Csicsman József-Sipos Szabó Eszter
csicsman@calculus.hu, siposeszti@gmail.com

Matematikai alapok az adatbányászati szoftverek első megismeréséhez

1.1 A statisztikai sokaság

A **statisztika** a valóság számszerű információinak megfigyelésére, összegzésére, elemzésére és modellezésére irányuló gyakorlati tevékenység és tudomány. Feladata a tömegesen előforduló jelenségek tömör, számszerű jellemzése.

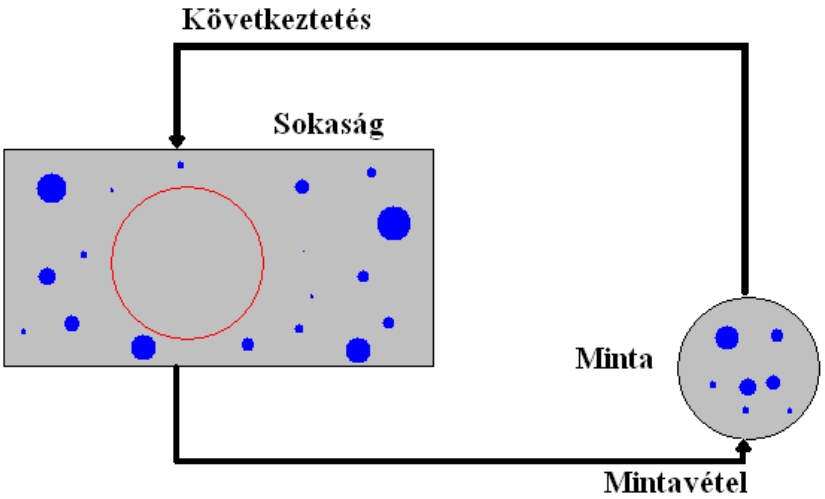
A statisztikai vizsgálat tárgya egy **rendszer**. Egy rendszernek elemei (**objektumai**) vannak, az objektumoknak **tulajdonságai** (objektumok lehetnek például egy iroda dolgozói; tulajdonság pedig a dolgozók fizetése). Az elemeknek mind közös, mind különböző tulajdonságokkal rendelkezniük kell, ez teszi lehetővé statisztikai célokra való csoportosításukat.

Egy rendszernek általában több objektuma, azoknak több, nem ritkán végtelen sok értékű tulajdonsága van. A rendszert alkotó objektumok, pontosabban azok tulajdonságait leíró (végtelen) sok jellemző változó adat alkotja az adatok **sokaságát**. A statisztikai sokaságnak két típusát szokták megkülönböztetni aszerint, hogy időpontban (**álló sokaság**) vagy időintervallumban (**mozgó sokaság**) vizsgáljuk-e. Álló sokaság például egy iroda dolgozóinak a fizetése 2005. január 1-jén; mozgó sokaság egy iroda dolgozóinak a fizetése 2005-ben.

A sokaság szabatos meghatározása fontos feltétele a statisztikai munkának, hiszen ez jelenti a feldolgozásra váró adatok pontos meghatározását.

1.2 A valóság és a statisztikai mérés kapcsolata

Általában nincs arra módunk, hogy a vizsgált rendszerről mindent megtudjunk, csak annak egy bizonyos állapotát figyelhetjük meg, azaz csak **mintát** vehetünk leíró adataiból, s majd a minta vizsgálatának eredményéből következtetünk a rendszerre. A minta vétele tehát az eredmények értéke szempontjából elsőrendűen fontos.



1. ábra A valóság és a statisztikai mérés kapcsolata

A jó minta:

- reprezentatív, összetételében helyesen képviseli a sokaságot, amelyből vették,
- véletlen, a mintaelemek egymástól függetlenül, egyenlő valószínűséggel kerülnek a mintába,
- elégséges méretű, elegendően nagy ahhoz, hogy a minta alapján levont következtetések kellően valószínűek legyenek.

A **nagy számok törvénye** szerint a mintaelemszám növelésével tetszőlegesen csökkenthető annak a valószínűsége, hogy a mintaátlag eltérése a sokaság várható értékétől meghaladjon egy rögzített hibakorlátot. Azaz minél nagyobb a minta, annál pontosabb lehet a becslés.

1.3 A sokaság adattípusai [1,2]

A sokaság (adathalmaz) elemzése az adatok megismerésével kezdődik. Lényeges információ, hogy milyen típusú, azaz milyen skálán mért alkotják az adathalmazt, mert bizonyos statisztikai jellemzők és matematikai–statisztikai eljárások csak bizonyos adatskálára számíthatóak, illetve alkalmazhatóak. Például pusztán statisztikai alapon sem állítható, hogy az 1. helyen végzett úszó 10-szer ügyesebb a 10. helyen végzettnél.

Az adattípusok bemutatásához legyen adott két megfigyelési egység, A és B, és ezekre vonatkozó két mért vagy számbavett jellemző ismérv (adat): X_a és X_b.

- **Nominális** skálán mért adatok

Vannak olyan jellemzők, amelyek esetében csak az dönthető el, hogy A és B jellemző szerint megegyezik egymással vagy különbözik egymástól. E skála esetében két vagy több értéke, kimenete lehet a változóknak.

A nominális skálára példa a nemzetiségi hovatartozás vagy a tüdődaganatok szövettani beosztása (kissejtes rák, nagysejtes rák, mirigyhám eredetű rák, laphámrák). Ezekben az

esetekben az egyes kategóriák között nincsen mennyiségi összefüggés, nem lehet azt mondani, hogy az egyik kategóriába tartozó elem nagyobb, több, stb., mint a másikba tartozó.

A nominális skála esetében a skálaértékek előfordulásának gyakorisága és a módusz vizsgálható. Azonban sem medián, sem átlag nem értelmezhető.

- **Ordinális** skálán mért adatok

Ez a nominális skála rokonának tekinthető, de ebben az esetben az egyes kategóriák sorba rendezhetők, meg tudjuk mondani, melyik a „jobb" vagy „több". Azt azonban a számértékek nem tüntetik fel, hogy az objektumok közötti eltérés mértéke mekkora.

Az ordinális skálára példa a Mohs-féle skála. A Mohs-féle skála azon szempont szerint rangsorolja az ásványokat, hogy melyik karcolja a másikat. A gyémánt áll a legmagasabb helyen, mert az összes ismert ásványt karcolja, míg őt magát semmilyen más ásvány nem karcolja. A skála azonban arról nem ad felvilágosítást, hogy a gyémánt mennyivel keményebb más ásványoknál. (Mérő László: A pszichológiai skálázás matematikai alapjai. Tankönyvkiadó, Budapest, 1992)

E skálatípus esetében medián vizsgálható, átlagról ellenben itt nincs értelme beszélni. Mediánról beszélhetünk akkor, ha például azt mondjuk, hogy a bazalt az a keménységi fok, aminél egy adott ásványminta fele puhább, másik fele pedig keményebb.

- **Intervallum**skálán mért adatok

Még finomabb összevetésre van mód, amikor a megegyezés – meg nem egyezés, valamint a sorrendiség megállapításán túl az is kideríthető az adatokból, hogy mennyivel nagyobb az egyik érték a másiknál (értelmezhető az adatok különbsége is). Az intervallumskála nullapontjának és egységpontjának a meghatározása is megállapodás kérdése.

Az intervallumskálára példa a hőmérsékletmérés (Celsius- vagy Fahrenheit skála).

Itt már számolhatunk átlagot, mivel a nullapont eltolása nem változtatja meg az átlag relatív helyét az átlagolt számok között.

- **Arány**skálán mért adatok

Az arányskála az intervallumskála jellemzőivel rendelkezik, emellett értelmezhető az is, hogy egyik adat hányszorosa a másiknak. Tartalmaz egy abszolút nullapontot is, mely rögzítve van. Ugyanakkor a skála egysége itt is szabadon megválasztható: például mérhetjük méterben vagy yardban, ez a két távolság arányát nem befolyásolja.

A darabszámmal vagy intenzitással rendelkező mennyiségek tipikus arányskálát képviselnek.

Az arányskálára a számokra vonatkozó összes művelet alkalmazható.

Mérési skálák	**Tulajdonság**	**Értelmezhető relációk**	**Sajátosságok**	**Jellemző példák**
Nominális	Megkülönböztetés	$X_a=X_b$ vagy $X_a \neq X_b$	Nem számszerű	Név, születési hely, nem

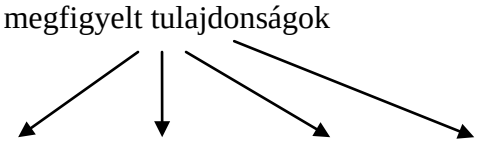
Ordinális	Megkülönböztetés, sorrend	$X_a=X_b$ vagy $X_a\neq X_b$ és $X_a\geq X_b$ vagy $X_a<X_b$	Nehezen mérhető, csak sorrendbe állítható	Sorrendek, (katonai) rangok
Intervallum	Megkülönböztetés, sorrend, különbség	$X_a=X_b$ vagy $X_a\neq X_b$ és $X_a\geq X_b$ vagy $X_a<X_b$ Értelmezhető X_a-X_b	Pozitív és negatív értékek	Hőmérséklet
Arány	Megkülönböztetés, sorrend, különbség, arány	$X_a=X_b$ vagy $X_a\neq X_b$ és $X_a\geq X_b$ vagy $X_a<X_b$ Értelmezhető X_a-X_b valamint X_a/X_b	Van elméleti minimum, azonos előjelű	Népességszám, jövedelem

1. táblázat A különböző mérési skálájú adatok tulajdonságai [1]

1.3 A statisztikai sokaság adatszerkezete

Az adathalmaz, azaz a statisztikai sokaság adatszerkezete alapvetően kétféle, **mátrixos** vagy **gráfos** lehet. Az alfejezet a mátrixos adatszerkezet leírását tartalmazza, a másik adatszerkezetről bővebben a Gráfos adatszerkezet című fejezetben olvashat.

Mátrixos adatszerkezetnél a statisztikai sokaság N objektummal és az objektumokhoz tartozó K mért adatottal írható le. Az adathalmaz megjeleníthető egy táblázatban, ahol a táblázat sorai az egyes objektumokra vonatkozó megfigyeléseket tartalmazzák, az oszlopok pedig a tulajdonságokra vonatkoznak (változók).



	1. tulajdonság	2. tulajdonság	…	k. tulajdonság
1-es objektum	X_{11}	X_{12}	…	X_{1k}
2-es objektum	X_{21}	X_{22}	…	X_{2k}
3-as objektum	X_{31}	X_{32}	…	X_{3k}
…	…	…	…	…
n-es objektum	X_{n1}	X_{n2}	…	X_{nk}

2. táblázat Mátrixos adatszerkezet

Az *2. táblázat* jelöléseit használva az n db objektum k db megfigyelt tulajdonsága alkotja az n*k méretű statisztikai sokaságot ($X_{11,}$ $X_{12,\ ...,}$ $X_{1k,\ ...,}$ X_{nk}). A sorok szerint minden objektumhoz (i-es objektum) hozzárendelhető egy vektor ($X_{i1,}$ $X_{i2,\ ...,}$ X_{ik}), aminek a komponensei az egyes megfigyelt tulajdonságokra vonatkoznak.

Az n objektumhoz n db k változós vektor tartozik. Ha a tulajdonságok (változók) intervallum-, illetve arányskálán mérhetőek, akkor az n db vektor elhelyezhető egy k dimenziós koordináta rendszerben. A könnyebb áttekinthetőség kedvéért érdemes az egymáshoz hasonló objektumokat megkeresni, azaz az adatokat csoportosítani a klaszteranalízis módszerével (bővebben lásd 4. fejezet). Így a továbbiakban elég n-nél kisebb számú csoporttal foglalkozni. Ha egyes tulajdonságok (változók) ordinális, illetve nominális skálán mért adatok, akkor más a helyzet. Ezen változók szerinti hasonló objektumokat úgy kereshetünk, hogy megnézzük, mely objektumok sorolhatóak azonos kategóriákba.

Nem csak az objektumok számát csökkenthetjük a könnyebb áttekinthetőség érdekében. Többváltozós adatelemzés során nagyon hasznos egy olyan eljárást alkalmazni, mely arra irányul, hogy a kezelhetőség érdekében a nagyszámú korreláló változókból új, kisebb számú, korrelálatlan változókat képezzünk. Egy ilyen eljárás a főkomponens-analízis (bővebben lásd 3. fejezet).

2 A leíró statisztika

A statisztikának alapvetően két nagy területe ismeretes; ezek között azonban sok találkozási pont, sőt átfedés figyelhető meg:

A **leíró statisztika** célja egy már rendelkezésre álló, valóságra vonatkozó adathalmaz összefoglalása, elemzése, egyszóval az információtömörítés.

A **következtető (matematikai) statisztika** célja a megfelelő – vagyis a sokaság egészének paramétereit legjobban tükröző, reprezentáló – minta kiválasztása, a sokasági paramétereknek a minta paramétereivel történő becslése, illetve a sokasági paraméterekre vonatkozó feltételezések, hipotézisek elfogadása vagy elvetése. Foglalkozik továbbá a valóság összefüggéseinek egyszerűsített megragadására törekvő modellekkel is, mint az idősor- és regressziós modellek.

