Leíró statisztika

A minket körülvevő világban számos olyan folyamat található, amely nem egy előre meghatározott matematikai zárt egyenlet szerint írható le. Ezekre általában a véletlentől függő tényezők hatnak, ami megnehezíti az ilyen folyamatok megismerését. Ilyenek például a termelési folyamatok. Képtelenség teljes mértékben megjósolni, hogy egy adott napon egy gép hány selejtes alkatrészt gyárt akkor is, ha jól ismerjük a gép paramétereit és rendelkezésre állnak korábbi termelési adatok. Könnyű hasonló példák találni a gazdasági, műszaki, orvosi és társadalomtudományban.

Ha egy ilyen folyamatot szeretnénk elfogadható szinten megismerni, nem marad más választás, mint információt szerezni. Olyan folyamatok képezhetik vizsgálódásunk tárgyát, amelyek tömeges jelenségeket generálnak, és ezekből adatokat gyűjthetünk. Segítségükkel további adatokat származtathatunk, elemzésükből levonhatjuk a megfelelő következtetéseket.

A statisztika tárgya

A *statisztika* a valóság tömör, számszerű jellemzésére szolgáló tudományos módszertan és gyakorlati tevékenység.

A statisztika kifejezés a latin status (állam, állapot), illetve az olasz statista (köztisztviselő, politikus) szavakra vezethető vissza; elsőként a német Gottfried Achenwall használta 1749-ben, az állam tevékenységével kapcsolatos adatok elemzésére.

A statisztikai sokaság

Fontos, hogy az adatokat mindig egy adott folyamatból gyűjtsük, illetve a különböző folyamatokhoz tartozó adatgyűjtéseket ne keverjük össze. Természetesen, az adataink mindig a megfigyelés tárgyát képező folyamatban található egyedek egy előre meghatározott tulajdonságcsoportjához kell, hogy tartozzanak. Az ilyen adatok összességet *statisztikai sokaságnak* nevezzük.

statisztikai ismérv

Az adatgyűjtésből nyert adatok tehát mindig a statisztika sokaság elemei. A tulajdonságcsoport minden egyes elemét *statisztika ismérvnek* nevezzük. Az adatgyűjtési folyamatot minden esetben megelőzi egy olyan elméleti alapokon álló terv kidolgozása, amely igazolja az adatgyűjtésen alapuló következtetések helyességet.

Gyakran előfordul, hogy egy statisztikai sokaság elemszáma nagyon nagy, szinte korlátlan, így a sokaság teljes megismerése túl sok időt vagy pénzt követel. Sőt, számos folyamat a jövőben is folytatódik és így az elméletileg kiterjesztett sokaság jelentős részét nem ismerhetjük meg a megfigyelés pillanatában, többek között nem tudhatjuk mekkora lesz ez az elemszám. Ilyen esettel találkozunk amikor egy alkatrészt gyártó gép hatékonyságára egynapos megfigyelés alapján következtetünk. A előző esetekben az egyedek csak egy szűkebb csoportját figyelhetjük meg.

A leíró statisztika tárgya

A *leíró statisztika* a megfigyelt adatok összegzésével, elemzésével, összehasonlításával, ábrázolásával, tömör számszerű jellemzésével foglalkozik, de nem készít belőlük következtetéseket a teljes sokaságra. Ezzel a *matematikai statisztika* foglakozik, ami valószínűségszámítási modellekkel dolgozik a megfelelő megbízhatósági szintek elérésére.

1. Adatelemzés

Leszármaztatott adat

Az adatgyűjtésből nyert adatokból a folyamat által generált elemeihez bizonyos tulajdonságait, ismérveit rendeltük. Ezek az ú.n. *alapadatok* nem feltétlenül számszerű adatok, pl. ha megnézzük milyen hallgatók tanulnak egy adott kurzuson és a nemüket, mint adat, jegyezzük fel. Ezekből további adatok származtathatók. *Leszármaztatott adat* egy olyan valós függvény eredménye, amely értelmezési tartománya a gyűjtött adatok halmaza vagy ennek részhalmaza. Ilyen függvény lehet a darabszám, összeg, átlag, maximum, minimum, de olyan egyszerűbb művelet eredménye is, mint az osztás. Ilyen függvényekkel az Excel szép számban rendelkezik.

Az adatelemzés célja az adatgyűjtéssel kapott adatok leírása bizonyos leszármaztatott adatok segítségével. Az elemzések között találjuk az adatok csoportosítását, a viszonyszámokat, és különböző átlagolási módszereket.

1.1. Csoportosítás

Induljunk ki a következő táblázatból.

1. táblázat. A Nyíregyházi Főiskola PTI szakon Analízis II előadására jelentkezett hallgatók nemek szerinti megoszlása a 2010/2011-es tanévben

nem	fő
fiú	70
lány	5
Összesen	75

A fenti táblázat elkészítéséhez összegyűjtöttünk az említett kurzusra jelentkező hallgatók névsorát. Egyetlen ismérvet vettük figyelembe, a személyek nemét, amelyek az alapadatainkat alkotják. Ezekből olyan leszármaztatott adatokat képeztünk, mint az egyes nemek darabszámát és az összes adat darabszámát. Így tömör, világosabb képet kapunk a kurzusra jelentkezett hallgatók nemek szerinti megoszlásáról.

csoportosítás

Az előző feladat egy nagyon egyszerű példát mutat adatok csoportosítására. A *csoportosítás* lényege az alapadatok felosztása egymástól világosan elkülönülő osztályokra egy vagy több ismérv figyelembevételével és a különböző osztályokhoz tartozó alapadatok megszámlálása (gyakorisága). Ezzel egy *csoportosított adatsort* kapunk. Sok esetben egy jelentős mennyiségű alapadatokkal rendelkező adatgyűjtés legegyszerűbb megadási módja egy csoportosított adatsor. A csoportosítások másik előnye az, hogy megkönnyíthetik a további leszármaztatott adatok kiszámítását.

Az Excelben a `DARABTELI` függvénnyel tudunk a legegyszerűbben csoportosítást készíteni. Lássuk a következő példát!

3	2	2	2	3	3	1	1	1	1	1
1	2	3	4	1	2	3	2	2	3	2
3	2	2	2	1	2	5	2	2	1	2
3	3	4	2	2	2	4	2	4	2	4
1	2	2	2	1	2	3	3	4	2	1
2	1	1	2	3	3	3				

2. táblázat. A Nyíregyházi Főiskola PTI szakon Analízis II vizsga végeredményei a 2009/2010-es tanévben

Nyissa meg az analizis-vizsga.xls fájlt! Az `A` oszlop tartalmazza az összes a vizsgán elért eredményt hallgatókra lebontva. Az érvényes szabályok értelmében ez legfeljebb 3 vizsga alkalom lehet hallgatónként, melynek a legutolsó eredménye a döntő. Ezek alkotják a 2. táblázat adatait és az adatgyűjtés alapadatainak tekintjük őket. Ezek a `B` oszlopban találhatók.

