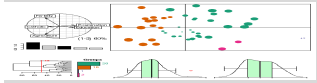
Ehhez elég belátni, hogy az E függvény szigorúan konvex. Ugyanis, ha egy szigorúan konvex függvénynek van lokális minimuma az szükségképpen globális minimum is.

Az E szigorú konvexségéhez elég ellenőriznünk, hogy a Hesse-mátrixának sarokfőminorai pozitívak, azaz azt, hogy $2\sum_{i=1}^{n} x_i^2 > 0$ és

$$4\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) = 4n^2(\overline{x^2} - (\overline{x})^2) > 0.$$

Ezek teljesülnek, hiszen az X–re vett minta nem konstans.

Ezzel beláttuk, hogy a normál egyenletek egyértelmű megoldása valóban globális minimumhelyet határoz meg.



A regressziós egyenes

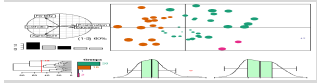
Felhasználva, hogy $c(X,Y) = \overline{xy} - \overline{x} \cdot \overline{y}$ és azt, hogy a Steiner–formula alapján $s^2(X) = \overline{x^2} - \overline{x}^2$, a regressziós együtthatókat a következő alakban is felírhatjuk:

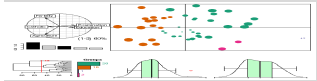
$$\widehat{a} = \varrho(X,Y)\frac{s(Y)}{s(X)}, \qquad \widehat{b} = \overline{y} - \varrho(X,Y)\frac{s(Y)}{s(X)}\overline{x}.$$

A lineáris regresszió feladatát megoldó egyenes egyen–letét, az ún. regressziós egyenest az alábbi alakokban írhatjuk fel:

$$Y = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}X + \overline{y} - \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}\overline{x},$$

$$Y = \varrho(X,Y)\frac{s(Y)}{s(X)}X + \overline{y} - \varrho(X,Y)\frac{s(Y)}{s(X)}\overline{x}.$$





Feltételezve, hogy az Y–ra vett minta sem konstans, a másodikat átrendezve kapjuk, hogy

$$\frac{Y-\overline{y}}{s(Y)} = \varrho(X,Y)\frac{X-\overline{x}}{s(X)},$$

azaz a két minta standardizálása után egy origón átmenő egyenest kapunk, melynek meredeksége a tapasztalati korrelációs együttható.

Nemlineáris regresszió

Csak azzal a speciális esettel foglalkozunk, amikor az Y függő változót szeretnénk az X magyarázó változó hatvány függvényével közelíteni:

$$Y \approx b \cdot a^X,$$

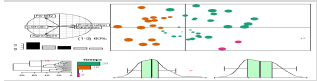
ahol feltételezzük, hogy a valódi a és b paraméterek pozitívak.

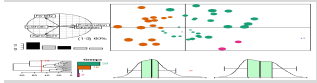
Milyen a és b valós együtthatókat válasszunk az (X, Y)-ra vonatkozó n elemű minta ismeretében?

A legkisebb négyzetek elve alapján olyan a és b értékeket választunk, melyekre az alábbi összesített hiba minimális:

$$\sum_{i=1}^{n} (y_i - ba^{x_i})^2.$$

Ezen nemlineáris szélsőértékszámítási feladat megoldása helyett azonban sokszor az alábbi lineáris (így egyszerűbb) feladatot oldjuk meg.