2.1 Statisztikai jellemzők

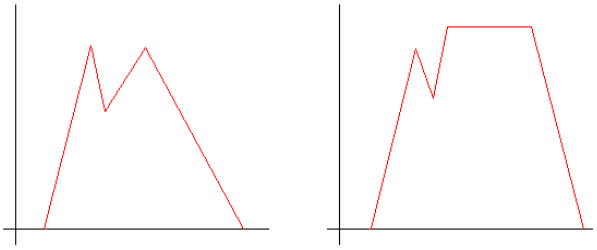
A statisztikai leírás célja a minta adatainak áttekinthető formába történő rendezése, tömörítése, és egyes jellemző értékeinek meghatározása. A jellemzőket általában 4 fő csoportba soroljuk:

1. **helyzetmutatók (középértékek)**: módusz, medián, átlag, kvantilis értékek
2. **szóródási mutatók**: terjedelem, szórás, relatív szórás
3. **koncentráció**: Lorenz-görbe

A fogalmak bemutatásához legyen adott a vizsgált X jellemzőre vonatkozó X_1, X_2,…,X_n n elemű minta.

Helyzetmutatók (középértékek):

- Módusz:
  Az eloszlás legnagyobb gyakoriságú értéke. Ez azonban nem mindig határozható meg egyértelműen.



3.ábra Nem egyértelmű móduszú eloszlások eloszlásfüggvényei [3]

A *3. ábrán* látható eloszlások arra figyelmeztetnek, hogy a mintánk nem volt homogén.
Ha a minta nem az, akkor valószínű, hogy a populáció sem volt az; ezek szerint - ha pl. az ország lakosainak testmagasságáról van szó – ez azt jelenti, hogy a férfiaknál és a nőknél az átlagos testmagasság értéke nem ugyanaz. Ilyenkor az eloszlásgörbe alapján érdemes "szétszedni" a mintánkat (többcsúcsúnál - nyilván - több részre).

- Medián: Véges sok elem (egy véges populáció) mediánján a következőt értjük:

  Ha páratlan elemszámú a sokaság, akkor a medián az értékek rendezett sokaságában a középső ((n+1)/2)-dik elem.

  Ha páros, akkor a rendezett minta két középső elemének számtani közepe.

  Példák [2]:

  1. Páratlan elemszám esetén:

  1 2 5 4 3 1 4 3 3 4 3 5 1
  A rendezett sokaság:
  1 1 1 2 3 3 3 3 4 4 4 5 5
  A medián a középső elem:
  1 1 1 2 3 3 **3** 3 4 4 4 5 5

  2. Páros elemszám esetén:

  1 4 2 4 2 3 5 3 1 1
  A rendezett sokaság:
  1 1 1 2 **2 3** 3 4 4 5
  A medián a középső elemek számtani közepe: 2,5.

- Átlag: mintaelemek számtani átlaga $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$

  Meg kell azonban jegyezni, hogy átlagolás absztrakciója folytán nyert átlag-adat nagyon sok esetben *nem is található meg* az eredeti eloszlásban. Erre példa az "Átlag János és családja"-jellegű statisztikai/szociológiai adathalmazok és a rajtuk alapuló elemzések.

- Kvantilis értékek
  A mennyiségi ismérv értékeit nagyság szerint növekvő sorrendbe állítva, majd az értékeket k számú egyenlő gyakoriságú csoportba osztva az egyes csoportok felső határai a kvantilis értékek.
  A különböző számú részekre való osztásokhoz a kvantilisek konkrét elnevezései tartoznak:

k	elnevezés
2	medián
4	kvartilis
5	kvintilis
10	decilis
100	percentilis

  Példa:
  alsó kvartilis (k_a) – olyan érték, amelyik a rendezett minta alsó 25%-át választja el a felső 75%-ától.
  felső kvartilis (k_f) – olyan érték, amelyik a rendezett minta alsó 75%-át választja el az alsó 25%-ától.

Szóródási mutatók:
Sok adatból álló minta egy adattal – pl. az átlaggal – való jellemzésekor/helyettesítésekor, mindenképpen információt veszítünk. Ez az ára annak, hogy a könnyebb intellektuális feldolgozhatóság érdekében több adat (egy eloszlás) egy adatba lett "belesűrítve". Az átlag, önmagában, nem jellemez elégséges pontossággal egy eloszlást.

Példa [3]:
Legyen két tanulónk: X és Y. (X nagyon okos, de lusta és szeszélyesen készül: ami érdekli, azt megtanulja; de amit nem érez magához közelállónak, azt semmi pénzért. Ezzel szemben Y képességeit tekintve átlagos - de szorgalmas.) Egy tárgyból szerzett érdemjegyeik egy adott időszakban:

X: 1 és 5 átlaguk: $\bar{x}$ = 3

Y: 3 és 3 átlaguk: $\bar{y}$ = 3

Látható, hogy $\bar{x}$ és $\bar{y}$ megegyeznek egymással (mindkettő közepes). Ám semmit sem mondanak arról, hogy ez a közepes átlagérték minek az eredményeként alakult ki: X-nél a két végletet "zsugorítja össze", míg Y esetében egyforma adatokat helyettesít.

Fentiekből látható, hogy $\bar{y}$ pontosan, míg $\bar{x}$ rosszul jellemzi a mintát, azaz az eloszlást. Ezért érdemes a szóródás mérőszámait is igénybe venni a minta jellemzéséhez.

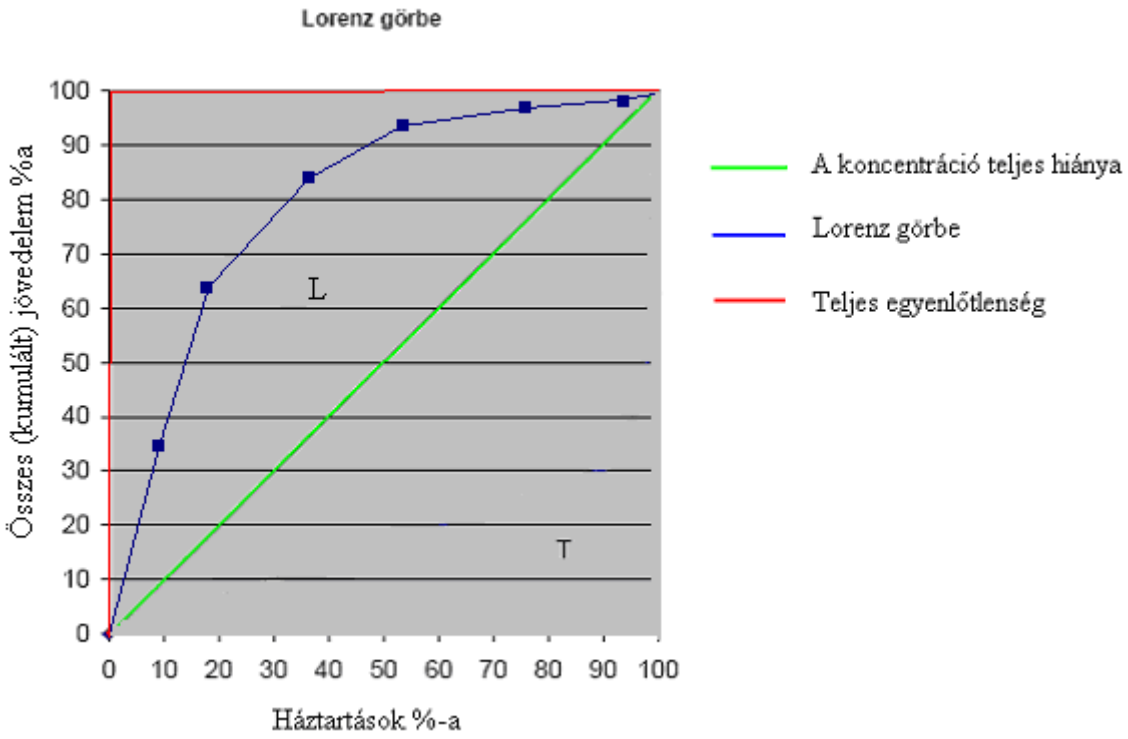
- Terjedelem
  A terjedelem a legnagyobb és a legkisebb mintaelem különbsége: $d = x_{max} - x_{min}$

- Szórás
  A szórás azt mutatja, hogy átlagosan milyen mértékben szóródnak a mintaelemek a minta átlaga körül:
  $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$, ami egyszerűsítés után: $\sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}}$

- Relatív szórás
  A minta szórása (s) viszonyítva a minta átlagához ($\bar{x}$): $V = \frac{s}{\bar{x}}$
  A relatív szórás mértékegység nélküli szám, amely megmutatja, hogy a minta értékei mennyire szóródnak átlagértékükhöz képest.

Koncentráció:

- Lorenz-görbe
  A Lorenz görbe a relatív koncentráció általános elemzési és szemléltetési eszköze. Annak bemutatására, hogy hogyan szerkeszthető meg a görbe, és pontosan mi is olvasható le róla, nézzünk példát!
  Legyen a statisztikai sokaságunk a magyar háztartások és a háztartások nettó jövedelmei! Rendezzük a háztartásokat nettó jövedelmeik szerint úgy, hogy az első háztartás legyen a legkisebb jövedelmű, az utolsó pedig a legnagyobb! Ezután már kiszámíthatjuk a Lorenz görbe pontjait (x,y), ha megnézzük, hogy a háztartások első x%-a a háztartások összes (kumulált) jövedelmének hány százalékával (y%) rendelkezik. Végül kössük össze ezeket a diszkrét pontokat! (*4. ábra*)



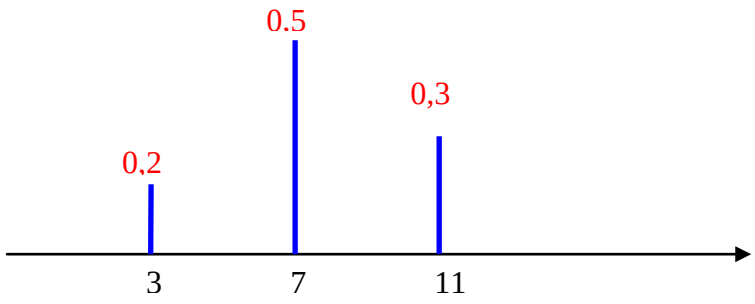
4.ábra Lorenz görbe

A Lorenz görbe tehát azt mutatja meg, hogy a háztartások alsó x%-a, az összes jövedelem hány százalékával rendelkezik. A főátló olyan elméleti helyzetet jelez, amelyben nincsenek jövedelemkülönbségek, tehát minden háztartásnak pontosan ugyanannyi a nettó jövedelme. A jövedelem koncentráció annál nagyobb, minél jobban eltávolodik a görbe az átlótól.
A Lorenz-görbe relatív értékeket tartalmaz, ezért összehasonlításokra a különböző népcsoportok, területek, országok jövedelemegyenlőtlenségének összehasonlítására, illetve időbeli összehasonlításokra jól alkalmazható.

- A Gini együttható
  A Gini együtthatót a koncentráció tömör jellemzésére használják. Azt mutatja meg, hogy a vizsgált sokaság koncentrációja mennyire tér el egy egyenletes eloszlású sokaság koncentrációjától.
  Úgy számítjuk ki, hogy a főátló és a Lorenz görbe közötti területet (*4. ábrán* L-lel jelölt terület) elosztjuk a főátló alatti területtel (*4. ábrán* T-vel jelölt terület). Tehát G=L/T. Mivel azonban a főátló felezi a négyzet területét, T=0,5, így G=2L.
  Ha minden háztartás jövedelme azonos, akkor G=0; míg ha egyetlen háztartásnak van csak jövedelme, akkor G=1.

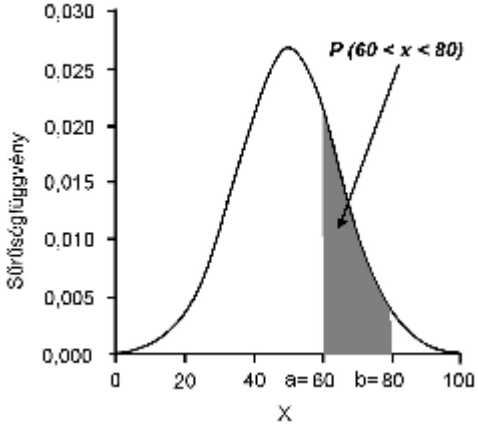
2.2 Nevezetes eloszlások

Ha tekintjük a populációban az egyedek adott tulajdonságának (X) összes lehetséges előforduló értékét, majd kiszámítjuk ezek relatív gyakoriságát (valószínűségét), úgy kapjuk a valószínűségi változó (X) **eloszlását**. Ez a megközelítés azonban csak **diszkrét valószínűségi változó** eloszlásának leírására alkalmazható.



5.ábra diszkrét valószínűségi változó eloszlása

Folytonos valószínűségi változó esetében annak a valószínűsége, hogy ez a változó pontosan egy adott értéket vesz fel, végtelenül kicsi szám, tehát folytonos valószínűségi változó esetében az eloszlás fenti értelmezése nem használható. Ezért egy folytonos valószínűségi változó eloszlásának leírására a **sűrűségfüggvényt** használjuk. A sűrűségfüggvény egy olyan függvény, melynek *a* és *b* közötti grafikon alatti területe megadja annak a valószínűségét, hogy a valószínűségi változó *a* és *b* számok közötti értéket vesz fel.



6.ábra folytonos valószínűségi változó sűrűségfüggvénye

A sűrűségfüggvényre teljesül az, hogy - ∞ és +∞ között a grafikon alatti területe 1-gyel egyenlő, hiszen ez annak a valószínűségét adja meg, hogy a valószínűségi változó értéke - ∞ és + ∞ között van, ami a biztos esemény, tehát valószínűsége 1.