A következőkben megmutatjuk, hogy milyen lépéseket érdemes végrehajtani egy csoportosított adatsor elkészítésekor. Közben olyan jelöléseket vezetünk be, melyek segítenek rövidebben leírni az Excelben végrehajtott lépéseket.

a) *Nevezzük el x-sszel az adattartományt!* Az Excel-ben név megadásával tudunk a legegyszerűbben abszolút módon hivatkozni egy tartományra. Ezt a tartomány kijelölés után a Beszúrás → Név → Név megadása menüvel érhetjük el, vagy közvetlenül a név beírásával a név mezőben. A név megadására a $\doteq$ jelet használjuk, ez esetünkben

   `x` $\doteq$ `B2:B63`

   A név megadása nem egy kötelező lépés, hiszen közvetlenül hivatkozhatunk egy tartományra a szokásos módon, de áttekinthetővé teszi a munkát.

b) *Határozzuk meg az osztályokat a csoportképző ismérv szerint!* Tudjuk, hogy egy vizsga 1-től 5-ig terjedő érdemjeggyel értékelhető. Írjuk be ezeket a számokat a `D2:D6` tartományra! Ha egyszerű értékeket, adatokat írunk

be egy cellába, akkor az = jelet használjuk. Esetünkben

`D2` = 1 `D3` = 2 ... `D6` = 5

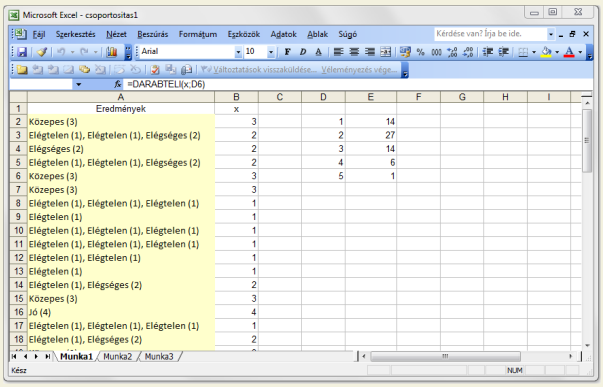
c) *Számoljuk ki az osztályok gyakoriságát!* Ez a `DARABTELI` függvénnyel történik. Először az `E2` cella szerkesztőlécébe írjuk az `=DARABTELI(x;D2)` utasítást! Munkánkban erre a ← jelet használjuk, ez esetünkben

`E2` ← `DARABTELI(x;D2)`

Végül az `E2` cellát másoljuk a hiányzó `E3:E6` tartományra! A másolás és beillesztés folyamatára a ⇝ jelet használjuk, tehát

`E2` ⇝ `E3:E6`

A végeredmény a következő ábrán látható.



1. ábra. A csoportosítás eredménye az analízis vizsgák esetén

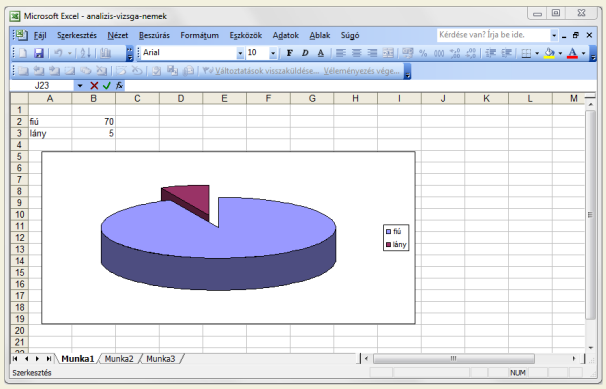
1.2. Grafikus ábrázolás

Grafikus ábrázolást akkor alkalmazunk, amikor szemléltetni szeretnénk az alap- vagy leszármaztatott adatokat. Az adatok közötti arányok bemutatásával áttekinthetővé válik a folyamat szerkezete és szabályrendszere. Ehhez nagyon fontos a megfelelő diagram kiválasztása.

Az Excel számos diagramtípust kínál. Ezek között találunk oszlop-, sáv-, pont- és kördiagramot. Elkészítésük nagyon egyszerű. Például, ha szemléltetni szeretnénk a 1. táblázat adatait, akkor érdemes egy kördiagramot készíteni. Ehhez hozzuk létre az analizis-vizsga-nemek.xls fájlt, beírjuk az

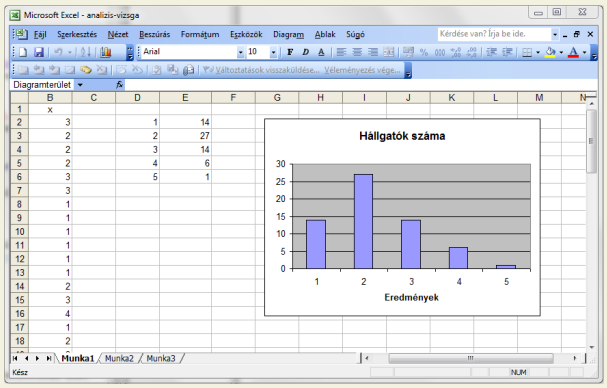
`A2` = fiú `A3` = lány `B2` = 70 `B3` = 5

adatokat és a Beszúrás → Diagram menüvel kiválasztjuk a megfelelő diagramot és elkészítjük. Az eredmény a következő ábrán látható.



2. ábra. A Nyíregyházi Főiskola PTI szakának Analízis II előadására jelentkezett hallgatók nemek szerinti megoszlása a 2010/2011-es tanévben

A 2. táblázat esetében oszlopdiagramot készítünk a analizis-vizsga.xls fájlban. Az eredmény a következő ábrán látható.



3. ábra. A Nyíregyházi Főiskola PTI szakosai Analízis II vizsgájának végeredményei a 2009/2010-es tanévben

1.3. Összehasonlítás

Az *összehasonlítás* két vagy több alap- vagy leszármaztatott adat egymáshoz való viszonyítása. A leggyakoribb összehasonlítás két adat hányadosa vagy különbsége, melynek eredménye egy újabb leszármaztatott adat. Adatok összehasonlításával tudjuk leírni a folyamat tulajdonságait.

összehasonlítás

Az összehasonlítás speciális esetei a *viszonyszámok*. Viszonyszám alatt két alap- vagy leszármaztatott adat hányadosát értjük. A viszonyszámokat szokás a következő módon megkülönböztetni:

viszonyszámok

megoszlási viszonyszám: az egyes adatokat az összes alapadat számához viszonyítjuk,

koordinációs viszonyszám: két azonos fajta adatot egymáshoz viszonyítunk,

intenzitási viszonyszám: két különböző fajta adatot egymáshoz viszonyítunk,

dinamikus viszonyszám: két időszakhoz tartozó adatot egymáshoz viszonyítunk.

Az Excelben könnyedén tudunk ilyen összehasonlításokat elvégezni. A szöveges leírások elhelyezhetők a munkalapon az `ÖSSZEFŰZ` és `SZÖVEG` függvények segítségével. Például, a 1. táblázat adataival számítsuk ki a fiúk arányát, a lányok arányát (megoszlási viszonyszámok), valamint az, hogy egy fiúra hány lány jut (koordinációs viszonyszám). Ez megtekinthető az analizis-vizsga-nemek.xls fájl „Viszonyszámok” nevű munkalapján. A lépések a következők:

`A2` = fiú `A3` = lány `B2` = 70 `B3` = 5

`A5` = Összesen `B5` ← SZUM(B2:B3) `n` ≐ `B5`

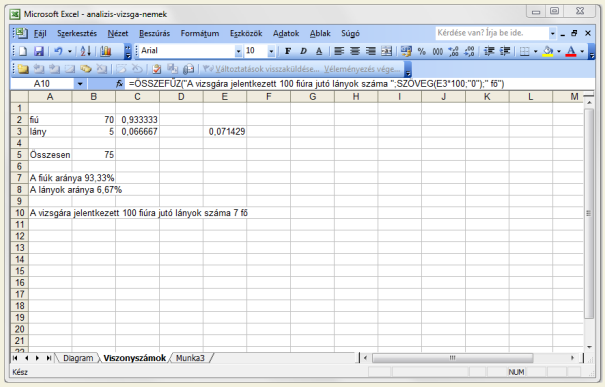
`C2` ← `B2/n` `C2` ⇝ `C3` `E3` ← `B3/B2`

`A7` ← `ÖSSZEFŰZ("A fiúk aránya ";SZÖVEG(C2*100;"0,00");"%")`

`A8` ← `ÖSSZEFŰZ("A lányok aránya ";SZÖVEG(C3*100;"0,00");"%")`

`A8` ← `ÖSSZEFŰZ("A vizsgára jelentkezett 100 fiúra jutó lányok száma ";SZÖVEG(E3*100;"0");" fő")`

Az eredmény a következő ábrán látható.