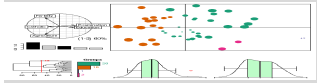


Az eredeti nemlineáris regressziós feladatot visszavezetjük egy lineáris regressziós feladat megoldására, linearizálunk:

$$\ln(Y) \approx \ln(a)X + \ln(b).$$

Az $(x_i, \ln(y_i)), i = 1, \ldots, n$ minta ismeretében az $\ln(a)$ és $\ln(b)$ paraméterek legkisebb négyzetes becslése teljesíti az alábbi normálegyenletet:

$$\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} \ln(b) \\ \ln(a) \end{pmatrix} = \begin{pmatrix} \overline{\ln(y)} \\ \overline{x \ln(y)} \end{pmatrix}.$$



A korábbiak alapján, ha az X-re vett minta nem konstans, akkor az $\ln(a)$ és $\ln(b)$ paraméterek legkisebb négyzetes becslése egyértelmű:

$$\widehat{\ln(a)} = \frac{\overline{x\ln(y)} - \overline{x}\cdot\overline{\ln(y)}}{\overline{x^2} - \overline{x}^2},$$

$$\widehat{\ln(b)} = \overline{\ln(y)} - \frac{\overline{x\ln(y)} - \overline{x}\cdot\overline{\ln(y)}}{\overline{x^2} - \overline{x}^2}\overline{x}.$$

Így az a és b paraméterek becslése:

$$\widehat{a} = \mathrm{e}^{\widehat{\ln(a)}}, \qquad \widehat{b} = \mathrm{e}^{\widehat{\ln(b)}}.$$

A nemlineáris regressziós feladat és linearizáltjának a kapcsolata

Legyenek az (X, Y) párra vonatkozó megfigyeléseink (x_i, y_i), $i = 1, \ldots, n$, ahol $y_i > 0$, $i = 1, \ldots, n$, és az X-re vett minta nem konstans.

Vizsgáljuk meg az alábbi két feladat közötti kapcsolatot.

I. Legyen $E_1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$E_1(u, v) := \sum_{i=1}^{n} (\ln(y_i) - (ux_i + v))^2, \qquad u, v \in \mathbb{R}.$$

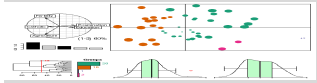
Minimalizáljuk E_1-et és jelölje $(\widehat{u}, \widehat{v})$ a minimumhelyet.

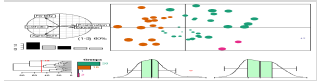
II. Legyen $E_2 : (0, \infty) \times (0, \infty) \to \mathbb{R}$,

$$E_2(a, b) := \sum_{i=1}^{n} (y_i - ba^{x_i})^2, \qquad a, b > 0.$$

Minimalizáljuk E_2-t és jelölje $(\widehat{a}, \widehat{b})$ a minimumhelyet, melyről feltételezzük, hogy egyértelmű.

Igaz-e, hogy $(\widehat{a}, \widehat{b}) = (\mathrm{e}^{\widehat{u}}, \mathrm{e}^{\widehat{v}})$?





Az I. feladat megoldása:

$$\widehat{u} = \frac{\overline{x\ln(y)} - \overline{x}\cdot\overline{\ln(y)}}{\overline{x^2} - \overline{x}^2},$$

$$\widehat{v} = \overline{\ln(y)} - \frac{\overline{x\ln(y)} - \overline{x}\cdot\overline{\ln(y)}}{\overline{x^2} - \overline{x}^2}\overline{x}.$$

A II. feladat megoldása: a

$$\frac{\partial E_2}{\partial a}(a,b) = 0, \qquad \frac{\partial E_2}{\partial b}(a,b) = 0,$$

egyenletrendszer az alábbi alakot ölti:

$$\sum_{i=1}^{n}(y_i - ba^{x_i})x_i a^{x_i} = 0, \qquad \sum_{i=1}^{n}(y_i - ba^{x_i})a^{x_i} = 0.$$

Ezt általában nem tudjuk explicit módon megoldani.

Az $(\widehat{a},\widehat{b}) = (e^{\widehat{u}}, e^{\widehat{v}})$ összefüggés általában nem igaz.

Példát adunk arra is, mikor teljesül és arra is, mikor nem.

