Statisztikai elemzések táblázatkezelővel

A statisztikai minta

A statisztikai elemzés tárgya az úgynevezett **statisztikai sokaság**, , az elemzéshez kiválasztott adathalmaz pedig a **statisztikai minta**. Az elemzés célja az, hogy a mintából kiszámított statisztikai jellemzők alapján az eredeti sokaság törvényszerűségeire következtessünk. Torzítatlan eredményt csak akkor kapunk, ha az eredeti sokaságot „lefedő", **reprezentatív mintát** választottunk a vizsgálat kiindulópontjául. A matematikai statisztika egyik legfontosabb feladata éppen az, hogy a közelítő értékek „jóságának" valószínűségét meghatározza. Az Excel függvényei által meghatározott értékek az úgynevezett **empirikus**, vagy **tapasztalatai** mutatószámok, amelyek csak bizonyos valószínűséggel közelítik a statisztikai sokaság valódi mutatószámait. A minta kiválasztása és a következtetés a kutatást végző ember felelőssége.

Az alábbiakban az Excel táblázatkezelő program statisztikai elemzések során használható szolgáltatásait, függvényeit ismertetjük. A függvények alkalmazásához szükséges matematikai statisztikai összefüggéseket ismertnek tételezzük.

Alapstatisztikák

Az adatok elemzésének legegyszerűbb eszközei a viszonyszámok, a grafikus ábrák, a középértékek, és az adatok szóródásának mérőszámai.

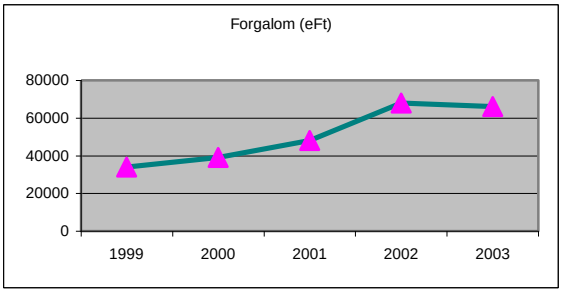
Viszonyszámok

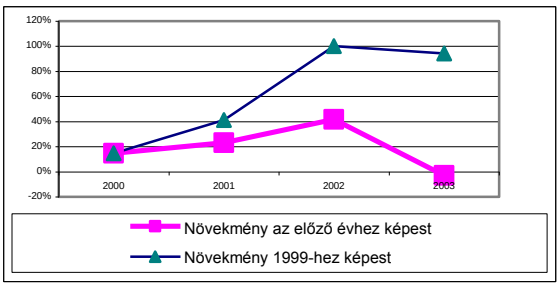
A viszonyszám két egymással logikai kapcsolatban álló adat hányadosa. Ilyen például a népsűrűség (fő/terület), a gépkocsi fogyasztása 100 km-en (l/km*100), stb. A legegyszerűbb viszonyszámok meghatározásához nincs szükség függvényre, elég egy egyszerű képlet.

Feladat.
Az alábbi táblázatban egy kereskedés forgalmának növekedését vizsgáljuk. Az első oszlopban az előző évhez, a második oszlopban, pedig mindig 1999-hez, mint bázis évhez viszonyítjuk az adott év forgalmát.

Év	Forgalom (eFt)	Növekmény az előző évhez képest	Növekmény 1999-hez képest
1999	34000		
2000	39000	15%	15%
2001	48000	23%	41%
2002	68000	42%	100%
2003	66000	-3%	94%
Átlagos növekedés		**19%**	**63%**

Év	Forgalom (eFt)	Növekmény az előző évhez képest	Növekmény 1999-hez képest
1999	34000		
2000	39000	=B3/B2-1	=B3/$B$2-1
2001	48000	=B4/B3-1	=B4/$B$2-1
2002	68000	=B5/B4-1	=B5/$B$2-1
2003	66000	=B6/B5-1	=B6/$B$2-1
Átlag		**=ÁTLAG(C3:C6)**	**=ÁTLAG(D3:D6)**





Középértékek

Átlag

Az adathalmaz elemzésének általában a legelső lépése az, hogy kiszámítjuk az adatok átlagát. Az átlag a matematikai statisztikában leggyakrabban használatos középérték, a **várható érték** közelítő értéke. Az átlag az a középérték, amelytől az adatok előjeles eltéréseinek összege nulla.
Legyen az adatok száma **n**, az adatok pedig rendre $\mathbf{x_1, x_2, \ldots x_{i,\ldots}, x_n}$, ahol $\mathbf{x_i}$ az i-dik adat. Ekkor az átlag jelölése és képlete:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Excelben egy adatblokk számadatainak átlagát az **ÁTLAG(blokkhivatkozás)** függvénnyel határozhatjuk meg.

Fontos tudni, hogy az **ÁTLAG** függvény kiszámításakor az Excel az üres vagy szöveget tartalmazó cellákat figyelmen kívül hagyja.

=ÁTLAG(L2:L5) számítása: $\frac{3+2+1}{3}$, míg az

=ÁTLAG(M2:M5) számítása: $\frac{3+2+0+1}{4}$.

	L	M
2	3	3
3	2	2
4		0
5	1	1
6	**2**	**1,5**

További függvények:

ÁTLAGA(blokkhivatkozás): ua. mint az ÁTLAG függvény, de az argumentumként megadott blokkban szereplő szöveget és HAMIS logikai értéket 0-nak, az IGAZ logikai értéket 1-nek tekintve számolja az átlagot.