4. ábra. A Nyíregyházi Főiskola PTI szakának Analízis II előadására jelentkezett hallgatók nemek szerinti leírása a 2010/2011-es tanévben

1.4. Átlagok

Az átlagok vagy más néven közepek azonos fajta számadatokon értelmezett függvény számszerű eredménye, amivel tömören jellemezhetjük ezeket az adatokat. Négyféle átlaggal foglalkozunk. Legyen

$$x_1,\ x_2,\ x_3,\ \ldots,\ x_n$$

n darab valós szám. Ezeknek a számoknak számtani, mértani, harmonikus és négyzetes átlagát a következő módon értelmezzük és jelöljük:

Számtani átlag:

$$\overline{x} := \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Mértani átlag:

$$\overline{x}_g := \sqrt[n]{\prod_{k=1}^{n} x_k} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \cdots \cdot x_n}$$

Harmonikus átlag:

$$\overline{x}_h := \frac{n}{\sum\limits_{k=1}^{n} \frac{1}{x_k}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}}$$

Négyzetes átlag:

$$\overline{x}_q := \sqrt{\frac{1}{n}\sum_{k=1}^{n} x_k^2} = \sqrt{\frac{x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2}{n}}$$

Az Excelben azonnal ki lehet számolni a fenti átlagokat beépített függvények segítségével. Számoljuk ki az átlagokat a 2. táblázatban található értékek esetén! A lépések a következők:

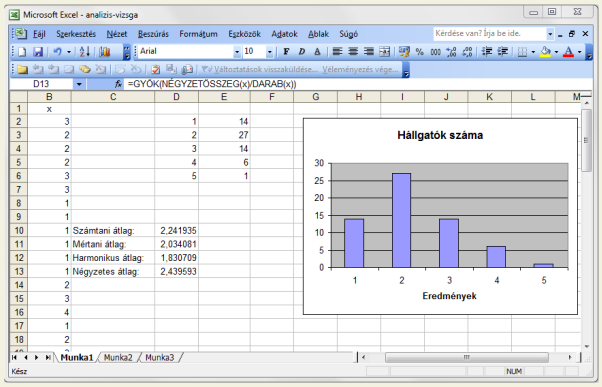
`C10` = Számtani átlag: `D10 ← ÁTLAG(x)`

`C11` = Mértani átlag: `D11 ← MÉRTANI.KÖZÉP(x)`

`C12` = Harmonikus átlag: `D12 ← HARM.KÖZÉP(x)`

`C13` = Négyzetes átlag: `D13 ← GYÖK(NÉGYZETÖSSZEG(x)/DARAB(x))`

Az eredmény az analizis-vizsga.xls fájlban és a következő ábrán látható.



5. ábra. A Nyíregyházi Főiskola PTI szakos hallgatói Analízis II vizsgája végeredményeinek átlaga a 2009/2010-es tanévben

Csoportosított adatsorok esetén az átlagoknak egy ú.n. *súlyozott* számítási formáját alkalmazzuk. Tegyük fel, hogy a számadatok között vannak egyforma értékű adatok és elkészítjük a csoportosításukat. Jelölje

$$x_1,\ x_2,\ x_3,\ \ldots,\ x_N$$

a különböző számadatokat, és

$$f_1,\ f_2,\ f_3,\ \ldots,\ f_N$$

a számok gyakoriságát. Előfordul, hogy a gyakoriságok helyett megoszlásukat, ú.n. *relatív gyakoriságukat* ismerjük, azaz a

relatív gyakoriság

$$g_i = \frac{f_i}{\sum_{k=1}^{N} f_k} \qquad (i = 1,\ 2,\ \ldots,\ N)$$

számokat. Ekkor a következő módon számítjuk ki az átlagokat:

Számtani átlag:

$$\overline{x} = \frac{\sum_{k=1}^{N} x_k f_k}{\sum_{k=1}^{N} f_k} = \sum_{k=1}^{N} x_k g_k$$

Mértani átlag:

$$\overline{x}_g = \sqrt[\sum_{k=1}^{N} f_k]{\prod_{k=1}^{N} x_k^{f_k}} = \prod_{k=1}^{N} x_k^{g_k}$$

Harmonikus átlag:

$$\overline{x}_h = \frac{\sum_{k=1}^{N} f_k}{\sum_{k=1}^{N} \frac{f_k}{x_k}} = \frac{1}{\sum_{k=1}^{N} \frac{g_k}{x_k}}$$

Négyzetes átlag:

$$\overline{x}_q = \sqrt{\frac{\sum_{k=1}^{N} x_k^2 f_k}{\sum_{k=1}^{N} f_k}} = \sqrt{\sum_{k=1}^{N} x_k^2 g_k}$$

Számítsuk ki az előző példában szereplő átlagokat súlyozott formában! A 2. táblázatban található értékekből nyertük a következő csoportosított adatokat.

x_k	f_k
1	14
2	27
3	14
4	6
5	1
Összesen	62

3. táblázat. A Nyíregyházi Főiskola PTI szakosai Analízis II vizsgájának csoportosított végeredményei a 2009/2010-es tanévben

Nyissuk meg az analizis-vizsga.xls fájlt, nevezünk meg egy üres munkalapot „Átlagok” névvel és végezzük el a következő lépéseket!

`A2` $= 1$ `A3` $= 2$ ... `A6` $= 5$ `xk` $\doteq$ `A2:A6`

`B2` $= 14$ `B3` $= 27$ `B3` $= 14$ `B4` $= 6$ `B5` $= 1$ `fk` $\doteq$ `B2:B6`

Jelölje n az összes elem számát azaz $n = \sum_{k=1}^{N} f_k$.

`A8` $=$ n$=$ `B8` $\leftarrow$ `SZUM(fk)` `n` $\doteq$ `B8`

Mivel relatív gyakorisággal egyszerűbb számolni, ezért bevezetjük a g_k oszlopot.

`C2` $\leftarrow$ `B2/n` `C2` $\rightsquigarrow$ `C3:C6` `gk` $\doteq$ `C2:C6`

A részszámításokat a többi oszlopon végezzük.

`D2` $\leftarrow$ `xk*gk` `D2` $\rightsquigarrow$ `D3:D6` `D8` $\leftarrow$ `SZUM(D2:D6)`

`E2` $\leftarrow$ `xk^gk` `E2` $\rightsquigarrow$ `E3:E6` `E8` $\leftarrow$ `SZORZAT(E2:E6)`

`F2` ← `gk/xk` `F2` ⇝ `F3:F6` `F8` ← `SZUM(F2:F6)`

`G2` ← `xk^2*gk` `G2` ⇝ `G3:G6` `G8` ← `SZUM(G2:G6)`

Ezekből megkapjuk a keresett átlagokat.