1. Példa: Legyen $n = 3$ és

$$(x_1, y_1) := (1, 6), \quad (x_2, y_2) := (2, 18), \quad (x_2, y_2) := (3, 54).$$

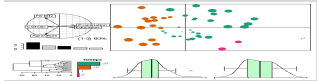
Ekkor

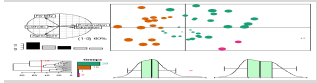
$$\widehat{u} \approx 1.09861, \qquad \widehat{v} \approx 0.693139, \qquad \widehat{a} = 3, \qquad \widehat{b} = 2,$$

és teljesül az $(\widehat{a},\widehat{b}) = (e^{\widehat{u}}, e^{\widehat{v}})$ összefüggés.

Továbbá,

$$E_1(\widehat{u},\widehat{v}) \approx 0, \qquad E_2(\widehat{a},\widehat{b}) = 0.$$





2. Példa: Legyen $n = 3$ és

$$(x_1, y_1) := (1, 4), \quad (x_2, y_2) := (2, 22), \quad (x_2, y_2) := (3, 50).$$

Ekkor

$$\widehat{u} \approx 1.26286, \quad \widehat{v} \approx 0.270724, \quad \mathrm{e}^{\widehat{u}} \approx 3.53553, \quad \mathrm{e}^{\widehat{v}} \approx 1.31091,$$

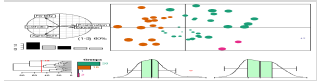
$$\widehat{a} \approx 2.60736, \qquad \widehat{b} \approx 2.84918,$$

és nem teljesül az $(\widehat{a}, \widehat{b}) = (\mathrm{e}^{\widehat{u}}, \mathrm{e}^{\widehat{v}})$ összefüggés.

Továbbá,

$$E_1(\widehat{u}, \widehat{v}) \approx 94.873, \qquad E_2(\widehat{a}, \widehat{b}) \approx 18.929,$$

azaz a nemlineáris regressziós modell „illeszkedik” jobban.



Az előző példából legalább két tanulság is leszűrhető:

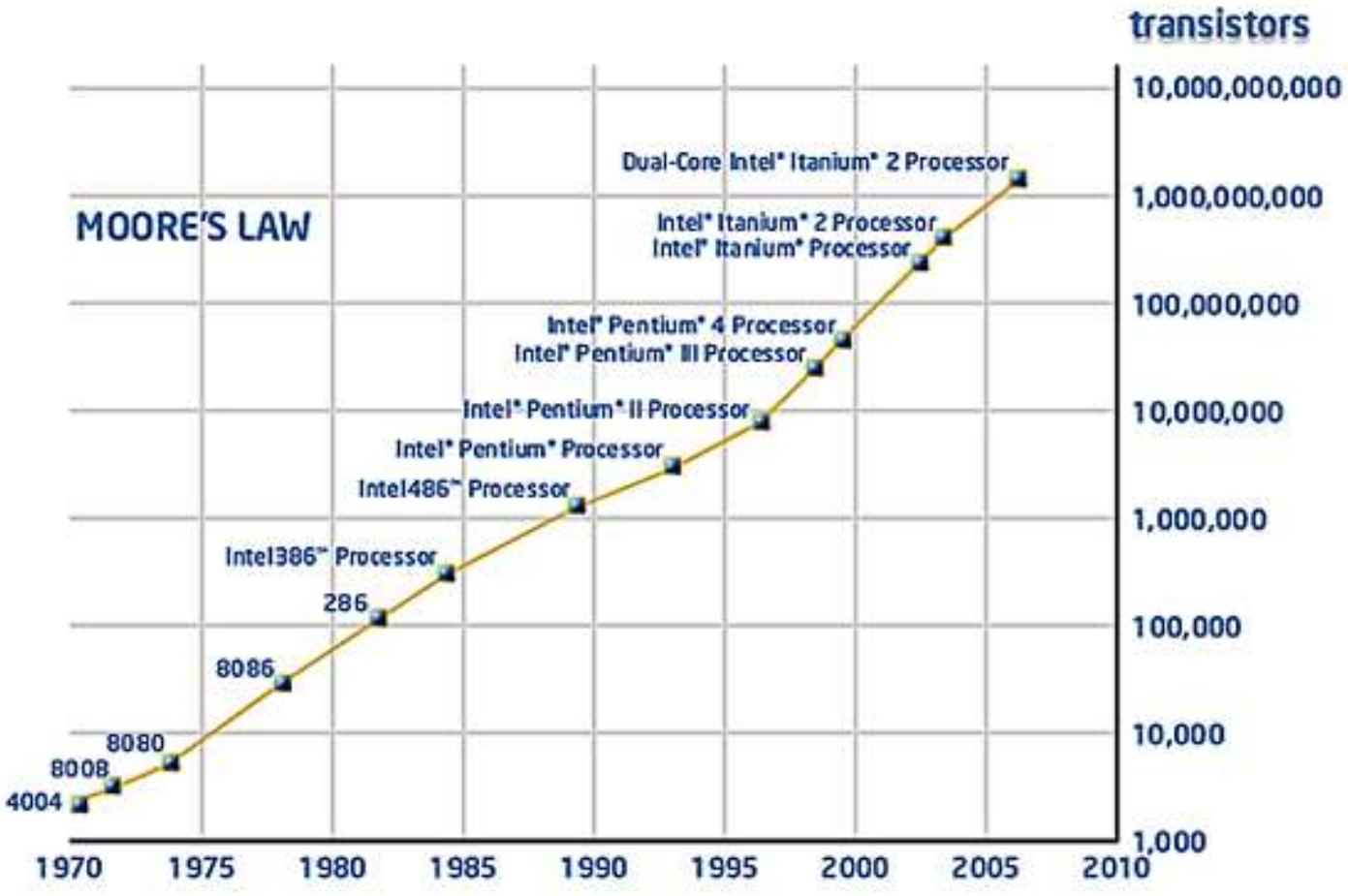
- a nemlineáris regressziós feladatnak és linearizáltjának a megoldása nem ekvivalens egymással.
- alapértelmezés szerint egyik módszer sem tekinthető jobbnak a másiknál, hiszen az egyik esetben egy nemlineáris egyenletrendszert kell megoldanunk, a másik esetben pedig linearizálunk.

Azt, hogy melyik módszerrel kapott becslést fogadjuk el további vizsgálatok, szakmai megfontolások dönthetik el.

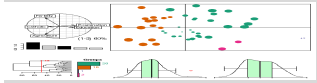
Példa: Moore–törvény

Gordon Moore, az Intel egyik alapítója fogalmazta meg 1975–ben:

Az integrált áramkörökben lévő tranzisztorok száma minden 24. hónapban (azaz 2 évente) megduplázódik.



Copyright: Intel Corporation. A függőleges tengelyen a tranzisztorok számának logaritmusa van ábrázolva.



Diszkrét és folytonos változó kapcsolata a változók közötti ok–okozati viszony esetén

- Ha a magyarázó változó diszkrét, akkor azt faktornak nevezzük.
- Ha a magyarázó változó folytonos, akkor a kovariáns elnevezéssel élünk.

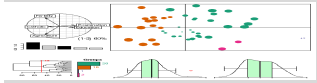
Statisztikai módszerek:

- Diszkrét magyarázó és folytonos függő változó: szórásanalízis.
- Folytonos magyarázó és diszkrét függő változó: osztályozási feladat.

Grafikus módszerek:

- Diszkrét magyarázó és folytonos függő változó: csoportosított hisztogram és doboz ábra.
- Folytonos magyarázó és diszkrét függő változó: halmozott hisztogram.

©Barczy Mátyás és Ispány Márton 2010



Szórásfelbontás I.

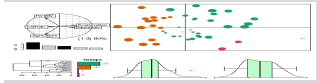
Figyeljük meg az Y folytonos változót és az F (diszkrét) faktort (ez a korábban X–szel jelölt magyarázó változó).