2.2.1 Néhány nevezetes diszkrét eloszlás

- **Binomiális eloszlás:** A diszkrét eloszlások nagyon sok esetben megállapítható változók viselkedését írják le jól. Abban a – legegyszerűbb – esetben, ha a változó csak két értéket vehet fel, akkor az értékek eloszlása binomiális eloszlást határoz meg. Néhány példa binomiális eloszlásra: a balkezesség előfordulása, dichotóm (igen vagy nem) választ kérő tesztekben a hibás válaszok száma, tömeggyártásban a „selejt" – „nem selejt" előfordulása, „alkalmas" – „nem alkalmas", „megfelelt" – „nem felelt meg" személyek előfordulása.

  Ha a változó az egyik értéket (x_1) *p*, a másikat (x_2) pedig *(1-p)* valószínűséggel veszi fel, akkor annak a valószínűsége, hogy egy ebből a sokaságból véletlenszerűen vett *n* elemű minta éppen *k* darab x_1 elemet tartalmaz, a binomiális eloszlás alapján*:*

$$p_k = \mathrm{P}\,(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

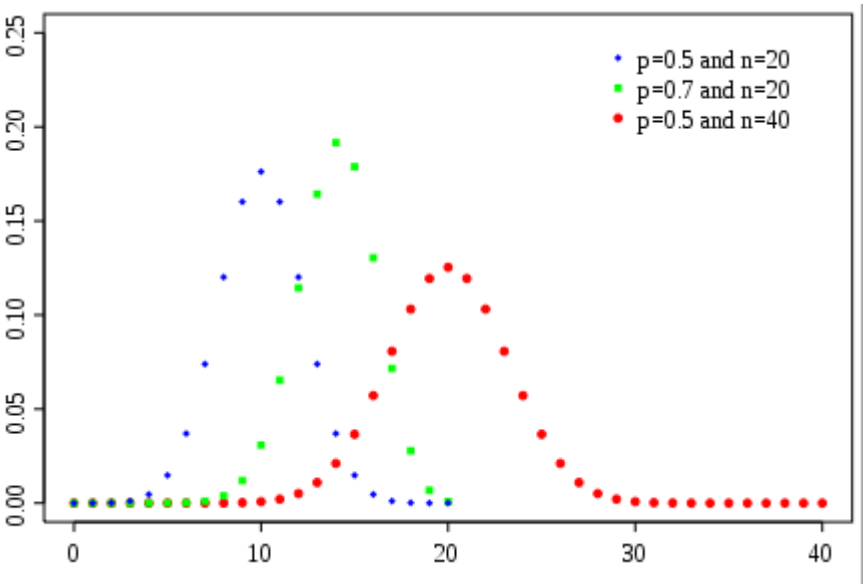
  Ezen paraméterek segítségével felírható a binomiális eloszlás szórása és várható értéke is, ahol a véges **várható érték**:

$$\mathrm{E}(X) = \sum_k x_k \cdot p_k$$

  feltéve, hogy ez az összeg véges. (Megjegyzés: X várható értéke az X által felvett értékek súlyozott számtani közepe. A valószínűségi változó a várható értéke körül mutat véletlen ingadozást.)

  A binomiális eloszlású valószínűségi változó **várható értéke**: *np*

  A binomiális eloszlású valószínűségi változó **szórásnégyzete**: *np(1-p)*



7. ábra Különböző paraméterű binomiális eloszlások

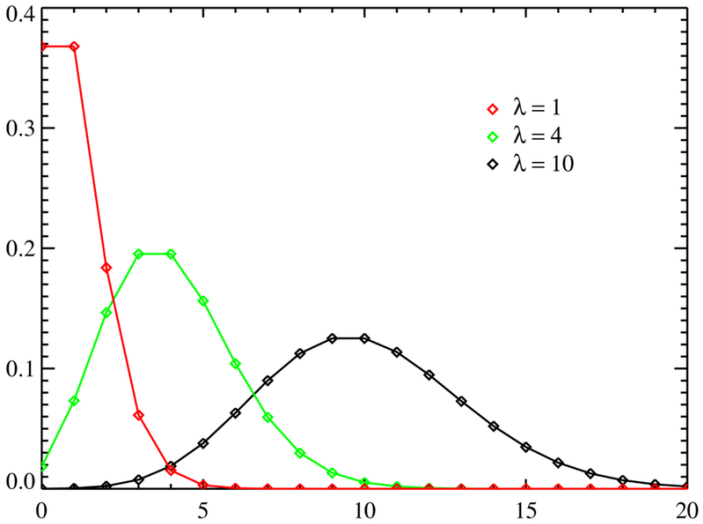
- **Poisson eloszlás**: A Poisson eloszlás a binomiális eloszlás határeloszlása, abban az értelemben, hogy ha binomiális eloszlások olyan sorozatát vesszük, melyben az eloszlások *n* paramétere úgy tart a végtelenbe, hogy közben az *np* szorzat konstans marad (*p* így nyilván a 0-hoz tart), akkor határeloszlásként Poisson eloszlást kapunk. A Poisson eloszlás kifejezi az adott idő alatt ismert valószínűséggel megtörténő események bekövetkezésének számát. Például: egy telefonközpontba adott időszakban és időtartamban beérkezett telefonhívások száma, egy radioaktív anyag adott idő alatt elbomló atomjainak száma, leszálló porszemek száma a tiszta papírlapon, egy számítógépes raktárkészlet-nyilvántartó rendszer napi "lefagyásainak" száma, egy hónapra jutó biztosítási események (betörés, tűz, viharkár, stb.) száma, egy számítógépes adatrögzítő operátor által egy óra alatt vétett adatbeviteli hibák száma, egy vasszerkezeten található hibás hegesztési varratok száma, egy öntvényben található zárványok száma.

  Az X valószínűségi változó λ paraméterű Poisson-eloszlású ha:

$$\mathbf{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

  ahol $0 < \lambda$ konstans.

  A Poisson eloszlású valószínűségi változó **várható értéke**: λ
  A Poisson eloszlású valószínűségi változó **szórásnégyzete**: λ



8. ábra Különböző paraméterű Poisson eloszlások

2.2.2 Néhány nevezetes folytonos eloszlás

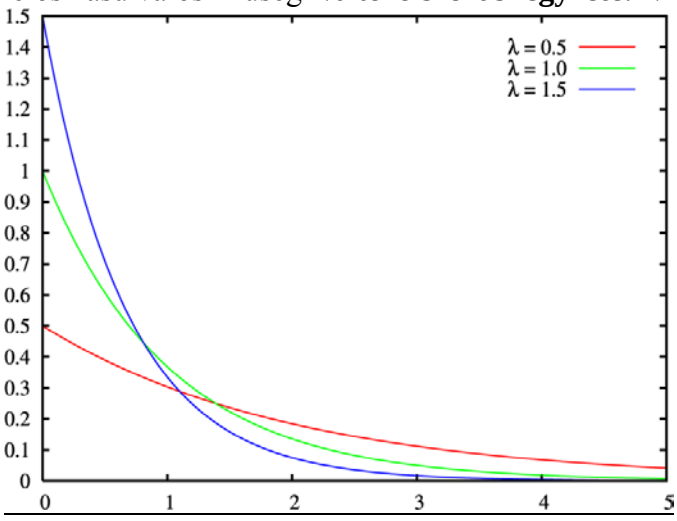
- **Exponenciális eloszlás**: a különböző dolgok élettartamának vizsgálatára, véletlen időtartamok (pl. várakozási idő) modellezésére használt eloszlás.

  Az exponenciális eloszlás kapcsolata a Poisson eloszlással:
  - Ha egy jelenség bekövetkezéseinek száma Poisson eloszlást követ, akkor a jelenség bármely két egymás utáni bekövetkezése között eltelt idő exponenciális eloszlású.
  - Ha tudjuk, hogy egy esemény a bekövetkezései között eltelt idő exponenciális eloszlású, akkor a bekövetkezések száma Poisson eloszlást követ.

  Az exponenciális eloszlás sűrűségfüggvénye:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

  Az exponenciális eloszlású valószínűségi változó **várható értéke**: $1/\lambda$
  Az exponenciális eloszlású valószínűségi változó **szórásnégyzete**: $1/\lambda^2$



9. ábra Különböző paraméterű exponenciális eloszlású változók sűrűségfüggvényei

- **Normális eloszlás**: A statisztikában a legfontosabb és leggyakrabban alkalmazott eloszlás. Normális eloszlással jellemezhetőek az olyan valószínűségi változók, melyek mért értékei a várható érték körül csoportosulnak. Ilyen például a magyar felnőtt férfiak magassága, aminek a várható értéke 179,1 cm. A legtöbb magyar felnőtt férfi magassága közel van ehhez az értékhez. A várható értéktől távolodva egyre kevesebb olyan férfit találunk, akinek a magassága nagyon eltér ettől pozitív vagy negatív irányban.
  A természetben, az orvostudományban nagyon sok mért paraméter normális eloszlással írható le, mint például az egyének vérnyomása, súlya, stb. A normális elnevezés is arra utal, hogy a mért adatainktól ezt várjuk, mert ez a természetes viselkedésük. Másik tipikus példa normális eloszlású valószínűségi változókra a mérési hibák.
  A gyakorlatban a **centrális határeloszlás tétel** azt mondja ki, hogy ha egy mennyiség (X) „valódi értékére" úgy próbálunk következtetni, hogy mért adatainkat átlagoljuk ($1/n(X_1+X_2+\ldots+X_n)$), akkor ez **az átlag tekinthető normális eloszlású valószínűségi változónak**, akármilyen is a mért adatok tényleges eloszlása.
  A centrális határeloszlás tétel következménye az is, hogy ha egy valószínűségi változó értékét nagyszámú, egymástól függetlenül ható véletlen tényező határozza meg úgy, hogy az egyes tényezők külön-külön csak igen kis mértékben járulnak hozzá az összes véletlen hatásból eredő ingadozáshoz, és az egyes tényezők hatásai összeadódnak, akkor általában normális eloszlású valószínűségi változót kapunk. (pl. egy tömeggyártás során a termék valamely numerikus jellemzőjét az anyagi jellemzők, a megmunkálási technológia, a tárolás, a szállítás, a hőmérséklet, a dolgozó felkészültsége, stb. együttesen határozza meg, a numerikus termékjellemzők gyakran normális eloszlásúak.)

  A normális eloszlású valószínűségi változó sűrűségfüggvénye:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}},$$

  ahol a két paraméter, m és σ valós számok, valamint $\sigma > 0$. m a normális eloszlású valószínűségi változó **várható értéke**; σ pedig a normális eloszlású valószínűségi változó **szórása.**

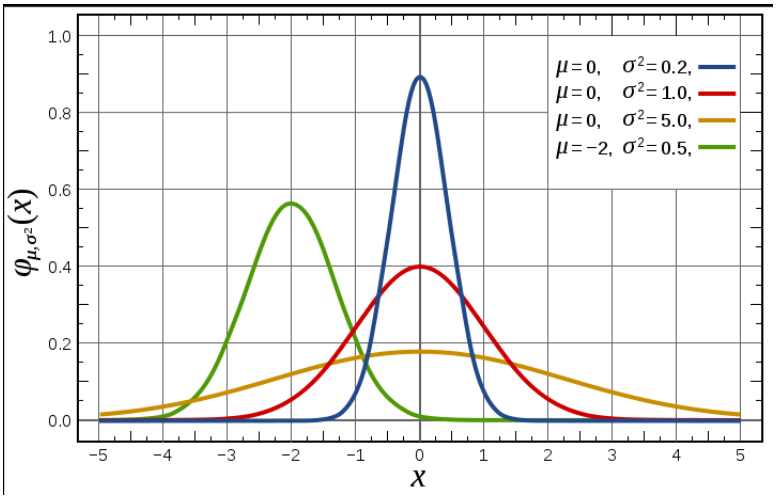
  Azt, hogy az X valószínűségi változó normális eloszlást követ, a következő módon szoktuk jelölni:

$$X \sim \mathcal{N}(m, \sigma^2).$$

  Speciálisan, ha $X \sim N(0, 1)$, akkor X-et standard normális eloszlásúnak nevezzük.

  Normális eloszlású valószínűségi változók néhány fontos tulajdonsága:
  - Ha $X \sim N(m, \sigma^2)$, akkor bármilyen nullától különböző valós a és bármilyen valós b szám esetén az $Y = aX + b$ valószínűségi változó is normális eloszlást követ, pontosabban $Y \sim N(am + b, a^2\sigma^2)$.
    Az eloszlás eme tulajdonságán alapul a **standardizálás** módszere: ha $X \sim N(m, \sigma^2)$, akkor $(X–m)/\sigma \sim N(0, 1)$.

- Normális eloszlású független valószínűségi változók összege is normális eloszlású. Pontosabban, ha $X_1 \sim N(m_1, \sigma_1^2)$ és $X_2 \sim N(m_2, \sigma_2^2)$ független valószínűségi változók, akkor $X_1 + X_2 \sim N(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.
  Fordítva: ha X_1 és X_2 független valószínűségi változó, és $X_1 + X_2$ normális eloszlású, akkor X_1 is és X_2 is normális eloszlású.