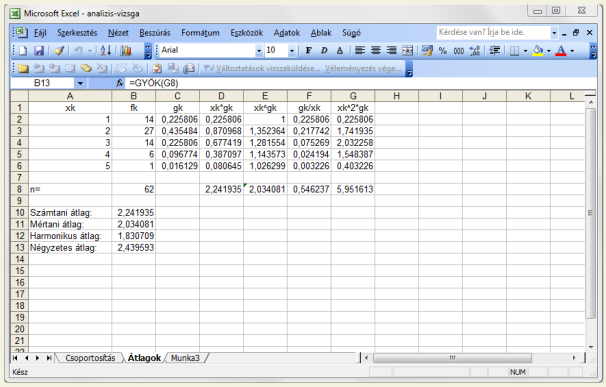
`A10` = Számtani átlag: `B10` ← `D8`

`A11` = Mértani átlag: `B11` ← `E8`

`A12` = Harmonikus átlag: `B12` ← `1/F8`

`A13` = Négyzetes átlag: `B13` ← `GYÖK(G8)`

Az eredmény a következő ábrán látható.



6. ábra. A Nyíregyházi Főiskola PTI szakosai Analízis II vizsgája végeredményeinek átlaga a 2009/2010-es tanévben

2. Gyakorisági sorok

rangsor

Az adatgyűjtések közül fontos szerepet játszanak azok, amely egyetlen egy mennyiségi ismérvből állnak, azaz számadatok hosszú listája. Egy számokból álló alapadat sorból további számokat képezhetünk (pl. átlagokat számíthatunk), amik segítenek jobban leírni a vizsgált sokaságot. Ezenkívül rangsorolhatjuk őket. A *rangsor* a mennyiségi ismérv értékeinek monoton sorozata. Sok esetben a rangsor jó szemléltető eszköznek bizonyul.

gyakorisági sor

Nagyon sok alapadat esetén a rangsor nem nyújt megfelelő segítséget, hiszen ez ilyenkor nehezen áttekinthető. Szükség lehet arra, hogy csoportosítást végezzünk a rangsorolt adatokból. Az ilyen csoportosítást *gyakorisági sornak* nevezzük. Azonban a csoportosítás elkészítésére nem mindegy, hogy az alapadatok milyen számokat tartalmaznak.

diszkrét sokaságok

Lehetséges, hogy a folyamat, amit vizsgálunk, jól elkülönülő értékeket vehet fel (azaz megszámlálható számosságú halmaz). Az ilyen folyamatokból nyert sokaságokat *diszkrét sokaságoknak* nevezzük. Ilyen sokaságok mellett

egyszerűen elkészíthetjük az értékek szerinti csoportosítást, ha a sokaság kevés számú értékből áll. Ez történik például, ha az alapadatok érdemjegyek, melyek 1-től 5-g terjedő egész értékek lehetnek.

folytonos sokaságok

Más a helyzet, ha a hosszú adatsorunk alig tartalmaz egyforma értékeket, hiszen majdnem olyan hosszú gyakorisági sort kapunk, ha érték szerint csoportosítanánk. Ez jellemző a nem diszkrét, azaz *folytonos sokaságokra*. Az ilyen sokaságok értékei akár egy intervallum tetszőleges elemei is lehetnek (pl. bizonyos legyártott termékek súlya), így annak a valószínűsége, hogy két egyforma értéket kapjuk nullával egyenlő. A gyakorlatban azonban minden mérés csak bizonyos hibahatár mellet végezhető el, de így is kis hibahatárnál alig kapunk azonos értékeket. Egyforma értékeket alig tartalmazó hosszú adatsorokat kaphatunk diszkrét sokaságoknál is.

osztályközös gyakorisági sor

A fenti esetekben érdemes olyan osztályokat készíteni, melyek egymásba nem nyúló és a teljes adatsort tartalmazó intervallumokból áll. Az ilyen csoportosítást *osztályközös gyakorisági sornak* nevezzük.

Lássuk egy példát! A 2010-ponthatarok.xls fájl tartalmazza a felvételi ponthatárokat a Nyíregyházi Főiskola összes alap- és felsőfokú szakképzési szakán a 2010-es tanévben. A következő táblázat tartalmazza ezeket az adatokat.

354	240	401	320	357	374	356	371	236	244
200	200	241	400	318	400	400	323	326	366
382	401	420	205	196	288	250	205	202	398
398	398	402	406	211	252	251	447	278	294
224	202	204	248	288	288	358	358	358	318
280	260	209	206	196	400	408	400	342	318
211	211	326	208	211	206	367	366	211	226
248	248	408	248	244	248	306	407	407	

4. táblázat. Felvételi ponthatárok a Nyíregyházi Főiskola a 2010-es tanévben

Látható, hogy az adatsor sok különböző értéket tartalmaz. Ebben az esetben egy osztályközös gyakorisági sort készítünk az adatok jobb szemléltetése érdekében. Először dönteni kell arról, hogy milyen legyen a csoportosítás. Pontosabban meg kell határozni azokat az intervallumokat, mely szerint a csoportosítást végezzük. Ebben a esetben olyan gyakorisági sort készítünk, amely 50 pontonként csoportosítja a ponthatárokat 200 ponttól 400 pontig, és külön összeszámolja még a kis és a nagy értékű ponthatárokat. Ehhez az Excel `GYAKORISÁG` függvénye hasznos segítség jelent a csoportosított adatsor elkészítéséhez.

Végezzük el a következő lépéseket! A 2010-ponthatarok.xls fájl „Adatsor” munkalapján találjuk a csoportosítandó adatokat. Ezeket x-szel nevezzük el.

x ≐ D3:D81

Ezek után hozzunk létre egy „Gyakorisági sor” munkalapot és ott végezzük el a csoportosítást a h segédváltozóval , amivel megadjuk az intervallumok alsó ak és felső bk határát.

A1 = h= B1 = 50 h ≐ B1

B3 = 200 A4 = 200 B4 ← B3+h B4 ⇝ A5:B8

A3 ← MIN(x) B8 ← MAX(x) ak ≐ A3:A8 bk ≐ B3:B8

A GYAKORISÁG függvénnyel fogjuk elvégezi a csoportosítást. Két argumentuma van, az első az adattömb, melynek csoportosítását végezzük, a másik az intervallumok határát tartalmazó tömb, mely szerint a csoportosítást végezzük. Ez azt jelenti, hogy elég lett volna a bk felső határokat tartalmazó adattömböt kiszámolni. A keletkező intervallumok alulról nyíltak és felülről zártak lesznek.

D3 ← GYAKORISÁG(x;bk) D3 ⇝ D4:D8

Mivel a GYAKORISÁG tömböt ad eredményül, tömbképletként kell megadnunk. Ehhez jelöljük ki a D3:D8 tartományt, nyomjuk meg az $F2$ gombot, végül a Ctrl+Shift+Enter billentyűkombinációval kapjuk meg a gyakorisági oszlopot. Csak az adatsor elemszámának meghatározása marad hátra.

C10 = n= D10 ← SZUM(fk) n ≐ D10 fk ≐ D3:D8

Ponthatárok	Kurzusok száma
196 — 200	4
201 — 250	27
251 — 300	9
301 — 350	9
351 — 400	20
401 — 447	10
Összesen	79

5. táblázat. Felvételi ponthatárokat a Nyíregyházi Főiskolán a 2010-es tanévben

Az eredmény a 5. táblázatban látható.

A GYAKORISÁG függvénnyel készített osztályközös gyakorisági sor olyan intervallumokkal dolgozik, melyek hézagmentesen illeszkednek egymáshoz, azaz az előző intervallum felső határa megegyezik a következő intervallum alsó határával. Kérdés, hogy a határra eső értékek melyik osztályhoz tartoznak. A GYAKORISÁG függvény az ilyen értékeket az előző intervallumhoz sorolja. Ezt úgy szokás közölni a gyakorisági sorokon, hogy egy kicsivel megnövelik az alsó határokat a második osztálytól kezdődően. Ez a kicsi szám függ az

adatsor értékeitől. A példánkban ez a kicsi szám 1, hiszen egész értékekről van szó.