A minta:

$$(y_1, f_1), (y_2, f_2), \ldots, (y_n, f_n).$$

Tegyük fel, hogy az F faktor k értéket vehet fel, amelyeket nyilván kódolhatunk az $1, 2, \ldots, k$ számokkal.

Ezeket az értékeket szinteknek nevezzük.



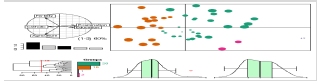
Szórásfelbontás II.

A k szint alapján a mintát az alábbi módon oszthatjuk fel:

$$y_{11}, y_{12}, \ldots, y_{1n_1}$$
$$y_{21}, y_{22}, \ldots, y_{2n_2}$$
$$\vdots$$
$$y_{k1}, y_{k2}, \ldots, y_{kn_k}$$

Az i–edik sor előállítása: az (y_ℓ, f_ℓ), $\ell = 1, \ldots, n$ mintaelemek közül megkeressük azokat, melyeknél $f_\ell = i$, a hozzájuk tartozó y–ok alkotják az i–edik szinthez tartozó megfigyeléseket: y_{ij}–nél az első index a szintet, a második pedig a szinten belüli sorrendet jelöli. Nyilván, $\sum_{i=1}^{k} n_i = n$.

Kérdés: van–e szerepe a faktornak, mennyire magyarázza a mintaelemeknek a (teljes) mintaátlag körüli szóródását?



Szórásfelbontás III.

Vezessük be az alábbi jelöléseket:

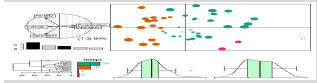
$$\overline{y_{i\cdot}} := \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad \text{az } i\text{–edik szint átlaga},$$

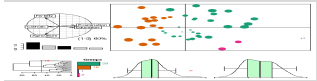
$$\overline{y_{\cdot\cdot}} := \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} \quad \text{teljes átlag},$$

$$Q_T := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{\cdot\cdot}})^2 \quad \text{teljes négyzetösszeg},$$

$$Q_K := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})^2 \quad \text{külső négyzetösszeg},$$

$$Q_B := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i\cdot}})^2 \quad \text{belső négyzetösszeg}.$$





Megjegyzés

- $$Q_K = \sum_{i=1}^{k} n_i (\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})^2.$$
- használatosak még az alábbi jelölések is:

$$Q_T = SST, \qquad Q_K = SSK, \qquad Q_B = SSB,$$

ahol SS=Sum of Squares és T=teljes, K=külső, B=belső.

Tétel (Szórásfelbontás)

$$Q_T = Q_K + Q_B \qquad (SST = SSK + SSB).$$

Bizonyítás. Alkalmazzuk a teve–szabályt:

$$(y_{ij} - \overline{y_{..}})^2 = ((y_{ij} - \overline{y_{i.}}) + (\overline{y_{i.}} - \overline{y_{..}}))^2$$
$$= (y_{ij} - \overline{y_{i.}})^2 + (\overline{y_{i.}} - \overline{y_{..}})^2 + 2(y_{ij} - \overline{y_{i.}})(\overline{y_{i.}} - \overline{y_{..}})$$

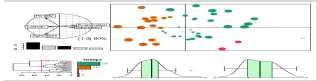
Világos, hogy elég belátni

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i.}})(\overline{y_{i.}} - \overline{y_{..}}) = 0.$$

Ez következik abból, hogy mivel $\overline{y_{i.}} - \overline{y_{..}}$ nem függ j–től, ezért a belső szummából kiemelhető és $\overline{y_{i.}}$ definíciója alapján

$$\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i.}}) = \sum_{j=1}^{n_i} y_{ij} - n_i\overline{y_{i.}} = 0.$$

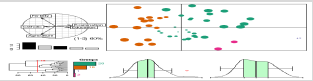
□



Heurisztika

A szórásfelbontásból következtethetünk a két változó közötti függőség erősségére.