10. ábra Különböző paraméterű normális eloszlású változók sűrűségfüggvényei

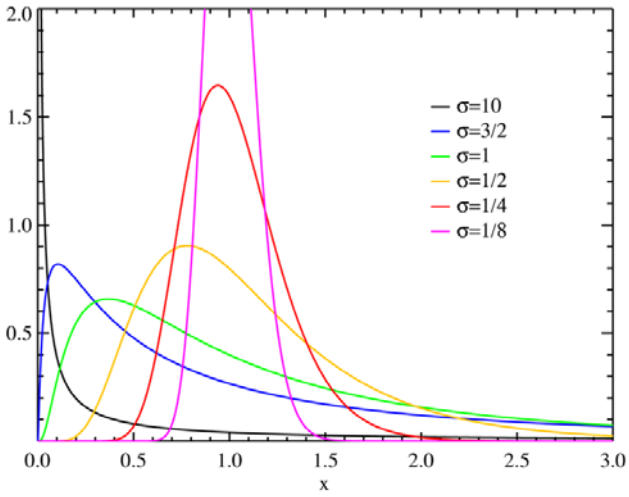
- **Lognormális eloszlás**: Ha az Y=lnX változó normális eloszlású, akkor az X valószínűségi változót lognormális eloszlásúnak nevezzük.
  A természetben bizonyos törési-aprítási folyamatoknál az őrlemény szemcsedarabjainak nagyság szerinti megoszlása lognormális eloszlást mutat. Ugyancsak jól közelíthető lognormális eloszlással egyes foglalkozási rétegek jövedelem-eloszlása, illetve az olyan "méretet" kifejező gazdasági jellemzők, mint az ügyfelek összes vásárlásai, a cégek árbevétele, a foglalkoztatottak száma, a telefonbeszélgetések időtartama stb.

  A lognormális eloszlás sűrűségfüggvénye:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0$$

  A lognormális eloszlású valószínűségi változó **várható értéke**: $e^{\mu+\sigma^2/2}$

  A lognormális eloszlású valószínűségi változó **szórásnégyzete**: $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$



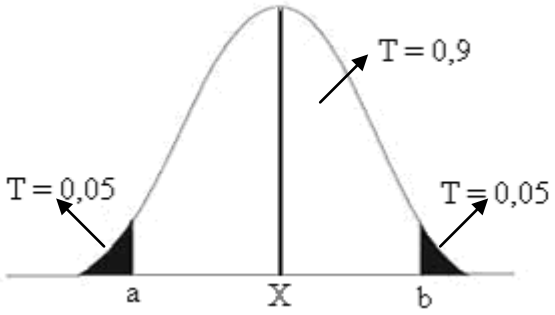
11. ábra Különböző paraméterű lognormális eloszlású változók sűrűségfüggvényei

2.3 További eloszlásjellemzők [2]

Az elméleti eloszlásjellemzőket becsülhetjük **pontbecsléssel** vagy **intervallumbecsléssel**.

Pontbecslés (pl. az elméleti átlagot a mintaátlaggal becsülve egy pontot jelölünk ki) esetén egy véges mintából következtetünk egy akár végtelen nagyságú sokaságra, ezért szinte biztos, hogy tévedünk.

Intervallumbecslés esetén egy olyan ún. **konfidencia** (megbízhatósági) **intervallumot** jelölünk ki az adott jellemző lehetséges helyéül egy alsó (a) és egy felső (b) határ megadásával, amely megadott (a 12. ábrán 90%-os) biztonsággal tartalmazza azt. A konfidencia intervallum csak akkor tartalmazza 100%-os biztonsággal a jellemzőt, ha annak teljes értékskálájára kiterjed (ekkor azonban semmitmondó).



12. ábra Az adott sűrűségfüggvényű valószínűségi változó mért értéke 90%-os biztonsággal fog az [a,b] intervallumba esni, azaz az [a,b] intervallum 90%-os szintű konfidenciaintervallum.

A konfidencia-intervallum szoros összefüggésben áll a **szignifikancia**teszttel. Ha a paraméter pontbecslése X, és az $[a,b]$ intervallum P szintű konfidenciaintervallum, akkor az intervallumon kívüli elemek (a-nál kisebb vagy b-nél nagyobb elemek) **szignifikánsan különböznek** X-től az **1-*P* szinten** ugyanazon feltevések mellett, mint amikkel a konfidencia-intervallumot előállítottuk.

3. A változók megismerése

A változók megismerése a 2. fejezetben, a 2.1 (Statisztikai jellemzők) alfejezetben leírt számítások elvégzésével kezdődik. Ezek az egyes változókat külön-külön vizsgálják, megmutatják azok alap statisztikai jellemzőit.
A változók közötti kapcsolatok feltérképezéséhez érdemes összefüggés vizsgálatot végezni. A mennyiségi jellemzők közötti kapcsolatok szorosságának jellemzésére a **korreláció**, a kapcsolatok típusának jellemzésére a **regresszió** használható.

A változók közötti kapcsolat lehet:

- **függvényszerű kapcsolat:** az egyik változó és a függvénykapcsolat egyértelműen és pontosan meghatározzák a második változó értékét.
- **valószínűségi (szochasztikus) kapcsolat**: a kapcsolat nem írható le egy függvénnyel, mert az a függvényszerű kapcsolat mellett még a véletlentől is függ.
- **egymástól nem függő (független) kapcsolat**

Az összefüggés vizsgálat lépései:

1. Az összetartozó értékek ábrázolása pontdiagramon.
2. A diagram alapján annak megállapítása, hogy a mennyiségi jellemzők között van-e kapcsolat. (Az összefüggés vizsgálat során a pontdiagramos ábrázolással érzékeltetett tendenciát valamilyen analitikusan ismert függvénnyel próbáljuk leírni)
3. Ha van kapcsolat a mennyiségi jellemzők között, akkor a regressziós függvény típusának megállapítása, majd a regressziós együtthatók (paraméterek) meghatározása. Értelmezésük.
4. A kapcsolat szorosságának meghatározása. (korrelációszámítás)
5. A regressziós függvény illeszkedésének ellenőrzése. Az illeszkedés jóságának meghatározása.

Fontos megjegyezni, hogy a korrelációs és regressziós számítás a kapcsolatot jellemzi, de semmit nem mond az oksági viszonyról. Tehát két, vagy több változó közötti sztochasztikus kapcsolat megállapításából nem következik, hogy a változók oksági összefüggésben vannak, azaz, hogy egyik tényező változása oka a másik tényező változásának. Az oksági kapcsolatot csak alapos szakmai és statisztikai vizsgálattal lehet megállapítani**.**

3.1 Regresszió

A regressziószámítás során használt főbb függvénytípusok:

$\hat{y} = a + bx$	lineáris
$\hat{y} = a \cdot b^x$	exponenciális
$\hat{y} = a \cdot x^b$	hatvány
$\hat{y} = a + \frac{b}{x}$	hiperbolikus
$\hat{y} = a + bx + cx^2$	másodfokú

ahol a, b, c a ***regressziós együtthatók***.

A pontdiagram alapján megbecsülhető a regressziós függvény típusa. Következő lépésként a végtelen sok ilyen típusú (de különböző regressziós együtthatójú) függvény közül a legjobban illeszkedő a ***legkisebb négyzetek elve*** alapján választható ki.

A legkisebb négyzetek elve: Az a függvény lesz legjobban illeszkedő (az adott típusúak közül), amelyikre teljesül, hogy a mért és számított eredményváltozó-értékek különbségeinek négyzetösszege minimális.

$(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \rightarrow \min$. $E_{ij}C_{ij}$ $\hat{y}_i$ jelöli a számított, y_i a mért eredményváltozó-értékeket)

A regresszióanalízis során feltételezzük, hogy:

- y_i az x_i-k minden értékénél normális eloszlású, vagyis az $y_i - \hat{y}_i$ hogy a mért és számított eredményváltozó-értékek különbségei (azaz a mérési hibák) $N(0,s^2)$ normális eloszlásúak;
- a különbözõ *i* mérési pontokban elkövetett mérési hibák egymástól függetlenek.

A különböző függvénytípusoknál különbözőképp számíthatjuk ki a legjobban illeszkedő függvény regressziós együtthatóit. Bizonyítás nélkül álljon itt a legegyszerűbb, a lineáris függvény ($\hat{y} = a + bx$) regressziós együtthatóinak (a,b) kiszámítási módja:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad a = \bar{y} - b \cdot \bar{x}$$

A regressziós függvény illeszkedésének jósága a relatív hibával mérhető. **Ha a relatív hiba 10% alatti,** akkor a regressziós függvényt a vizsgált jelenség **matematikai modell**jének szokták tekinteni. A relatív hiba:

$V_{S_y} = \frac{s_y}{\bar{y}} \cdot 100\%$, ahol s_y a reziduális szórás: $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$, n pedig a minta elemszáma.

3.1.1 A logisztikus regresszió

Amikor a célváltozó bináris, azaz 2 lehetséges értéke van (pl. siker: 1, kudarc: 0), akkor a lineáris regresszió alkalmazhatatlan, mert folytonos, $-\infty$ és $+\infty$ között értelmezett magyarázó változók esetén a célváltozó becsült értéke nem feltétlenül fog 0 és 1 közé esni.

Az ilyen modellezésekhez használható a logisztikus regresszió [14]. A logisztikus regresszió két, egymást kölcsönösen kizáró kategória (siker: D = 1, kudarc: D = 0) bekövetkezési esélyeinek az egymáshoz való arányát, vagyis az ú.n. odds mértékét modellezi a magyarázó változók értékeinek ismeretében. Legyen a siker bekövetkezésének feltételes valószínűsége $P_x = P(D=1|x)$. (A siker bekövetkezésének feltételes valószínűsége azt jelenti, amikor a siker bekövetkezésének a valószínűségére vagyunk kíváncsiak, de ismeretes számunkra, hogy a magyarázó változók értéke x.) Így a kudarc valószínűsége $1- P_x$.

Ekkor a sikernek a kudarchoz viszonyított esélye:

$$odds_x = P_x/(1-P_x)$$

Ez a transzformáció a [0, 1] intervallumba eső P_x-et az $odds_x$-en keresztül a $(-\infty, +\infty)$ intervallumba képezi le. A logisztikus regresszió feltételezése szerint az odds logaritmusa a magyarázó változók lineáris függvénye. Így már használható lesz a lineáris regresszió.

$$\ln(odds_x) = \varepsilon_0 + \beta_1 x_1 + \ldots + \beta_n x_n,$$

ahonnan

$$odds_x = e^{\varepsilon_0 + \beta_1 x_1 + \ldots + \beta_n x_n} = e^{\beta x + \varepsilon_0}$$

A siker és a kudarc valószínűsége ezután a kétféle odds aránya:

$$P_x = P_x/((1-P_x)+P_x) = odds_x/(1+odds_x)$$
$$1-P_x = 1/(1+odds_x)$$

P_x egy kvantitatív mérőszáma lehet annak, hogy egy ügylet várhatóan mennyire sikeres. Ha P_x nagy, akkor várhatóan sikeres ügylet lesz, ellenkező esetben pedig kudarc.

A logisztikus regressziós modell előnyei [15]:
- A logisztikus regresszió nem követeli meg, hogy a független változók normális eloszlásúak legyenek.
- A P_x 0 és 1 közötti értéket vehet fel, így gyakorlatilag azonnal valószínűségi mutatóvá alakítható.

3.1.2 Számítógépes algoritmusokkal támogatott többváltozós regresszió

A többváltozós lineáris regresszió használatakor a számítógépes programok automatikus segítséget nyújtanak annak megállapítására, hogy mely független változók kerüljenek be a regressziós elemzésbe. A három legfontosabb ilyen algoritmus a **forward**, a **backward**, illetve a **stepwise** módszer.

Forward módszer

- A változók egyesével lépnek be.
- Elsőként belépő változó: az, amelyik a legerősebben korrelál a függő változóval.
- Majd egyesével megvizsgálja a kimaradt változókat, és azt veszi be először a modellbe, amelyik a legjobban növeli a modell magyarázóerejét. (Ennek mérőszáma a determinációs együttható, ami a korrelációs együttható négyzetével egyezik meg (képletét lásd a 3.2 (Korreláció) alfejezetben). A 0 és 1 közé eső determinációs együttható értéke megmutatja a függő változó független változók általi bejósolhatóságának milyenségét. Ha a determinációs együttható 1-hez közeli, akkor a modellben levő változók jól magyarázzák a függő változót.)
- Az algoritmus addig léptet be újabb változókat, míg azok szignifikánsan növelik a modell magyarázó erejét.

Backwards módszer

- Kezdetben minden független változó benne van a modellben.
- Egyesével megvizsgálja a bent levő változókat, és azt lépteti ki először a modellből, amelyik a legkevésbé csökkenti a modell magyarázóerejét
- Az algoritmus addig léptet ki újabb változókat, míg a még bent levő változók kiléptetése már szignifikánsan csökkentené a modell magyarázó erejét.

Stepwise módszer

- Kiinduló állapot: a forward módszerhez hasonló
- Elsőként belépő változó: az, amelyik a legerősebben korrelál a függő változóval.
- Azonban egyetlen változónak sincs bérelt helye: ha egy újonnan belépett változó miatt egy, a már modellben levő másik változó kiléptetése esetén nem csökkenne szignifikánsan a modell magyarázóereje, akkor az algoritmus ki is dobja ezt a változót.

A módszerek kritikája:
Ha a magyarázó változók nem függetlenek egymástól, hanem erősen korreláltak (multikollinearitás), akkor az egyes magyarázó változók hatását a függő változóra nem lehet szétválasztani, ezek átvehetik egymás szerepét a regressziós egyenletben. A változó-szelekció itt bemutatott módszereinek az a célja, hogy elkerüljük a multikollinearitást, azaz a változók számát csökkentse úgy, hogy a bent maradt magyarázó változók között ne legyen erős korreláció, mert ezek a független változók közölnek új, az előzőekből le nem vezethető információkat a függő változóról. A forward, backward, illetve stepwise módszerek végrehajtása során azonban fontos, hogy pontosan ismerjük az egyes magyarázó változók hatását a függő változóra, de épp a multikollinearitás miatt ezzel van a probléma, az egyes változók hatását nem lehet szétválasztani.
A másik érv ezen automatikus algoritmusok ellen, hogy csupán a statisztikai paraméterek alapján ejti ki, illetve vonja be a változókat, mely a valós tartalom rovására mehet. Ezért mindig nagy elővigyázatossággal kell kezelni az eredményeket. Erre egy analóg példát hoz Jeremy Miles és Mark Shevlin könyve, az Applying regression & correlation: a guide for students and researchers. Tegyük fel, hogy utazásra készülünk. Mi történne, ha stepwise algoritmussal pakolnánk be a bőröndünkbe a szükséges ruhadarabokat? Az eljárás először kiválasztja közülük a legfontosabbat, ami például a nadrágunk. Következő lépésként megvizsgálja, hogy ha már a nadrág bent van a bőröndben, akkor ezt alapul véve, melyik lesz a következő legfontosabb ruhadarabunk. Ekkor a stepwise algoritmus például az alsónadrágot már kevésbé fogja szükségesnek találni, mivel egy fajta nadrág már úgyis van a bőröndben. Az eljárás az alsónadrágot nem veszi be a legfontosabb ruhadarabjaink közé, így az otthon marad.