Az osztályközös gyakorisági sor elkészítése esetén az igazi probléma annak eldöntése, hogy melyik a legjobb csoportosítás. Ez abból ered, hogy minden osztályközös csoportosítás bizonyos mértékű adatvesztéssel jár, hiszen ilyen gyakorisági sor nem tartalmaz konkrét adatokat. Például, az 5. táblázat szerint 4 olyan kurzus van, melynek ponthatárai 196 és 200 között van, de az eredeti adatsor nélkül lehetetlen megmondani a pontos értéküket. Az adatvesztés csökkenthető az osztályok számának megnövelésével és azzal együtt az intervallumok hosszának csökkentésével. A gond az, hogy ez egy idő után a szemléltetés kárára megy. Meg kell találni a megfelelő kompromisszumot.

Ha más szakmai vagy egyéb indokok alapján nem tudunk dönteni, akkor javasoljuk a következő módszert. Osszuk el egyforma részekre az adatsor legkisebb és legnagyobb eleme közötti távolságát és ilyen osztályokat képezzünk! Az osztályok számát a következő módon határozzuk meg. Ha a számuk N és az adatsor elemszáma n, akkor N legyen az a legkisebb egész szám, amire

$$2^N > n$$

teljesül. Ha ilyen csoportosítás mellett döntöttünk volna, akkor a példánkban 7 osztályt kellet volna készíteni és az intervallumok hossza egyformán

$$h = \frac{\max(x) - \min(x)}{N} = \frac{447 - 196}{7} = 35,8571$$

értékkel lett volna egyenlő.

hisztogram

Az osztályközös gyakorisági sorokat szokás *hisztogrammal* ábrázolni. A hisztogram egy olyan oszlopdiagram, amivel hézagmentesen egymás mellé illesztett téglalapokból áll és értékei a gyakoriság vagy annak többszöröse (pl. relatív gyakoriság).

sűrűségi hisztogram

Ilyen diagramot nem nehéz az Excel programmal készíteni, azonban így csak azonos szelességű oszlopokat készíthetünk. Ez azt jelenti, hogy csak akkor kapunk egy arányos, jól szemléltethető hisztogramot, ha az osztályközös gyakorisági sor minden osztályának hossza megegyezik. Ellenkező esetben csak akkor biztosítható az arányosság, ha az oszlopok szélessége arányos az osztályok hosszával és az értékeket úgy súlyozzuk, hogy a gyakorisági értékeket elosztjuk az egyes intervallumok hosszával. Ilyen módon kapjuk például a *sűrűségi hisztogramot*, melynek értékei a relatív gyakoriságnak és az egyes intervallumok hosszának aránya.

gyakorisági poligon

Az Excel programmal készített hisztogramok nem lesznek minden esetben torzításoktól mentesek. Van azonban egy másik szemléltető diagram, ami jól elkészíthető, és ez a *gyakorisági poligon*. A gyakorisági poligon olyan vonalakkal összekötött pontdiagram, mely pontok x koordinátája az intervallumok közepe és y koordinátája az intervallumok hosszával súlyozott gyakorisági érték. Tehát az osztályközös gyakorisági sor minden f_k gyakoriságú $I_k = [a_k, b_k]$ intervalluma esetén képezzük az

$$x_k = \frac{a_k + b_k}{2}, \qquad y_k = \frac{f_k}{b_k - a_k}$$

koordinátájú pontot. Ezenkívül az első és utolsó pontot összekötjük az x tengelyen az első osztályközt megelőző (azzal azonos hosszúságú) osztályköz, ill. az utolsó osztályközt követő (azzal azonos hosszúságú) osztályköz középpontjával.

Készítsünk hisztogramot és gyakorisági poligont a Nyíregyházi Főiskola felvételi ponthatáraihoz! Először elkészítjük a diagramok adatforrásait. A hisztogramhoz szükséges a közölt határok megadása.

`F3` = 196 - 200 `F4` = 201 - 250 ... `F8` = 401 - 447

A hisztogram az az oszlopdiagram, mely egy adatsort tartalmaz, értékei `D3:D8` és az x tengely feliratai `F3:F8`. A gyakorisági poligonhoz szükséges a pontok koordinátái.

`H3` ← `bk-ak` `H3` ⇝ `H4:H8` `hk` ≐ `H3:H8`

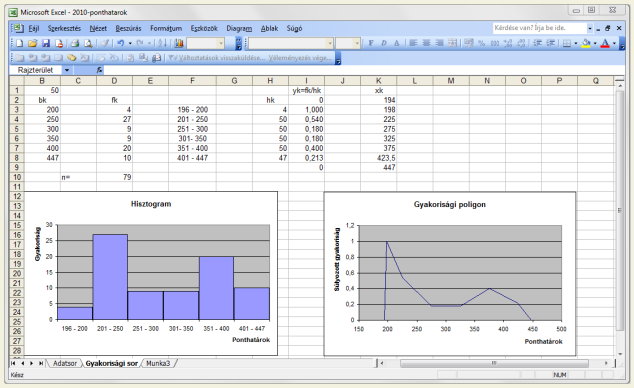
`I3` ← `fk/hk` `I3` ⇝ `I4:I8` `I2` = 0 `I9` = 0 `I1` = `yk=fk/hk`

`K3` ← `(ak+bk)/2` `K3` ⇝ `K4:K8` `K2` = `2*A3-K3` `K9` = `2*K8-A8`

`K1` = `xk`

A gyakorisági poligon az a vonallal összekötött pontdiagram, mely egy adatsor tartalmaz, x értékei `K2:K9` és y értékei `I2:I9`.

Az eredmény a következő ábrán látható.



7. ábra. Hisztogram és gyakorisági poligon a Nyíregyházi Főiskola 2010-es tanévére vonatkozó felvételi ponthatárokra

Végül szeretnénk megjegyezni, hogy gyakorisági sorokat és hisztogramokat az Excel Eszközök → Adatelemzés → Hisztogram menüjével is el lehet készíteni.

3. Helyzetmutatók

helyzetmutatók

Az alapadatok mennyiségi ismérv szerinti eloszlásának alakjáról sokat elárulnak az előző részben tanult ábrázolások, de a tömör jellemzés érdekében célszerű számszerű információt keresni a legfontosabb tulajdonságairól. Ha azt keressük, hogy az alapadatokhoz viszonyítva melyik helyen kapunk egy bizonyos tulajdonságot, akkor *helyzetmutatókról* beszélünk. Több helyzetmutatóval jellemezhetjük az alapadatok eloszlását. Ezek az átlag, módusz, medián és kvantilisek.

átlag

Az *átlag* nem más, mint az értékek számtani közepe. Jele: $\overline{x}$. Az 1.4. részben már foglalkoztunk vele. Az átlag a matematikai statisztika fontos eszköze, jelentőségét a következő fejezetekben fogjuk igazán látni.

módusz

A *módusz* az adatsor leggyakoribb eleme, vagyis az az elem, amely legtöbbször fordul elő az adatsorban. Természetesen az adatsornak több ilyen eleme is lehet, ekkor többmóduszú adatsorról beszélünk. A módusz a gyakorisági poligon maximum helye. Jele: M_o.

medián

A *medián* az adatsor azon értéke, amelynél ugyanannyi kisebb, mint nagyobb érték fordul elő, azaz a rangsorolt adatsor közepe. A medián az az érték, amely a gyakorisági poligon alatti területet két egyforma részre bontja. Jele: M_e.

kvantilisek

A *kvantilisek* olyan értékek, melyek a rangsorolt adatsort megadott arányban osztják ketté, és ugyanazt teszik a gyakorisági poligon alatti területtel. Ha ez az arány $q : (1-q)$, ahol $0 < q < 1$, akkor q-adik kvantilisről beszélünk. Jele: Q_q. A q érték függvényében különböző módon is nevezhetjük a kvantiliseket.