- Ha a Q_K külső négyzetösszeg relatíve nagy a Q_B belső négyzetösszeghez képest, akkor ez arra utal, hogy az Y–ra vett együttes mintában lévő ingadozás elsősorban a szintek közötti különbségből adódik. Így az F faktor hatással van az Y függő változóra.
- Ha viszont a Q_K külső négyzetösszeg relatíve kicsi a Q_B belső négyzetösszeghez képest, akkor ez arra utal, hogy az Y–ra vett együttes mintában lévő inga–dozást elsősorban a szinteken belüli ingadozások magyarázzák. Így az F faktor nincs hatással az Y függő változóra.



H^2–mutató

$$H^2 := \frac{Q_K}{Q_T} \in [0,1].$$

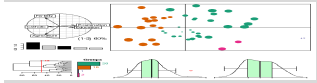
Értelmezés: H^2 az Y változó tapasztalati szórásnégyzetének az F faktor által megmagyarázott hányada.

A $H^2 = Q_K/Q_T$ tört számlálójában a Q_K külső négyzetösszeg annak felel meg, hogy minden egyes szint esetén a szint összes mintaelemét a szint átlagával helyettesítjük és, ha ez a négyzetösszeg „kicsi” a teljes négyzetösszeghez képest, akkor az F faktor csak „kicsit” magyarázza az Y mintaelemek szóródását.

A heurisztika pontosítására a későbbiekben kerül sor.

Nyilván,

$$H^2 = \frac{Q_K}{Q_T} = 1 - \frac{Q_B}{Q_T}.$$



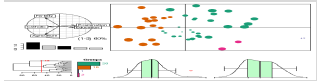
Megjegyzés

- $$H^2 = 0 \iff Q_K = 0 \iff \overline{y_{1\cdot}} = \cdots = \overline{y_{k\cdot}} = \overline{y_{\cdot\cdot}},$$

  azaz minden szint átlaga megegyezik a teljes minta-átlaggal. A szintek között átlag szempontjából nincs különbség, a faktornak nincs szerepe.

- $$H^2 = 1 \iff Q_B = 0$$
  $$\iff y_{ij} = \overline{y_{i\cdot}}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i.$$

  Ez abban az értelemben függvényszerű kapcsolat, hogy ha megmondjuk, hogy melyik szintről van szó és, hogy mennyi az adott szint átlaga, akkor ezzel az adott szinthez tartozó összes mintaelemet is megmondtuk.



Teljes, külső és belső szórásnégyzet

Teljes (tapasztalati) szórásnégyzet:

$$s_T^2 := \frac{Q_T}{n}$$

Külső (tapasztalati) szórásnégyzet:

$$s_K^2 := \frac{Q_K}{n}$$

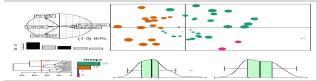
Belső (tapasztalati) szórásnégyzet:

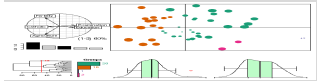
$$s_B^2 := \frac{Q_B}{n}$$

Az i–edik szinten belüli (tapasztalati) rész–szórásnégyzet:

$$s_i^2 := \frac{\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i\cdot}})^2}{n_i}, \qquad i = 1, \ldots, k.$$

©Barczy Mátyás és Ispány Márton 2010





Megjegyzés

- $s_T^2 = s_K^2 + s_B^2$,
- $$s_B^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_{i\cdot}})^2}{n} = \frac{\sum_{i=1}^{k} n_i s_i^2}{n},$$

  azaz a rész–szórásnégyzeteknek az egyes szintekhez tartozó mintaelemek számával súlyozott számtani közepe a belső szórásnégyzet, és nem a teljes szórásnégyzet.
- $$H^2 = \frac{s_K^2}{s_T^2} = 1 - \frac{s_B^2}{s_T^2}.$$

Osztályozási feladat

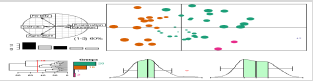
Ha az Y függő változó diszkrét, akkor osztályozási feladatról beszélünk. Ekkor ugyanis Y értékei egy–egy csoportot jelölnek ki, és az X folytonos (valós értékű) magyarázó változó (kovariáns) segítségével akarjuk azt eldönteni, hogy a megfigyelés (rekord) melyik csoportba tartozik.