3.1.1 A változók információs értéke (IV-information value)

A változószelekció alapjául szolgálhat a változó **információs értékének** a kiszámítása is, ahol az információs érték megmutatja, hogy ha csak az adott a változó lenne a modellben, mennyire tudná előrejelezni a célváltozó értékét.
Ha a célváltozó bináris, azaz két kimenete (jók – rosszak) van, akkor egy csoportosított magyarázó változó (C) információs értéke a következőképp számítható ki:

$$IV(C) = \sum (\%Jók_i - \%Rosszak_i)*WoE_i,$$

ahol i jelöli a C változó osztályait, WoE_i pedig az egyes osztályokra kiszámított **bizonyítéksúlyt** (weight of evidence): WoE_i = Ln(%Jók / %Rosszak)

Példa:

Gyermekek száma	% Jók (nincs fizetési hátraléka)	% Rosszak (fizetési hátraléka van)	WoE	IV
nincs adat	5%	6%	-0,182	0,002
0	5%	2%	0,916	0,027
1	20%	15%	0,288	0,014
2	30%	25%	0,182	0,009
3	24%	27%	-0,118	0,004
4+	16%	25%	-0,446	0,040
Összesen	100%	100%		0,097

3.táblázat Numerikus példa az IV kiszámítására

A példában az információs érték 0,097.
Empirikus szabály, hogy ha az információs érték kisebb, mint 0,02, akkor a változó előrejelző képessége rossz; 0,02 és 0,1 között gyenge; 0,1 és 0,3 között közepes; 0,3 felett pedig erős.

A WoE-hez hasonlóan egy másik mennyiség, a – szélesebb körben használt – **lift** érték is az egyes osztályok előrejelző képességét méri a célváltozóra vonatkoztatva. A lift megmutatja, hogy egy adott változó egyes osztályaiban hányszor nagyobb a célesemény bekövetkezésének aránya, mint a teljes mintában.

Vegyünk egy példát: Tegyük fel, hogy egy bank 1.000.000 ügyfele közül 100.000 rossz adós van. Vagyis a rossz adósok aránya a megfigyelt populációban 0,1. Csoportosítsuk az ügyfeleket asszerint, hogy ki hol él (1-es Bp., 2-es megyeszékhely, 3-as egyéb város, 4-es egyéb település), és tegyük fel, hogy a következő eloszlást kapjuk:

Hol lakik?	Ügyfelek száma a csoportban	Rossz adósok száma a csoportban (az összes rossz adós hány százaléka esik az adott csoportba)	Jó adósok száma a csoportban (az összes jó adós hány százaléka esik az adott csoportba)	Lift	WoE
1	500.000	20.000 (20%)	480.000 (53,3%)	0,4	-0,98
2	250.000	10.000 (10%)	240.000 (26,7%)	0,4	-0,98
3	125.000	30.000 (30%)	95.000 (10,6%)	2,4	1,04
4	125.000	40.000 (40%)	85.000 (9,4%)	3,2	1,448

Az 1. csoportba az ügyfelek fele került. Ha a rossz adósok aránya itt is a populációban megfigyelt lenne, akkor ebbe a csoportba 500.000*0,1=50.000 rossz adós került volna. A valóságban ennél kevesebb, csak 20.000 rossz adós van a budapestiek között. Így ennek a csoportnak a liftje 20.000/50.000=0,4.
Mivel a rossz adósokra vagyunk kíváncsiak (a lift értékeket is erre számoltuk ki), a célváltozó akkor pozitív, ha egy ügyfél rossz adós, így itt a WoE = Ln(%Rossz adósok / %Jó adósok).

A lift és a WoE közötti kapcsolatot képlettel is megfogalmazhatjuk:

$$WoE = \log \frac{1-p}{\frac{1}{lift} - p}, \qquad (1)$$

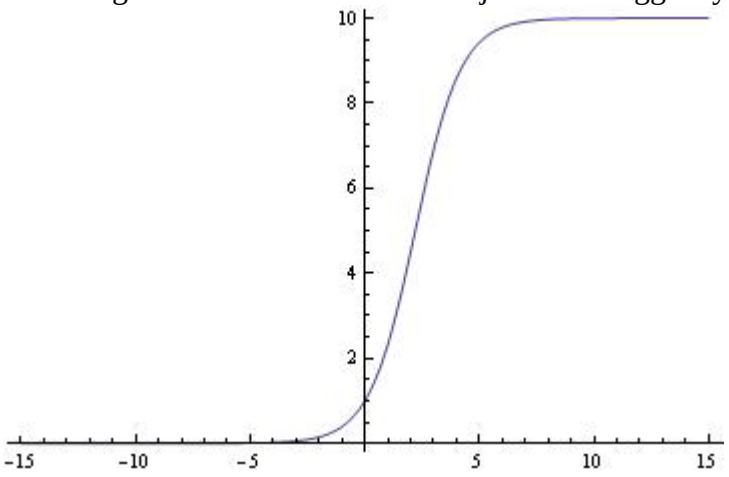
azaz:

$$lift = \frac{1}{(1-p)\,e^{-WoE} + p}, \qquad (2)$$

ahol p = (azon objektumok száma, melyekre a célváltozó értéke pozitív) / (összes objektum száma)
A példánál maradva p a rossz adósok aránya a mintában. ($0 \le p \le 1$)

Megjegyzések:

1. A lift és a WoE egymás szigorúan monoton függvényei, azaz ha egy osztály lift értéke nagyobb egy másik osztály lift értékénél, akkor ugyanez igaz lesz a WoE értékekre is. A *13. ábrán* látható görbe a lift értékeket ábrázolja a WoE függvényében. (itt p = 0,1)



13.ábra Lift értékek a WoE értékek függvényében

2. A különböző populációkban mért lift, illetve WoE értékek nem összehasonlíthatóak, mert a különböző populációkban vagy modellekben a p arány különböző lehet!
3. Az (1) matematikai kifejezés csak akkor valós, ha

$$\frac{1-p}{\frac{1}{lift} - p} > 0$$

, és

$lift \neq 0$.

Ezek a kikötések pontosan akkor teljesülnek, ha $p \neq 1$, illetve ha $lift \neq 0$.

- A lift = 0 akkor következik be, ha valamelyik osztályban nincs olyan objektum, melyre a célváltozó értéke 1. (Csak jó adósok kerültek az adott osztályba!)
- A p = 1 esetben a célváltozó értéke biztosan 1 (mindenki rossz adós). Ilyenkor a lift értékek 1-ek, a WoE-k pedig nem értelmezettek. Ekkor (1)-ben nem csak a számláló, hanem a nevező is 0, így az osztás nem értelmezhető.
- Minden más esetben a számláló és a nevező is pozitív. (A lift felső korlátja pontosan az 1/p. Tetszőlegesen közel kerülhet hozzá, de mindig igaz, hogy lift < 1/p, azaz p < 1/lift.)

4. Felhasználva, hogy

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \ldots + (-1)^{n+1}\frac{x^n}{n} + \ldots \quad \text{ha } |x|<1,$$

a WoE értékre kapjuk:

$$WoE = \log(1-p) + \log(lift) + p*lift + \left(\frac{(p*lift)^2}{2} + \frac{(p*lift)^3}{3} + \ldots + \frac{(p*lift)^n}{n} + \ldots\right)$$

Az előző pontban írtak alapján p*lift < 1, ha p < 1, így ekkor

```
WoE ~ log(1-p)+log(lift)+p*lift.
```

3.2 Korreláció

Két valószínűségi változó **korrelációja** annak mértéke, hogy az egyik változó megváltozása mennyire mutat összefüggést a másik változó megváltozásával. Fontos hangsúlyozni, hogy a korreláció lineáris kapcsolatot mér. A korrelációs együttható (r) a következő módon számítható ki:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{másképp:} \qquad r = \frac{\sum_{i=1}^{n}(x_i \cdot y_i) - n \cdot \bar{x} \cdot \bar{y}}{n \cdot s_x \cdot s_y}$$

A korrelációs együttható értékére mindig igaz, hogy $-1 \le r \le 1$. A kapcsolat szorosságára vonatkozó szokásos szóbeli értékelés:

Ha:		
	$\lvert r \rvert \le 0{,}25$	nincs, vagy nagyon gyenge a kapcsolat
	$0{,}25 < \lvert r \rvert \le 0{,}5$	gyenge,
	$0{,}5 < \lvert r \rvert \le 0{,}75$	közepes,
	$0{,}75 < \lvert r \rvert < 1$	erős,
	$\lvert r \rvert = 1$	függvényszerű, azaz ax+b a kapcsolat.

Az r előjele a kapcsolat irányára utal. Ha az egyik változó megváltozása a másik változó azonos irányú megváltozásával függ össze, akkor köztük **pozitív korreláció** van, ekkor r pozitív, és **negatív a korreláció**, ha a megváltozások ellentétes irányúak, ekkor r negatív. Ha a két változó független egymástól, azaz az egyik megváltoztatásával a másik változó értéke nem változik, akkor r=0.

Az objektumok csoportosításánál (lásd 4. fejezet) a független változókat, változócsoportokat keressük, mert ezek közölnek új, az előzőekből le nem vezethető információkat az objektumokról, míg a regressziószámításnál épp a célváltozóval összefüggő változókra van szükség, mert ezek képesek előrejelezni a célváltozó értékét.

3.3 Főkomponens analízis

A **főkomponens analízis** (PCA) a többváltozós statisztikai eljárások egyik eszköze. Több változó esetén érdemes felderíteni, hogy melyek azok a változók, amelyek a leginkább meghatározzák a vizsgált tulajdonság (a célváltozó) értékét. Ezek lesznek a fontos változóink. A nem fontos változókat elhagyva csökkenthető a dimenziószám, kényelmesebbé és letisztultabbá válhat a további elemzés, s ezen változók elhagyásával veszítjük a legkevesebb információt.

A cél formálisan: adott *k* változó (X_1, X_2, …, X_k), és keressük ezeknek olyan Z_1, Z_2, …, Z_k kombinációit (főkomponensek), amelyek nem korreláltak. A korrelálatlanság azért jó tulajdonság, mert azt jelenti, hogy az új változók az adatok különböző "dimenzióit" mérik.

Amikor főkomponens-analízist végzünk, abban bízunk, hogy a legtöbb főkomponens szórása olyan kicsi, hogy elhanyagolhatóak. Ekkor az adatokban meglévő változatosság néhány főkomponenssel jól leírható.

A főkomponens-analízissel azonban nem mindig lehetséges a nagyszámú változókat kisebb számú változókká alakítani. Sőt, ha az eredeti változók nem korrelálnak egymással, akkor egyáltalán nem lehet a változók számát csökkenteni. A legjobb eredmény akkor kapható, ha az eredeti változók erősen korrelálnak egymással - akár pozitív, akár negatív a korreláció. Ebben az esetben könnyen elképzelhető, hogy 20-30 eredeti változót adekvátan reprezentálhat 2-3 főkomponens. Ha pedig ez teljesül, akkor a fontosabb főkomponensek (melyek varianciája elég nagy) lesznek csupán érdekesek, elég ezekkel tovább dolgozni.

Fontos hangsúlyozni, hogy a főkomponensek dimenziótlan számok, emiatt fizikai jelentésük sem nyilvánvaló. Egyes esetekben hipotézis állítható fel arra vonatkozóan, hogy az első (a legnagyobb szórású) főkomponens az adatok hátterében meghúzódó, valamilyen fizikailag is azonosítható közös ok-változó. Máskor meg kell elégedni a főkomponens-analízis előnyeivel anélkül, hogy azonosítani tudnánk a háttérben ható okozati összefüggéseket.

Egy többváltozós adathalmaz elemzése során a valós adatok legtöbbször különböző klaszterekben, klaszterenként különbözően helyezkednek el. A „hagyományos” elemzési módszerek két fázisúak:

1. először az adattér pontjain klaszterezést végeznek,
2. utána főkomponens analízissel meghatározzák az egyes klaszterekhez tartozó főkomponenseket

4. Az objektumok megismerése

4.1 Előrendezés: osztályba sorolás

A mérési adatok a vizsgálat során nem valamilyen szempont szerinti rendezettségben követik egymást. Az adatok megismerése folyamán azonban ahhoz, hogy a viszonylag nagyszámú adatot át tudjuk tekinteni, érdemes őket *csoportosítani*.

A legegyszerűbb esetben, amikor adataink egy dimenziósak, azaz a számegyenesen helyezkednek el, adataink értékkészletét résztartományokra osztjuk, majd megszámoljuk, hogy egy ilyen részbe (**osztály**ba, vagy **csoport**ba) hány adat esik. Ezt a számot hívjuk az osztályhoz tartozó **gyakoriságnak**. Az osztályok, a hozzájuk tartozó gyakoriságokkal együtt alkotják a minta **gyakorisági eloszlását**.