Medián: Ha $q = \frac{1}{2}$, akkor $Q_{\frac{1}{2}} = M_e$.

Tercilisek: Ha $q = \frac{1}{3}$ vagy annak egész többszöröse

$$T_1 := Q_{\frac{1}{3}}, \qquad T_2 := Q_{\frac{2}{3}}$$

Kvartilisek: Ha $q = \frac{1}{4}$ vagy annak egész többszöröse

$$Q_1 := Q_{\frac{1}{4}}, \qquad Q_2 := Q_{\frac{2}{4}} = M_e \qquad Q_3 := Q_{\frac{3}{4}}$$

Hasonlóan beszélhetünk *Kvintilisekről*, ha $q = \frac{1}{5}$ vagy annak egész többszöröse, *decilisekről*, ha $q = \frac{1}{10}$ vagy annak egész többszöröse, vagy *percentilisekről*, ha $q = \frac{1}{100}$ vagy annak egész többszöröse.

A helyzetmutatók kiszámítása különbözik attól, hogy milyen adatsorral rendelkezünk.

3.1. Helyzetmutatók egyszerű adatsorok esetén

Tegyük fel, hogy egyszerű adatsorunk n darab számadatból áll és értékei az

$$x_1,\ x_2,\ x_3,\ \ldots,\ x_n.$$

Ekkor

Átlag $\overline{x} := \dfrac{1}{n}\sum_{k=1}^{n} x_k = \dfrac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$

Átlagot az Excel-ben az `ÁTLAG` függvénnyel érdemes számítani.

Módusz A rangsorolt adatsorból könnyebben meg tudjuk keresni a leggyakoribb elemet. Móduszt az Excel-ben az `MÓDUSZ` függvénnyel érdemes számítani.

Medián A medián kiszámítása előtt rangsorolni kell az adatsort. Ha az adatsor elemszáma páratlan, akkor a medián a rangsorolt adatsor középső eleme, melynek indexe $\dfrac{n+1}{2}$. Ha az adatsor elemszáma páros, akkor a medián a rangsorolt adatsor két középső elemének számtani közepe. Ekkor e két középső elem indexe $\dfrac{n}{2}$ és $\dfrac{n}{2}+1$. Mediánt az Excel-ben az `MEDIÁN` függvénnyel érdemes számítani, ekkor nem szükséges az elemek rangsorolása.

Kvantilisek A kvantilisek kiszámítása előtt rangsorolni kell az adatsort. Jelölje

$$y_1,\ y_2,\ y_3,\ \ldots,\ y_n.$$

a rangsorolt adatsort. Legyen továbbá m és t a $q(n+1)$ szám egész része és törtrésze, azaz

$$m := [q(n+1)], \qquad t := q(n+1) - m.$$

Ekkor

$$Q_q := y_m + t(y_{m+1} - y_m).$$

Kvantiliseket az Excel-ben a `PERCENTILIS` függvénnyel érdemes számítani, de `KVARTILIS` függvényt is találunk a függvények között.

Lássuk a következő példát! Számítsuk ki a 2 táblázatban található Analízis II vizsga végeredményeiből álló adatsor átlagát, móduszát, mediánját és kvantilisei közül az első tercilisét, kvartiliseit, első decilisét és tizennyolcadik percentilisét.

Ehhez nyissa meg az analizis-vizsga.xls fájlt és az első munkalapján írja be a következőket:

`C16` = Módusz: `D16` ← `MŰDUSZ(x)`

`C17` = Medián: `D17` ← `MEDIÁN(x)`

`C19` = 1. tercilis: `D19` ← `PERCENTILIS(x;1/3)`

`C21` = 1. kvartilis: `D21` ← `KVARTILIS(x;1)`

`C22` = 3. kvartilis: `D22` ← `KVARTILIS(x;3)`

`C24` = 1. decilis: `D24` ← `PERCENTILIS(x;1/10)`

`C25` = 18. percentilis: `D25` ← `PERCENTILIS(x;18/100)`

3.2. Helyzetmutatók csoportosított adatsorok esetén

Tegyük fel, hogy olyan gyakorisági sornál szeretnénk helyzetmutatókat meghatározni, ahol a különböző értékek szerint csoportosítunk. Jelölje

$$x_1,\, x_2,\, x_3,\, \ldots,\, x_N$$

a különböző számadatokat, és

$$f_1,\, f_2,\, f_3,\, \ldots,\, f_N$$

ezek gyakoriságát. Jelölje $n = \sum_{k=1}^{N} f_k$ az adatsor elemszáma. Továbbá érdemes elkészíteni a gyakoriságok halmozott összeadását, az ú.n. *kumulált gyakoriságot*. Jele: f'_k. Pontosabban:

kumulált gyakoriság

$$f'_1 = f_1, \qquad f'_k = f_1 + f_2 + \cdots + f_k \quad (k > 1).$$

Ekkor

Átlag A 1.4. részben alkalmazott súlyozott számítási formát alkalmazzuk.

$$\overline{x} = \frac{\sum_{k=1}^{N} x_k f_k}{n}$$

Ha az Excel-ben dolgozunk, akkor érdemes egy segédoszlopot készíteni, ami tartalmazza az $x_k f_k$ szorzatokat. Ezt *értékösszegsornak* nevezzük és $S_k := x_k f_k$ módon jelöljük. Így

értékösszegsor

$$\overline{x} = \frac{\sum_{k=1}^{N} S_k}{n}$$

Példát az analizis-vizsga.xls fájlban találunk.

Módusz A legnagyobb f_k gyakorisághoz tartozó x_k érték. A példában látható, hogy a legnagyobb gyakorisági érték $f_2 = 27$ és ehhez az $x_2 = 2$ tartozik. Ezért $M_o = 2$.

Medián Először számítsuk ki az $\frac{n+1}{2}$ egész részét és jelöljük ezt m-mel, azaz

$$m := \left[\frac{n+1}{2}\right].$$

Keressük meg azt a k sorszámú osztályt, amire igaz, hogy az osztály kumulált gyakorisági értéke nagyobb vagy egyenlő, mint m, de az előző osztály kumulált gyakorisági érteke kisebb, mint m, azaz

$$f'_k \geq m, \qquad f'_{k-1} < m.$$

Az így kapott k osztályt medián osztálynak nevezzük és a hozzá tartozó x_k értéket m_e-vel jelöljük. Ha az adatsor elemszáma páros, vagy $m < f'_k$, akkor $M_e = m_e$. Ellenkező esetben

$$M_e = \frac{x_k + x_{k+1}}{2},$$

azaz a medián osztályhoz és a következő osztályhoz tartozó értékek számtani közepe.