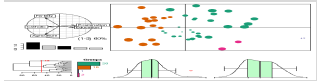
Bináris osztályozási feladat: Y értékei 0 vagy 1. Ekkor a mintát két csoportba, egy 0–ás és egy 1–es csoportba oszthatjuk. Az alábbiakban csak ezzel foglalkozunk.

Példa

- Tranzakciók vizsgálata (csalás – legális).
- Ügyfélszegmentáció (jó – rossz ügyfél).
- Betegség felismerés (igen – nem).

Döntésfüggvény: egy $d : \mathbb{R} \to \{0, 1\}$ függvény, amellyel a számegyenest két részre bontjuk. A két részhalmaznak feleltetjük meg ezután a két csoportot.





A feladat leírása: minden mintaelemet a 0–ás vagy 1–es csoportba szeretnénk besorolni. Már rendelkezésre áll egy 0–ás minta, illetve egy 1–es minta (azaz vannak mintaelemeink, melyeket a 0–ás, illetve az 1–es csoportba már besoroltunk).

Vezesük be az alábbi jelöléseket:

- $\overline{x_0}$: a 0–ás minta mintaátlaga,
- s_0^2: a 0–ás minta korrigálatlan emipirikus szórásnégyzete,
- $\overline{x_1}$: az 1–es minta mintaátlaga,
- s_1^2: az 1–es minta korrigálatlan emipirikus szórásnégyzete.

Lineáris döntésfüggvény

Tegyük fel, hogy $s_0^2 = s_1^2$.

Az x mintaelemet akkor soroljuk a 0–ás csoportba, ha

$$x\left(\mathbb{1}_{\{\overline{x_0}>\overline{x_1}\}} - \mathbb{1}_{\{\overline{x_0}\leq\overline{x_1}\}}\right) \geq \left(\mathbb{1}_{\{\overline{x_0}>\overline{x_1}\}} - \mathbb{1}_{\{\overline{x_0}\leq\overline{x_1}\}}\right)\frac{\overline{x_0}+\overline{x_1}}{2}. \tag{1}$$

Azaz $\overline{x_0} > \overline{x_1}$ esetén az x mintaelemet akkor soroljuk a 0–ás csoportba, ha

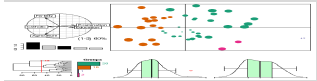
$$x \geq \frac{\overline{x_0}+\overline{x_1}}{2},$$

míg $\overline{x_0} \leq \overline{x_1}$ esetén akkor, ha

$$x \leq \frac{\overline{x_0}+\overline{x_1}}{2}.$$

Ekkor $d : \mathbb{R} \to \{0,1\}$, $d(x) := 0$, ha $x \in \mathbb{R}$ olyan, hogy (1) teljesül, egyébként pedig $d(x) := 1$.

©Barczy Mátyás és Ispány Márton 2010



Kvadratikus döntésfüggvény

Az x mintaelemet akkor soroljuk a 0–ás csoportba, ha

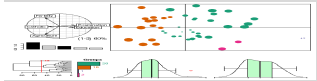
$$\left(\frac{x-\overline{x_0}}{s_0}\right)^2 - \ln s_0^2 \leq \left(\frac{x-\overline{x_1}}{s_1}\right)^2 - \ln s_1^2. \qquad (2)$$

Ekkor $d : \mathbb{R} \to \{0,1\}$, $d(x) := 0$, ha $x \in \mathbb{R}$ olyan, hogy (2) teljesül, egyébként pedig $d(x) := 1$.