Nagyon nem mindegy azonban, hogy hogyan jelöljük ki az osztályhatárokat. Különböző csoportosításokkal létrehozott grafikonok alapján egészen különböző, extrém esetekben ellentétes jelentésű állításokat is igazolni lehet. Erre egy tanulságos példa található Magyar Zsolt: Valószínűségszámítás és statisztika című könyvében.

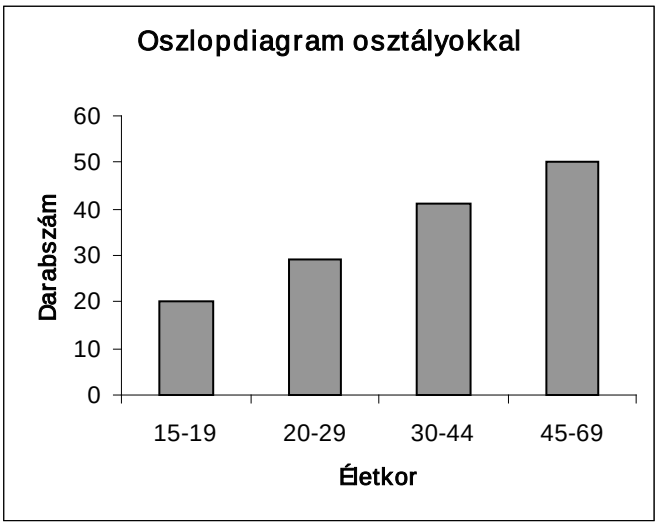
„Az adatok grafikus megjelenítésekor az adatsor ábrázolójának nagy lehetősége van manipulatív módszerek kiválasztására: pusztán a megjelenítés során sugallni tud valamit az adatsorról. Szokták mondani: statisztikai adatokkal minden be lehet bizonyítani, és az ellenkezőjét is. Nézzünk erre néhány példát!

1. példa: A politika és az életkor

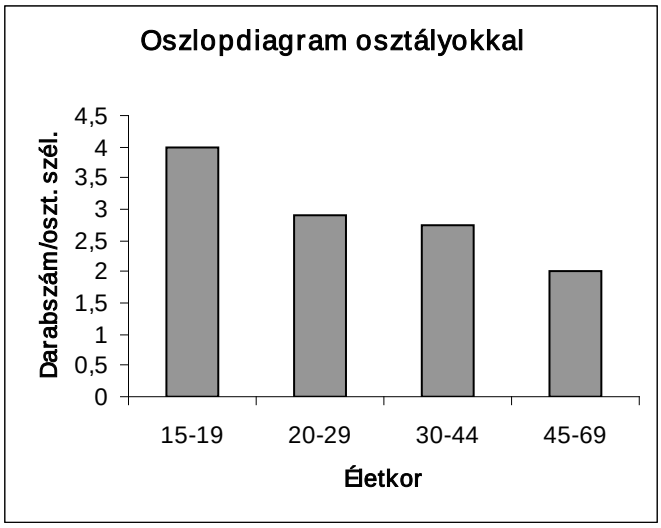
Egy vidéki városban tartott politikai rendezvényre 140 ember ment el. A résztvevők életkorát nagyság szerint közzétették (a jobb követhetőség érdekében összesítve közöljük, hogy az egyes életkorú emberekből hány volt jelen a rendezvényen, a zárójel előtti szám az életkor, a zárójelben álló szám a létszám):

15 (2), 16 (3), 17 (4), 18 (5), 19 (6), 20 (6), 21 (5), 22 (4), 23 (3), 24 (2), 25 (3), 26 (3), 27 (2), 28 (1), 29 (0), 30 (1), 31 (0), 32 (1), 33 (1), 34 (0), 35 (1), 36 (0), 37 (1), 38 (2), 39 (4), 40 (4), 41 (5), 42 (10), 43 (5), 44 (6), 45 (5), 46 (6), 47 (3), 48 (4), 49 (4), 50 (3), 51 (0), 52 (4), 53 (2), 54 (3), 55 (0), 56 (2), 57 (1), 58 (2), 59 (1), 60 (2), 61 (1), 62 (0), 63 (0), 64 (1), 65 (1), 66 (0), 67 (2), 68 (2), 69 (1)

Az első statisztikus azt az eredményt kapta, hogy a fiatalokat kevésbé érdekli a politika, és az időseket a legjobban. Az osztályokba sorolás alapján elkészítette az életkor szerinti részvételi létszám oszlopdiagramját:

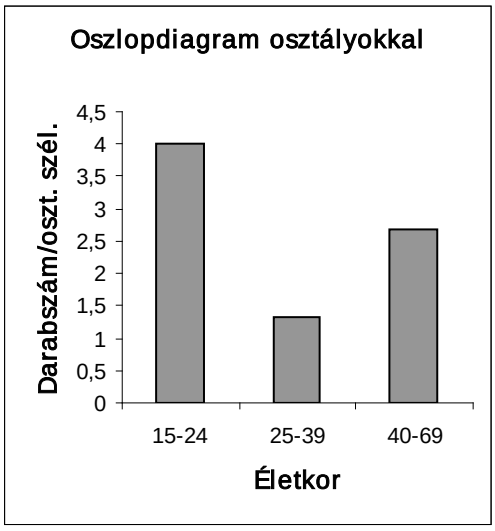


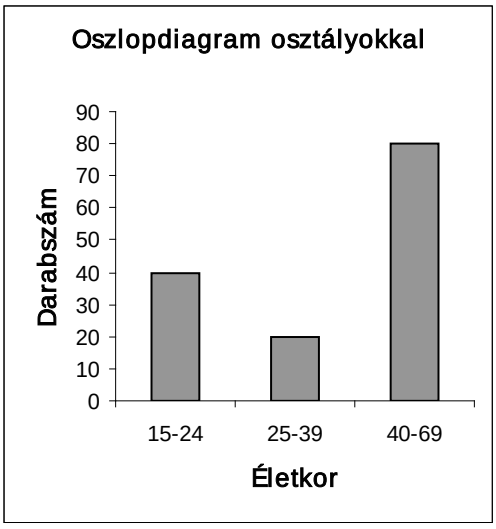
A második statisztikus fejét csóválva azt mondta: Nem jó, hiszen az osztályok nem egyforma életkori létszámról szólnak. Tehát figyelembe kell vennünk, hogy egy osztály hány évet ölel fel, és ezzel el kell osztanunk az adatokat. Így az első osztály létszámát 5-tel, a másodikét 10-zel, a harmadikét 15-tel, a negyedikét 25-tel osztjuk. A kapott értékeket ábrázoljuk oszlopdiagramon:



Ebből éppen az jött ki, hogy a fiatalokat érdekli a legjobban a politika, és az időseket legkevésbé.

A harmadik statisztikus azt mondta: Egyiknek sincs igaza, hiszen a grafikonokból származó kétféle, egymásnak ellentmondó eredmény azt mutatja, hogy rossz az osztálybasorolás. Olyan osztályokat kell keresni, ahol mindkét féle grafikonból ugyanazt az eredményt kapjuk. Mutatott is egy példát:





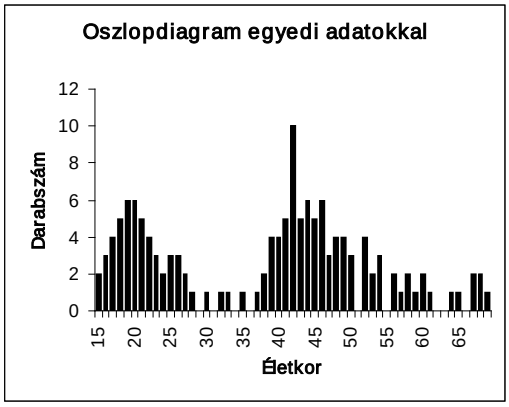
Azt ugyan nem lehet eldönteni a grafikonok alapján, hogy melyik korosztályt érdekli legjobban a politika, de az biztos, hogy a középkorúakat a legkevésbé, hiszen mindkét grafikon ezt alátámasztja.

A negyedik statisztikus azt mondta: Ne önállóan nézzük az adatokat, hanem próbáljuk meg a város lakosságához viszonyítani. El is kérte a nyilvántartásból a lakossági létszámokat, és a következőket kapta (az első esetben figyelembe vett osztálybasorolással dolgozott):

Életkor	**15-19**	**20-29**	**30-44**	**45-69**
Lakosok összes száma	359	518	735	894
Rendezvényen részt vett	20	29	41	50
Lakosok számához viszonyított %-os arány	5,57%	5,6%	5,58%	5,59%

Ebből viszont látszik, hogy érdeklődésben nincs jelentős különbség a korosztályok között.

Megjegyzésként hozzáfűzzük a feladathoz, hogy az adatok osztálybasorolás nélkül is ábrázolhatóak, érdemes elgondolkodni, hogy ebből milyen következtetés vonható le:



”

4.2 Klaszterezés [7]

Általános esetben is igaz, hogy úgy szeretnénk az adatok csoportosítását elvégezni, hogy hasonló elemek ugyanazon, míg eltérő elemek külön csoportba kerüljenek. Klaszterezéskor, azaz az elemek csoportokba sorolásakor meg kell adni, hogy miként definiáljuk az elemek hasonlóságát, és mi alapján csoportosítsunk.

A klaszterezés az adatbányászat legrégebbi és leggyakrabban alkalmazott része. Az egyik legdinamikusabban fejlődő terület, amikor vállalati ügyfeleket, vásárlókat csoportosítanak különböző jellemzőik alapján. A vállalati menedzsment végül ezen felderített közös tulajdonságaik szerint fogja kezelni az egyes csoportokat. Gyakran nem is az a kérdés, hogy az egyes elemeket melyik csoportba soroljuk, hanem az egyes csoportok közös tulajdonságainak a meghatározása.

Az elemek automatikus csoportosítására általában azért van szükség, mert az elemek túl nagy száma miatt a kézi kategorizálás túl nagy költséget jelentene. Az eljárás során azonban egy veszteséges adattömörítést végzünk, ez az ára annak, hogy a teljes adatbázist egy átláthatóbb adatcsoport adatbázissá alakítsuk.

Jónéhány klaszterező algoritmus létezik. Ezen matematikai algoritmusok alapja, hogy az elemek különbözőségének megállapítására egy függvényt, egy távolság metrikát definiál, majd miután értelmezhetővé válik két elem „távolsága”, az algoritmus megkeresi az egymáshoz „közeli” elemeket.

Érdemes meggondolni, hogy mennyire valóságszerű az elemek különbözőségét pusztán egy távolság szerint meghatározni. Ekkor az egy csoportba kerülő, azaz a hasonló elemek távolsága a saját csoportbeli elemeitől kisebb lesz, mint a más csoportban található elemektől való távolsága. Nem biztos azonban, hogy mindig ez a természetes csoportosítás. Vegyük például az alábbi síkban elhelyezkedő pontokat:

……..

……..

Ránézésre két csoport különböztethető meg, az alsó, illetve a felső pontsor. Még akkor is, ha tudjuk, hogy az alsó pontsor bal oldali szélső pontja messzebb van az alsó pontsor jobb oldali szélső pontjától, mint a felső pontsor bal oldali szélső pontjától, így valószínűleg egy klaszterező algoritmus nem a felső és az alsó pontsor szerint osztaná ketté a ponthalmazt.

Mivel az eredmények összehasonlítására nincs objektív mérték, nem létezik ideális klaszterező algoritmus. Az egyes alkalmazások jellegétől függ, hogy melyik algoritmust érdemes választani.

4.3 A klaszterező algoritmusok [7]

A klaszterező algoritmusoknak 5 fő típusa:

Partíciós módszerek: A partíciós módszerek a pontokat k diszjunkt csoportra osztják úgy, hogy minden csoportba legalább egy elem kerüljön. Egy kezdeti felosztás után megvizsgáljuk a kapott partíciókat, majd ebből kiindulva újra osztjuk őket, így javítva mindig a felosztás jóságán. A jóság megállapítására az egyes algoritmusok eltérő célfüggvényeket használnak. A folyamat akkor ér véget, mikor az algoritmus újra futtatva elemek már nem „mozognak” a partíciók között.

Hierarchikus módszerek: A hierarchikus módszerek a klaszterekből egy hierarchikus adatszerkezetet (fát) építenek fel.

Spektrál módszerek: A csoportok meghatározásához az adathalmazt reprezentáló mátrix sajátértékeit, illetve sajátvektorait használja fel.

Sűrűség alapú módszerek: a legtöbb klaszterező csak elliptikus alakú klaszterekt tud kialakítani. A sűrűség alapú módszerek ennek a hibának a kiküszöbölésére születtek meg. az alapvető ötlet az, hogy egy klasztert addig növesztenek, amíg a sűrűség a „szomszédságban” meghalad egy bizonyos korlátot. A sűrűség alapú módszereket a klaszterezés mellett a kívülálló elemek felderítésére is alkalmazzák.

Rács alapú módszerek: A rács alapú módszerek az elemeket rácspontokba képezik le, és a későbbiekben már csak ezekkel a rácspontokkal dolgoznak. Ezeknek az algoritmusoknak a gyorsaság a fő előnyük.

Ebben a fejezetben a partícionáló és a hierarchikus módszereket tárgyaljuk bővebben.