Kvantilisek A medián kiszámításához hasonlóan vegyük először a $q(n+1)$ szám egész részét és törtrészét, azaz

$$m := [q(n+1)], \qquad t := q(n+1) - m.$$

Keressük meg azt a k sorszámú osztályt, amire igaz, hogy az osztály kumulált gyakorisági értéke nagyobb vagy egyenlő, mint m, de az előző osztály kumulált gyakorisági értéke kisebb, mint m, azaz

$$f'_k \geq m, \qquad f'_{k-1} < m.$$

Az így kapott k osztályt kvantilis osztálynak nevezzük és a hozzá tartozó x_k értéket K_v-vel jelöljük. Ha $t = 0$, vagy $m < f'_k$, akkor $Q_q = K_v$. Ellenkező esetben

$$Q_q := x_k + t(x_{k+1} - x_k).$$

3.3. Helyzetmutatók osztályközös gyakorisági sorok esetén

Az osztályközös gyakorisági sorok általános sémáját a következő táblázat mutatja:

Osztályhatárok	Gyakorisság
a_1 — b_1	f_1
a_2 — b_2	f_2
a_3 — b_3	f_3
⋮ ⋮	⋮
a_N — b_N	f_N
Összesen	n

6. táblázat. Osztályközös gyakorisági sorok általános sémája

Már említettük, hogy az osztályközös gyakorisági sorok hézagmentesen illeszkedő osztályokat feltételez, de általában úgy közlik, hogy melyik osztályhoz tartoznak a határra eső értékek, hogy valamennyivel megnövelik az alsó határokat. Ebben az esetben a helyzetmutatók kiszámítása előtt vissza kell állítani az alsó határokat úgy, hogy megegyezzen az előző intervallum felső határával, azaz

$$a_k = b_{k-1}, \qquad (k = 2, 3, \ldots, N).$$

Másrészt előfordul olyan osztályközös gyakorisági sor, amelynek első vagy utolsó osztálya nyitott, azaz az a_1 vagy a b_N értékek nincsenek megadva. Ekkor magunk adjuk meg a hiányzó értékeket úgy, hogy a nyitott osztály hossza megegyezzen a közvetlen mellette levő osztály hosszával, és így

$$a_1 = b_1 + a_2 - b_2, \qquad b_N = a_N + b_{N-1} - a_{N-1}.$$

Az előző korrekciók mellet a medián és kvantilisek kiszámításához el kell készíteni az f' kumulált gyakoriságokat. Szeretnénk megjegyezni, hogy az osztályközös gyakorisági sorok nem tartalmaznak pontos adatokat, ezért a helyzetmutatókat csak becsülni tudjuk.

Átlag Először számítsuk ki az

$$x_k = \frac{a_k + b_k}{2} \qquad (k = 1, 2, \ldots, N)$$

osztályközepeket. Ezután úgy számítjuk ki az átlagot, mint egy csoportosított adatsor esetén, azaz

$$\overline{x} = \frac{\sum_{k=1}^{N} x_k f_k}{n}$$

Módusz Keressük meg a legnagyobb gyakorisággal rendelkező osztályt, amelyet modális osztálynak fogunk nevezni. Legyen i a modális osztály sorszáma. Ez azt jelenti, hogy f_i a legnagyobb gyakorisági érték az egész gyakorisági sorban. Jelölje

$m_o = a_i$, a modális osztály alsó határát,

$k_1 = f_i - f_{i-1}$, a modális és az azt megelőző osztály gyakoriságának különbségét (ha $i = 1$, akkor $k_1 := f_i$),

$k_2 = f_i - f_{i+1}$, a modális és az azt követő osztály gyakoriságának különbségét (ha $i = N$, akkor $k_2 := f_i$),

$h = b_i - a_i$, a modális osztály hosszát.

Ekkor

$$M_o = m_o + \frac{k_1}{k_1 + k_2} h.$$

Medián Legyen $m = \frac{n}{2}$ és keressük meg azt az osztályt, amire a kumulált gyakorisági érteke nagyobb vagy egyenlő, mint m, de az előző osztály kumulált gyakorisági érteke kisebb, mint m, azaz ha i ennek az osztálynak a sorszáma, akkor

$$f'_i \geq m, \qquad f'_{i-1} < m.$$

Az így kapott osztály tartalmazza a mediánt. Jelölje

$m_e = a_i$, a mediánt tartalmazó osztály alsó határát,

$f_{m_e} = f_i$, a mediánt tartalmazó osztály gyakoriságát,

$f'_{m_e-1} = f'_{i-1}$, a mediánt tartalmazó osztályt megelőző osztály kumulált gyakoriságát (ha $i = 1$, akkor $f'_{m_e-1} := 0$),

$h = b_i - a_i$, a mediánt tartalmazó osztály hosszát.

Ekkor

$$M_e = m_e + \frac{\frac{n}{2} - f'_{m_e-1}}{f_{m_e}} h.$$

Kvantilisek Legyen $m = qn$ és keressük meg azt az osztályt, amire a kumulált gyakorisági érteke nagyobb vagy egyenlő, mint m, de az előző osztály kumulált gyakorisági érteke kisebb, mint m, azaz ha i ennek az osztálynak a sorszáma, akkor

$$f'_i \geq m, \qquad f'_{i-1} < m.$$

Az így kapott osztály tartalmazza a kvantilist. Jelölje

$K_v = a_i$, a kvantilis tartalmazó osztály alsó határát,

$f_{K_v} = f_i$, a kvantilis tartalmazó osztály gyakoriságát,

$f'_{K_v-1} = f'_{i-1}$, a kvantilis tartalmazó osztályt megelőző osztály kumulált gyakoriságát,

$h = b_i - a_i$, a kvantilis tartalmazó osztály hosszát.

Ekkor

$$Q_q = K_v + \frac{m - f'_{K_v-1}}{f_{K_v}} h.$$

Példaként számítsuk ki helyzetmutatókat a felvételi ponthatárokból készített osztályközös gyakorisági sorban. A számításokhoz hozzunk létre egy „Mutatók” elnevezésű munkalapot a 2010-ponthatarok.xls fájlban.

Ponthatárok	Kurzusok száma
196 — 200	4
201 — 250	27
251 — 300	9
301 — 350	9
351 — 400	20
401 — 447	10
Összesen	79

7. táblázat. Felvételi ponthatárok a Nyíregyházi Főiskola 2010-es tanévében

Kezdjük azzal, hogy a „Gyakorisági sor” munkalap `A2:D10` tartományát átmásoljuk a „Mutatók” munkalap `B2:E10` tartományára. Megjegyezzük, hogy most nem kell névvel ellátni az a_k, b_k és f_k oszlopokat, mert ezt már ebben a fájlban megtettük. Azonban egy új feladat esetén érdemes azzal kezdeni. Ezek után hozzuk létre a kumulált gyakorisági sort.

`F3` ← `E3` `F4` ← `F3+E4` `F4` ⇝ `F5:F8` `F2` = f’k `kumf` ≐ `F3:F8`

Az átlag kiszámításához szükségesek az x_k osztályközepek és az S_k értékösszegsorra:

`H3` ← `(B3+C3)/2` `H3` ⇝ `H4:H8` `H2` = xk `xk` ≐ `H3:H8`

`J3` ← `xk*fk` `J3` ⇝ `J4:J8` `J2` = Sk=xk*fk `Sk` ≐ `J3:J8`

`A12` = Átlag: `B12` ← `SZUM(sk)/n`

A módusz, medián és kvantilis értékeket úgy szokás számolni, hogy ránézünk az adatsorra és megállapítjuk a képletben szereplő változók (m_0, m_e, k_1, k_2, h, stb.) értékeit, majd a megfelelő képletbe helyettesítve kiszámítjuk a helyzetmutatókat. A mátrix típusú függvények (`HOL.VAN`, `INDEX`, stb.) segítségével automatikusan is megkaphatjuk a keresett helyzetmutatókat.