Megj. Ellenőrizhető, hogy $\overline{x_0} \neq \overline{x_1}$ és $s_0^2 = s_1^2$ esetén a kvadratikus döntésfüggvénnyel is ugyanazt a döntést hozzuk meg, mint a lineáris döntésfüggvénnyel.
Ha $\overline{x_0} = \overline{x_1}$ és $s_0^2 = s_1^2$, akkor a kvadratikus döntésfüggvénnyel mindig a 0–ás csoportba soroljuk x–et, a lineáris döntésfüggvénnyel nem mindig.

A kérdéskörrel általánosabban foglalkozik a diszkriminancia analízis és a logisztikus regresszió.

©Barczy Mátyás és Ispány Márton 2010



Több változó kapcsolata

Ilyen kapcsolatok vizsgálatára a statisztika egy rendkívül szerteágazó eszköztárat hozott létre (többváltozós statisztikai módszerek). Valójában ezek a módszerek képezik a statisztikai szoftverek gerincét.

Ezek közül egy hatékony leíró, vizualizációs eszköz az ún. többdimenziós skálázás (MDS: multidimensional scaling). A módszer során a megfigyeléseinket (az adatbázis rekordjait) ábrázoljuk egy alacsony (2 vagy 3) dimenziós térben.

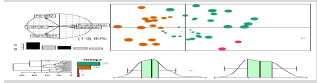
Legyenek a vizsgált statisztikai változók $X_1, X_2, \ldots, X_p$, melyeket egy vektorba írunk:

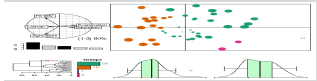
$$\boldsymbol{X} := (X_1, X_2, \ldots, X_p).$$

Az $\boldsymbol{X}$–re vonatkozó megfigyelések, melyek szintén vektorok (az adatbázis rekordjai vagy sorai), pedig legyenek:

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n.$$

Ez egy ún. többdimenziós minta.

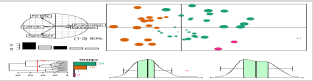




Határozzuk meg ezután a mintaelemek páronkénti távolságát valamilyen távolság definíció (metrika) alapján. Ez lehet a szokásos euklideszi távolság, de lehet más metrika is.

Távolság (metrika): olyan kétváltozós nemnegatív függvény, amely akkor és csak akkor 0, ha a két változója egybeesik, szimmetrikus és teljesül rá a háromszög–egyenlőtlenség.

Ha d–vel jelöljük a szóbanforgó metrikát, akkor elkészíthető az ún. távolságmátrix: $(d(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$



Konfiguráció: egy olyan n pontból álló pontrendszer a k dimenziós euklideszi térben (ahol k kicsi), amelynek pontjai kölcsönösen megfeleltethetőek a megfigyeléseinknek úgy, hogy a pontok közti euklideszi távolság közel van a megfigyelések között fent bevezetett távolsághoz.

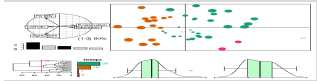
Megoldás–típusok:

- metrikus, ahol számít a távolság nagysága,
- nem–metrikus, ahol csak a távolságok sorrendje számít (Shepard–Kruskall algoritmus).

Alkalmazások: marketing, régészet, e–learning.

Irodalomjegyzék

1. Fazekas I.: (szerk.), Bevezetés a matematikai statisztikába. Kossuth Egyetemi Kiadó. Debrecen, 2003.
2. Hunyadi L., Vita L.: Statisztika közgazdászoknak. KSH, Budapest, 2002.



Összefoglalás

1. A kapcsolat típusai.
2. Két nominális skálán mért diszkrét változó: Cramér– és Csuprov–mutató.
3. Két ordinális skálán mért diszkrét változó: Spearman-féle rangkorrelációs együttható.
4. Két folytonos változó: korrelációs együttható, regressziós egyenes.
5. Egy diszkrét és egy folytonos változó: szórásfelbontás és osztályozási feladat.
6. Többdimenziós skálázás.