4.3.1 Partíciós módszerek

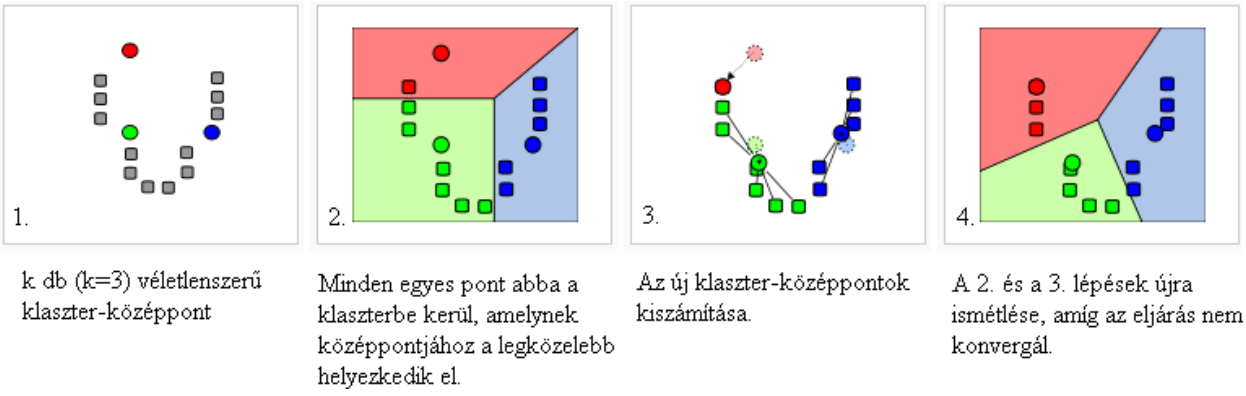
A partícionáló algoritmusokban közös, hogy a csoportok száma (k) előre adott.

Az egyik legegyszerűbb és legrégebbi klaszterező algoritmus a **k-közép** (k-means) **algoritmus**. Ennek lényege, hogy ha adott n különböző megfigyelésünk (X_1, X_2,…,X_n), ahol

minden megfigyelés egy d dimenziós valós vektor, akkor ezeket hogyan rakhatjuk k db csoportba (k<n) úgy, hogy az elemeknek a megfelelő csoportközéptől (súlyponttól) vett négyzetes hibaösszegei minimálisak legyenek. Ez az algoritmus tehát csak olyan elemek csoportosítására használható, amelyek vektortérben vannak megadva, hiszen értelmezni kell a csoportok középpontját. Nominális skálán mért változóknál ez az algoritmus nem alkalmazható.

Az algoritmus lépései a következőek (MacQueen, 1967):

- Kiválasztja a klaszterek számát (k).
- Véletlenszerűen létrehoz k számú klasztert, és meghatározza minden klaszter közepét, vagy azonnal létrehoz k véletlenszerű klaszter középpontot.
- Minden egyes pontot abba a klaszterbe sorol, amelynek középpontjához a legközelebb helyezkedik el.
- Kiszámolja az új klaszter középpontokat.
- Addig ismétli az előző két lépést (iterál), amíg valamilyen konvergencia kritérium nem teljesül (általában az, hogy a besorolás nem változik).



13. ábra A k-közép algoritmus [2]

Az algoritmus általában elég gyors, de nem biztos, hogy a megoldás a globális optimumhoz fog konvergálni. A kiinduló k db klaszter kijelölésétől függ, hogy az algoritmus a globális optimumhoz, vagy esetleg lokális optimumhoz fog vezetni.

4.3.2 Hierarchikus módszerek [2,7]

A hierarchikus klaszterező eljárásokban az adatokat hierarchikus adatszerkezetbe (fába, dendogram) rendezzük. Az adatpontok a fa leveleiben helyezkednek el. A fa minden belső pontja egy klaszternek felel meg, és ezek a klaszterek azokat a pontokat tartalmazzák, amelyek a fában alatta találhatók.

Két alapvető hierarchikus eljárás létezik: az egyik a felhalmozó, a másik a lebontó.

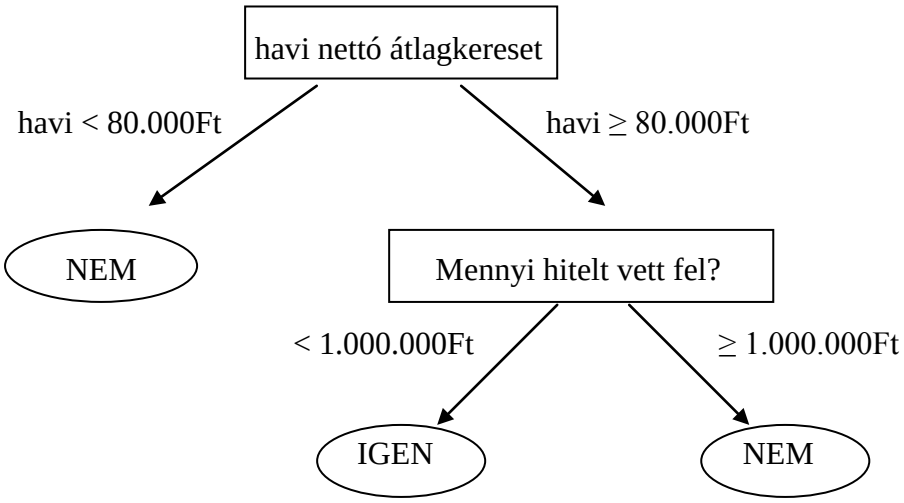
A **felhalmozó eljárásban** kezdetben minden adatelem egy klaszter, majd a legközelebbi klasztereket egyesíti az algoritmus, és a hierarchiában egy szinttel feljebb új klasztert alakít ki.

A **lebontó eljárásban** kezdetben egyetlen klaszter létezik, amelybe minden adatpont beletartozik, majd ezt tovább osztjuk. Az újabb klaszterek az előző finomításai lesznek.

Az eljárások akkor állnak meg, amikor vagy elérnek egy előre megállapított klaszterszámot, vagy a klaszterek közötti távolság egy előre megállapított mértéknél kisebbé válik.

Egy ilyen hierarchikus módszer a **döntési fát** felépítő rekurzív algoritmus. Ezt a módszert akkor érdemes használni, ha a megfigyelt objektumok (pl. ügyfelek) több dimenziósak, azaz több körülmény (pl. nettó kereset, felvett hitel nagysága,..) is befolyásolja a megfigyelt valóságot. (pl. mely ügyfelek tudják rendesen fizetni a hiteleik törlesztőrészletét; ez lesz a célváltozónk).

Ehhez a példához tartozhat az alábbi döntési fa:



14. ábra Egy egyszerű döntési fa

A döntési fa egy kifejezetten számítógépes alkalmazásra kifejlesztett osztályozó eljárás, amelynek végeredménye egy bináris jellegű osztályozó fa struktúra. A fát a gyökerétől kezdve építi fel. Az egyes döntési pontokon a minta mindig kétfelé válik egy adott változó alapján. (pl. nominális skálán mért változóknál a kategóriák szerint, míg intervallum skálán mért változóknál meghatározott értéknél válik szét a fa két ága) A fa építése során minden lépésben azt kell tehát megválaszolni, melyik változónál, és annak mely értékénél történjen a vágás. Döntési fák felépítésére többféle algoritmus is létezik. Egy ilyen például az ID3, melynek alapötlete a következő: válasszuk ki a célváltozót (alap esetben ez egy nominális vagy ordinális változó, amely meghatározza, hogy milyen osztályokba soroljuk az objektumokat). Következő lépés megkeresni azt a változót, amely a legjobban „meghatározza” a célfüggvény kimeneti értékét. Ez lesz a fa gyökere, és ezen változó lehetséges értekei lesznek az ágak. A fa ágán haladva a következő szinten újra megkeressük azt a változót, ami előzetes ismereteink alapján (hogy hol vagyunk a fában) a legjobban „meghatározza” a célfüggvény kimeneti értékét, és így tovább.

Tehát meg kell tudnunk határozni egy változóról annak **információs előnyét** vagy **leíróerejét** a célváltozóra nézve. Az ID3 a leíróerő alapján rendezi sorba a változókat, és ennek segítségével építi fel a fát.

Ahhoz, hogy definiálni tudjuk egy változó leíróerejét, még szükséges az **entrópia** fogalmának a bevezetése. Az entrópia egy rendszer rendezettségi fokát jellemzi, és a következőképp számolhatjuk ki:

$$Entrópia\ (S) = -\sum_{v_i \in R_S} p_{v_i} log_2(p_{v_i}),$$

ahol S jelöli a példahalmazt, azaz a tanuló adathalmazunkat, R_S az S halmaz véges értékkészletét, p_{vi} pedig a v_i érték előfordulási valószínűségét.

Az *entrópia* függvény minimális értéke 0, amit akkor kapunk, ha minden objektum a célváltozó alapján egy adott csoportba került. Ekkor a bizonytalanságunk nulla. Az entrópia akkor a legnagyobb ha az összes érték (csoport) előfordulási valószínűsége egyenlő. Ekkor a legnagyobb a bizonytalanságunk is, hiszen bármelyik csoportba ugyanakkora valószínűséggel kerülhet az objektum.

Az *Entrópia(S)* tehát az adott S halmaz „összevisszaságát” adja meg, és kíváncsiak vagyunk arra, hogy egy-egy változó szerinti rendezés mennyire csökkenti ezt az értéket. Ezt mutatja meg a leíróerő:

$$Leíróerő\ (S, A) = Entrópia(S) - \sum_{v \in R_A} \frac{|S_v|}{|S|} Entrópia\ (S_v),$$

ahol R_A az A változó értékkészletét, S_v pedig S azon részhalmazát jelöli, melyre az A változó értéke éppen v.

Az ID3 algoritmusa tehát a leíróerő alapján fogja eldönteni, hogy egy adott szinten melyik változó határozza meg „legjobban” a célt, és ez alapján fogja majd felépíteni a fát.

Ez a faépítés azonban csak akkor működik, ha változóink nem folytonos skálán mértek. Általánosabb esetben azonban a magyarázó változóink lehetnek folytonos értékűek, ekkor ezeket helyettesítjük egy dinamikusan megalkotott bináris változóval.

Például az A folytonos értékű változót egy A_x bináris változóval helyettesíthetjük. Vegyük példaként a következő táblázattal leírható esetet:

Ügyfél kora:	25	30	32	38	41	47
Tudta-e fizetni a hitele törlesztőrészleteit:	Nem	Nem	Igen	Igen	Igen	Nem

Egy olyan c értéket keresünk, amely a legnagyobb *leíróerő*vel bír. A példák kimenetei alapján megnézzük, hogy hol vannak azok a határok, ahol átlépünk egy másik csoportba. A példában ez a két határ a 32+38/2=35, valamint a 41+47/2=44 lesz. Ezek után az *Ügyfél kora* helyett két új változónk lesz: az *Ügyfél kora* $_{<35}$ és a *Ügyfél kora* $_{>44}$. Ezeket az új bináris változókat használjuk, és megnézzük, melyiknek a legnagyobb a *leíróereje*.

Más módszerek is léteznek, amelyek több bináris változót rendelnek egy folytonos attribútumhoz.

A döntési fa osztályozó algoritmus továbbfejlesztése a véletlen erdő módszer, melynek fő gondolatmenete: több döntési fát készítünk a véletlen mintavételezéssel létrehozott tanuló adatállományokra, majd megvizsgáljuk, hogy melyik döntési fa melyik osztályba sorolja az adott objektumot. Végül abba az osztályba kerül, amelyiket a legtöbb döntési fa javasolt.

5. A statisztikai modellezés és a modellek összehasonlítása

Minden statisztikai modellezési feladat a statisztikai sokaság és maga a feladat megismerésével kezdődik. Ilyen feladat lehet például, ha egy banki adatállományra alapozva kell előrejelezni, hogy az egyes új lakáshitel igénylők várhatóan tudják-e majd törleszteni a részleteket. Különböző, már bemutatott és egyéb statisztikai módszerekkel lehetőség nyílik a bankoknál felhalmozódott ügyfél és viselkedési adatokból olyan hasznos és rejtett összefüggések felfedésére, amelyek felhasználhatóak a leendő ügyfelek hitelvisszafizetési kockázatának megbecslésére, és statisztikai modellek építésére.
A modellezés során azonban több lehetséges megoldást is kaphatunk az adott feladatra, attól függően, hogy milyen változókat veszünk be a modellbe, illetve milyen statisztikai módszert használunk a modellépítésre. Cél ezek közül a leghatékonyabb modell kiválasztása.

Ahhoz, hogy az eredmények ellenőrizhetőek legyenek, a modellezés előtt véletlenszerűen partícionálni kell az adathalmazt egy tanuló (training) és egy ellenőrző (validation) állományra. A felosztás nem kell, hogy fele-fele arányú legyen, az ellenőrző állományra elegendő lehet az adatok 10%-a is, mert minél több adat kerül az ellenőrző állományba, annál kevesebb marad magában a tanuló állományban, amiből a modellt építjük. A felosztás célja, hogy a tanuló állományon megfigyelt összefüggéseket később tesztelni lehessen az ellenőrző állományon. Így lehet a létrehozott modellt kalibrálni. Ez azért fontos, mert előfordulhat, hogy a modell beállítása olyan jól „sikerül”, hogy bár a tanuló állományra tökéletes modellt hoztunk létre, de az annak az állománynak az egyedi sajátosságait tükrözi, ami eltér a sokaságétól. Azaz túl speciális lett a modell, ezt nevezik túlillesztésnek. Az ellenőrző állománnyal tehát a modell hatékonyságát tesztelhetjük, hogy hogyan teljesít az új adatokon. Ez alapján lehet majd kiválasztani a felépített modellek közül a legjobbat.

5.1 Tévesztési mátrix

Azoknál a feladatoknál, ahol cél az objektumok osztályba sorolása, a modellek összehasonlítására elkészíthető a tévesztési mátrix, ami a valóság és a modellezéssel előrejelzett eredmények kapcsolatát mutatja be.