A módusz:

B14 = imo B15 ← HOL.VAN(MAX(fk);fk;0) imo ≐ B15

C14 = mo C15 ← INDEX(ak;imo) mo ≐ C15

D14 = k_1 D15 ← HA(imo=1;INDEX(fk;imo);INDEX(fk;imo)-INDEX(fk;imo-1))
k_1 ≐ D15

E14 = k_2 E15 ← HA(imo=SOROK(fk);INDEX(fk;imo);INDEX(fk;imo)-INDEX(fk;imo+1))
k_2 ≐ E15

F14 = hmo F15 ← INDEX(bk;imo)-INDEX(ak;imo) hmo ≐ F15

A17 = Módusz: B17 ← mo+k_1/(k_1+k_2)*hmo

A medián:

A19 = m A20 ← n/2 m ≐ A20

B19 = ime
B20 ← HA(m<INDEX(kumf;1);1;HA(m=INDEX(kumf;HOL.VAN(m;kumf));HOL.VAN(m;kumf);HOL.VAN(m;kumf)+1))
ime ≐ B20

C19 = me C20 ← INDEX(ak;ime) me ≐ C20

D19 = f'me-1 D20 ← HA(ime=1;0;INDEX(kumf;ime-1))
kum_me_min ≐ D20

E19 = fme E20 ← INDEX(fk;ime) fme ≐ E20

F19 = hme F20 ← INDEX(bk;ime)-INDEX(ak;ime) hme ≐ F20

A22 = Medián: B22 ← me+(n/2-kum_me_min)/fme*hme

A kvantilisek esetén meg kell adni a q értékét. Most $q = \frac{1}{3}$, ami az első tercilisnek felel meg.

A24 = q= B24 ← 1/3

A25 = mq A20 ← n*q mq ≐ A20

B25 = imq
B26 ← HA(mq<INDEX(kumf;1);1;HA(mq=INDEX(kumf;HOL.VAN(mq;kumf));HOL.VAN(mq;kumf);HOL.VAN(mq;kumf)+1
imq ≐ B26

C25 = Kv C26 ← INDEX(ak;imq) Kv ≐ C26

D25 = f'Kv-1 D26 ← HA(imq=1;0;INDEX(kumf;imq-1))
kum_Kv_min ≐ D26

E25 = fKv E26 ← INDEX(fk;imq) fKv ≐ E26

F25 = hmq F26 ← INDEX(bk;imq)-INDEX(ak;imq) hmq ≐ F26

A28 = Kvantilis: B28 ← Kv+(mq-Kum_kv_min)/fKv*hKv

4. Szóródási mutatók

Két, ugyanazzal a helyzetmutatókkal rendelkező adatsor viselkedése nagyon eltérő is lehet. Ha például az analízis vizsga osztályzatai ugyanannyi egyes és ötösből áll, az nem ugyanaz, mintha mindenkinek hármasa lenne. Pedig mindkét esetben az átlag hárommal egyenlő. Ezért szükség van arra, hogy mérni tudjuk mennyire koncentrálódnak az adatsor értékei az átlag körül. Az ilyen mutatókat *szóródási mutatóknak* nevezzük.

Több ilyen mutató is van:

A szóródás terjedelme Jele: R. Az adatsor legnagyobb és legkisebb értékének különbsége, azaz

$$R = \max x - \min x.$$

Rangsorolt minta esetén megegyezik az utolsó és ez első elem különbségével. Osztályközös gyakorisági sorok esetén egyenlő az utolsó osztály felső határa és az első osztály alsó határának különbségével, azaz a 6. táblázat szerint

$$R = b_N - a_1.$$

Az interkvartilis terjedelme Jele: IQR. Az adatsor harmadik és első kvartilisének különbsége, azaz

$$IRQ = Q_3 - Q_1.$$

Azt adja meg, hogy melyik intervallumban találhatók az adatsor értékeinek középső fele.

A szórás Jele: s. Az adatsor értékeinek az átlagtól vett eltérései négyzetes átlaga, azaz

$$s = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \overline{x})^2}{n}}, \quad \text{vagy} \quad s = \sqrt{\frac{\sum_{k=1}^{N} f_k(x_k - \overline{x})^2}{n}}$$

attól függően, hogy egyszerű vagy csoportosított adatsorral van dolgunk. Osztályközös gyakorisági sorok esetén a jobboldali képlettel számolunk, ahol x_k az osztályközepeket jelöli.

A korrigált szórás Jele: s^*. A matematikai statisztikai számításokban (becslésekben, próbafüggvényekben, stb) a szórás képletében korrekciót végeznek el, eggyel kisebb számmal átlagolnak, azaz

$$s^* = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \overline{x})^2}{n-1}}, \quad \text{vagy} \quad s^* = \sqrt{\frac{\sum_{k=1}^{N} f_k(x_k - \overline{x})^2}{n-1}}.$$

Sőt szórás alatt gyakran korrigált szórást értenek. Például az Excel `SZÓRÁS` függvénye valójában korrigált szórást számol. Nyilvánvaló, a szórás és a korrigált szórás között a következő kapcsolat áll fenn:

$$s = \sqrt{\frac{n-1}{n}}s^*, \qquad s^* = \sqrt{\frac{n}{n-1}}s.$$

A relatív szórás Jele: V. Gyakran nagyobb értékek esetén nagyobb szóródás is megengedhető, ezért olyan mutatóra is szükség lehet, ami a szórást az átlaghoz viszonyítja, azaz

$$V = \frac{s}{\overline{x}}.$$

A relatív szórást százalékos formában szokták megadni.

Számítsuk ki a szóródási mutatókat az analizis-vizsga.xls fájlban található példában.

`C28` = Szóródás terjedelme: `D28` ← `MAX(x)-MIN(x)`

`C29` = Interkvartilis terjedelme: `D29` ← `KVARTILIS(x;3)-KVARTILIS(x;1)`

`C30` = Korrigált szórás: `D30` ← `SZÓRÁS(x)`

`C31` = Szórás: `D31` ← `GYÖK((n-1)/n)*SZÓRÁS(x)`

`C32` = Relatív szórás: `D32` ← `SZÓRÁS(x)/ÁTLAG(x)*100` `E32` = %

Lássuk, hogyan tudunk szóródási mutatókat számolni osztályközös gyakorisági soroknál! Ehhez vegyük a 2010-ponthatarok.xls fájlban található példát.

`A31` = Szóródás terjedelme: `B31` ← `NDEX(bk;SOROK(bk))-INDEX(ak;1)`

A már bemutatott kvantilis-számítással megállapíthatjuk az első és harmadik kvartilist, az értékeket beírjuk az `E32` és `G32` cellákba, hogy ki tudjuk számolni az adatsor interkvartilis terjedelmét.

`D32` = Q1= `E32` = 229,17 `quart1` ≐ `E32`

F32 = Q3= G32 = 357,63 quart3 ≐ G32

A32 = Interkvartilis terjedelme: B32 ← =quart3-quart1

A szórás kiszámításához nevezzük el a B12 cellát atl néven, hiszen ez a cella már tartalmazza az adatsor átlagát. Szükségünk van egy segédoszlopra, ahol az értékek és az átlag különbségének négyzete található:

atl ≐ B12

K2 = (xk-atl)^2 K3 ← (xk-atl)^2 K3 ⇝ K4:K8 K10 ← SZUM(K3:K8)

A33 = Korrigált szórás: B33 ← GYÖK(K10/(n-1))

A34 = Szórás: B34 ← GYÖK(K10/n)

A35 = Relatív szórás: B35 ← B34/atl*100 C35 = %