		Aktuális feltétel	
		A feltétel teljesül	A feltétel nem teljesül
Teszt eredmény	Pozitív	A feltétel teljesül + pozitív teszt = TP (True Positives)	A feltétel nem teljesül + pozitív teszt = FP (False Positives)
	Negatív	A feltétel teljesül + negatív teszt = FN (False Negatives)	A feltétel nem teljesül + negatív teszt = TN (True Negatives)

15. ábra Általános tévesztési mátrix

Példa: banki ügyfelek hitelei várhatóan be fognak-e dőlni		Aktuális helyzet	
		Pozitív (valóban bedőlt)	Negatív (nem dőlt be)
Teszt eredmény	A teszt szerint pozitív (várhatóan bedől)	TP	FP
	A teszt szerint negatív (várhatóan nem dől be)	FN	TN

16. ábra Tévesztési mátrix banki példán

Akiknél az aktuális feltétel valóban teljesül: **P** = TP + FN, és azok száma, akiknél valóban nem: **N** = TN + FP.
Többféle széles körben használt mérőszám képezhető ezek alapján a besorolási pontosságra:
TPR (True Positive Rate) vagy **Sensitivity** = TP/(TP+FN) = TP/P: a pozitívak helyesen felismert aránya
TNR (True Negative Rate) vagy **Specificity** = TN/(FP+TN) = TN/N: a negatívak helyesen felismert aránya
FPR (False Positive Rate): FP/N
FNR (False Negative Rate): FN/P
YR: (TP+FP)/(P+N)
Accuracy = (TP+TN)/(TP+TN+FP+FN): a helyesen felismert adatok aránya
Precision = TP/(TP+FP): a pozitívként felismertek hanyad része volt valóban pozitív

A besorolásnál kétféle hibát követhetünk el:
FP: elsőfajú hiba; FN: másodfajú hiba
Az adott feladattól függ, hogy mikor melyik hibát szeretnénk minimalizálni. Például egy orvosi tesztnél fontos, hogy minden beteget ki tudjunk szűrni. Nagyobb hiba, ha egy betegről nem derül ki, hogy kóros, mint ha egy egészségesnél is pozitívat jelez a teszt. Így az ilyen feladatoknál a másodfajú hiba minimalizálása a cél. Egy bírósági eljárás során viszont súlyosabb vételnek tekinti a társadalom, ha egy ártatlan ember kerül börtönbe, itt a bíró az elsőfajú hiba minimalizálására fog törekedni. De előfordulhatnak olyan esetek is, ahol pl. min(FP+FN) a cél.

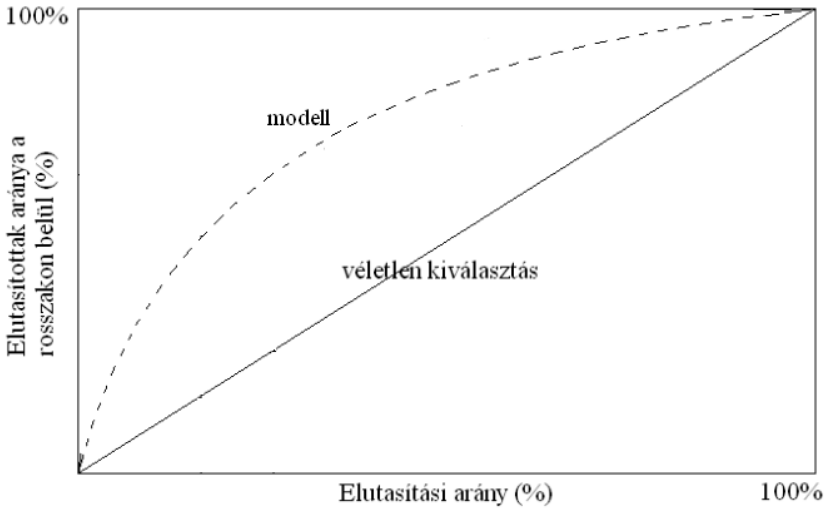
Költség-elemzés:
Már láttuk, különböző feladatoknál eltérően ítélhetjük meg az egyes hibák súlyát, illetve az első és másodfajú hibák költségei igen különbözhetnek egy feladaton belül is. Ahhoz, hogy az egyes osztályozó eljárásaink eredményei összehasonlíthatóak legyenek, a hibás osztályozáshoz hozzárendelhetjük az egyes hibák költségét. Ezt tartalmazza a költségmátrix. Cél az összes hibaköltség minimalizálása.

Költségmátrix		Aktuális helyzet		
		1-es osztály	2-es osztály	3-as osztály
Teszt eredmény	1-es osztály	0	2	5
	2-es osztály	4	0	1
	3-as osztály	2	3	0

Az összes hibaköltség:
$\sum_{i=1}^{n}\sum_{j=1}^{n} E_{ij}C_{ij}$, ahol E_{ij} azon elemek száma, amelyek valójában a j-edik osztályba tartoznak, de hibásan az i-edikbe soroltuk őket; C_{ij} pedig ennek a rossz besorolásnak a költsége a költségmátrixból.

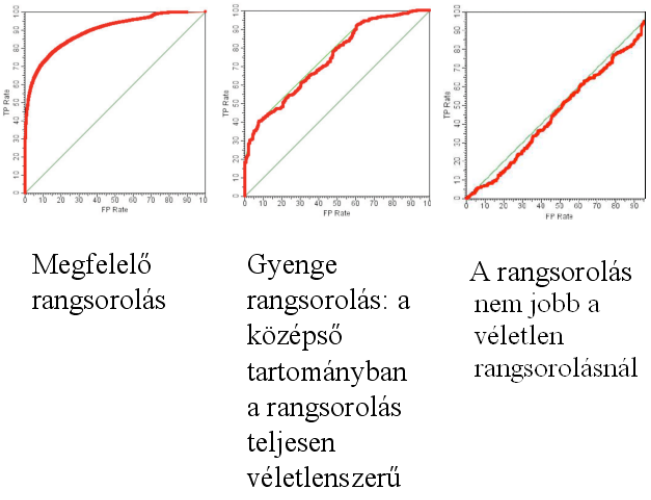
5.2 ROC chart
A banki példánál maradva, valószerű, hogy a banki ügyfelei között időben változhat a valóban jó és valóban rossz adósok számának aránya. Ha ettől az aránytól függetlenül szeretnénk mérni a modellünk teljesítményét, azaz amikor valakit jó adósnak jósolt, akkor milyen arányban volt jó, illetve rossz az előrejelzés, akkor érdemes a ROC chart grafikont használni. Modellünk teljesítményét egy görbe írja le, melynek pontjai megmutatják, hogy a jó ügyfelek valahány százalékának a kizárásával a rossz ügyfelek hány százalékát zárjuk ki?

Formálisan: a ROC chart egy olyan görbe, melyre:
x = FPR(t) = 1-Specifity(t), azaz 1- a negatívak helyesen felismert aránya,
y = TPR(t) = Sensitivity(t), azaz a pozitívak helyesen felismert aránya.



17. ábra ROC chart

Gyakran a modellhez tartozó görbe alatti területtel mérik a modellek teljesítményét. Így a különböző modellek teljesítménye összemérhetővé válik.



18. ábra Különböző modellek teljesítményének értékelése a ROC chart segítségével

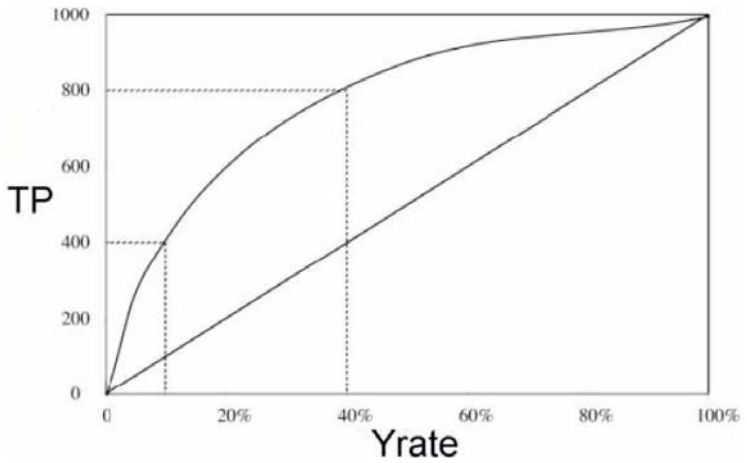
Kapcsolat a Lorenz görbe és a ROC chart között:
Mindkét grafikonon az abcissza (x) tengely mutatja az (1 – a negatív pontok) kumulatív eloszlását (a példákban: Lorenz görbe – (1 – a legkevesebb jövedelemmel rendelkező háztartások) = a legtöbb jövedelemmel rendelkező háztartások x%-a; ROC chart – 1 – a negatívok helyesen felismert aránya). De míg a Lorenz görbe ordinátáján (y tengely) az összes megfigyelés kumulatív eloszlása látszik, addig a ROC chartnál az ordináta a pozitív pontok kumulatív eloszlását mutatja. Ha kevés a negatív pont, a két görbe alakja nagyon hasonló lesz, mert ekkor a pozitív pontok kumulatív eloszlása közelít az összes megfigyelés kumulatív eloszlásához.

5.3 Lift chart

A Lift chart grafikont általában marketing kampányok döntés-előkészítésénél használják. Vegyünk egy példát. Egy áruházlánc új termék bevezetését tervezi, és hogy felhívja erre a háztartások figyelmét, marketing leveleket küld nekik. Az áruházlánc marketingosztálya ismeri a háztartások különböző adatait. Minden kiküldött levélre ki kell fizetni a postaköltséget, de ez mégis kifizetődő az áruháznak, ha megveszik az új termékét. Az áruházlánc ezért minimalizálni szeretné a kiküldött leveleket, de ugyanakkor maximalizálni is az eladott termékeket.
Ilyenkor a valóban lehetséges vásárlók száma (P) nem ismert, ezért nem tudjuk kiszámítani TPR-t, és így a ROC chart nem ábrázolható. Ilyenkor lehet hasznos a lift chart, ami segítségével kiválaszthatóak azok a legígéretesebb háztartások, ahova érdemes marketing levelet küldeni.

Formálisan: a lift chart egy olyan görbe, melyre: x = YR(t), y = TP(t).

A ROC charttól görbéitől eltérően a lift chart görbéi nem függetlenek a P/N aránytól.



19. ábra Lift chart

A ROC charthoz hasonlóan a görbe alatti terület itt is használható a különböző modellek teljesítményének az összehasonlítására.

Hivatkozások:

[1] Regionális elemzési módszerek, ELTE Regionális Földrajzi Tanszék (2005)
2. Adatkezelés, statisztikai és számítástechnikai alapok
Forrás: http://geogr.elte.hu/REF/REF_Kiadvanyok/REF_RTT_11/

[2] Wikipedia www.en.wikipedia.org és www.hu.wikipedia.org

[3] Rákóczi István Péter: Az informatika matematikája (51. modul-tankönyv)
Forrás: w3.szikszi.hu/~rip/statisz.pdf

[4] Elek István (2005): Az adatbányászat osztályozási eljárásainak alkalmazása a vektoros térinformatikában, Geodézia és Informatika, 57. évf. 11. szám pp.12-17

[5] Münnich Á., Nagy Á., Abari K.: Többváltozós statisztika pszichológus hallgatók számára (2006), Debrecen: Bölcsész Konzorcium ISBN 963 9704 04 0
Forrás: http://psycho.unideb.hu/statisztika

[6] Magyar Zsolt: Valószínűségszámítás és statisztika
Forrás: http://sulinet.hu/matek/magyar_zs/vsz_1.doc

[7] Dr. Bodon Ferenc (2009): Adatbányászati algoritmusok
Forrás: http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf

[9] Fazekas I. (szerk.) (1997): Bevezetés a matematikai statisztikába. Egyetemi jegyzet. Kossuth Lajos Tudományegyetem, Debrecen (1997)

[10] Lukács Ottó (2002): Matematikai statisztika (Bolyai-könyvek sorozat), Műszaki Könyvkiadó ISBN 963-16-3036-6

[11] Miha Vuk, Tomaˇz Curk (2006): ROC Curve, Lift Chart and Calibration Plot Metodoloˇski zvezki, 2006, Vol. 3, No. 1, pp. 89-108

[12] Szűcs Imre (2006): Adatbányászati módszerek alkalmazása a pénzügyi szektorban, ELTE, szakdolgozat, Budapest
Forrás: miau.gau.hu/miau/97/szucs_isgt.doc

[13] Buday Balázs (1999): Induktív tanulás – az ID3 algoritmus, Hallgatói Esszé (ELTE)
Forrás: people.inf.elte.hu/saci/MI/esszek/**ID3-algoritmus**.doc

[14] Info-Datax Kft. Tanulmánya (2006) Módszertani elemzés a nemfizetési valószínűség modellezéséhez
Forrás: http:// www.pszaf.hu/data/cms65481/pszafhu_palyamunka_nemfizetes.pdf

[15] Többváltozós statisztikai szeparáció
Forrás: ME-GTK Pénzügyi Tanszék honlapja
http://193.6.3.238/gtk/ui/uipz/hallgatoi/tobbvaltozos_statisztikai_szeparacio.pdf

[16] Hajdú Ottó: *Segédlet a „Bevezetés az ökonometriába” c. tárgy oktatásához*
Forrás: http://portal.uni-corvinus.hu/fileadmin/user_upload/hu/tanszekek/kozgazdasagtudomanyi/tsz-statisztika/tantargyak/kvanti_op/LogitKieg.pdf