Tegyük fel, hogy négy gyerek matematika jegyeinek átlaga 3. Az átlag alapján feltételezhetjük, hogy közepes képességű gyerekekről van szó. Pedig az eredeti jegyek lehettek a 3,5,3,5 de éppúgy lehettek a 3,3,3,3 is. Az első esetben van a gyerekek között nagyon jó, és közepes matekos, a második esetben mindenki közepes.

Általában igaz, hogy az átlag önmagában nem hordoz elegendő információt az adatok elemzéséhez. Semmit nem mond arról, hogy mekkora az adathalmaz terjedelme, mennyire „szórnak” az adatok.

Az 5, 6, 7, az 1, 6, 11 és a 6, 6, 6 háromelemű adatsor átlaga egyaránt 6.

A helyes következtetések további jellemzők, középértékek meghatározására van szükség.

Terjedelem

A **terjedelem** az adathalmaz legnagyobb és legkisebb elemének különbsége. Excelben az alábbi egyszerű képlettel határozható meg:

=MAX(blokkhivatkozás)-MIN(blokkhivatkozás)

A képlet eredményeként kapott érték megadja, hogy mekkora az adathalmaz összes elemét tartalmazó intervallum.

Medián

A **medián** a nagyság szerint sorba rendezett adathalmaz középső értéke. A meghatározásból következik, hogy az adatok 50%-a kisebb, 50%-a pedig nagyobb a mediánnál. Ha az adatsor elemeinek száma páratlan, akkor a medián pontosan a középső érték, ha pedig páratlan, akkor a középső két érték átlaga. A kiszámítására szolgáló Excel függvény:

MEDIÁN(blokkhivatkozás)

Módusz

A **módusz** az adathalmazban előforduló leggyakoribb érték. A kiszámítására szolgáló Excel függvény:

MÓDUSZ(blokkhivatkozás)

A módusz az adathalmaz sűrűsödési pontja. Ha több azonos gyakoriságú elem van az adathalmazban, akkor a MÓDUSZ függvény ezek közül a legkisebbet adja vissza, ha pedig az adathalmazban nincs két azonos érték, akkor a #HIÁNYZIK! hibaüzenetet kapjuk.
Szimmetrikus eloszlások esetén az átlag, medián és módusz közel esik egymáshoz. A statisztikai elemzés során célszerű mindhárom középértéket meghatározni. Az alábbi adathalmaz esetében például ha csak az átlagot számoljuk ki, kifejezetten megtévesztő következtetésekre juthatunk. Az átlagot egy-egy szélsőséges adat nagyon erősen befolyásolja!

Sorszám	Adat
1	2,0
2	2,3
3	3,0
4	2,0
5	2,1
6	12,0
Átlag	**3,9**
Medián	2,3
Módusz	2,0

Szórás

A **szórás** a matematikai statisztika alapvetően fontos, az adatok szóródására jellemző mérőszáma, az adatok átlagtól való, az adatok átlagtól való átlagos eltérésének közelítő értéke[1].
A szórás meghatározására az Excel két függvényt tartalmaz:

Függvény	Képlet	Jelentés
Korrigált empirikus szórás: **SZÓRÁS (blokkhivatkozás)**	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	A statisztikai sokaság szórását mintából becsüljük.
Tapasztalati szórás **SZÓRÁSP (blokkhivatkozás)**	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$	A sokaság minden adata (teljes populáció) ismert.

ahol $\mathbf{x_i}$ az i-dik adat, **n** az adatok száma, $\bar{x}$ pedig az adatok átlaga.

[1] A szórás képletében az átlagtól való eltérések négyzetét összegezzük. Négyzetre emelés nélkül a számláló nulla lenne, hiszen az adatok átlagtól való eltéréseinek összeg – $\sum_{i=1}^{n}(x_i - \bar{x})$ – mindig nulla.

A SZÓRÁS függvényben használt képlet a statisztikai sokaság elméleti szórásának torzítatlan becslése.

Megjegyzés:

Az Excel az adattartományban előforduló logikai értékeket (IGAZ vagy HAMIS), és a szöveget a függvény értékének kiszámítása során figyelmen kívül hagyja. Ha a logikai értékeket és a szöveget is számításba szeretnénk venni, használjuk a **SZÓRÁSA** függvényt, amely az **ÁTLAGA** függvényhez hasonlóan veszi figyelembe a nem numerikus adatokat.

Variációs együttható

A gyártási feladatokban a szórást úgy értelmezhetjük, mint egy gép működése során fellépő véletlen hatások okozta átlagos hibát. Bármennyire kiváló műszaki paraméterekkel rendelkezik az adagoló gép, számolnunk kell azzal a ténnyel, hogy a tejes dobozba sohasem kerül hajszálpontosan 1 liter tej. Az azonban már egyáltalán nem mindegy, hogy a várt 1 literhez képest a töltési hiba 1 milliliter, vagy 1 deciliter. Az utóbbi esetben a töltőgép szisztematikus hibájára kell gyanakodnunk. A **relatív szórás,** vagy más néven **variációs együttható** megmutatja, hogy a szórás hány százaléka az átlagnak. Meghatározásához nincs függvény, értéke egyszerű képlettel számolható.

Sorszám	A csoport	B csoport
1	2	3
2	5	2
3	2	3
4	5	3
5	4	3
6	2	2
7	2	3
8	2	4
9	4	4
10	3	5
11	2	3
12	5	3
Átlag	**3,17**	**3,17**
Módusz	**2,00**	**3,00**
Medián	**2,50**	**3,00**
Szórás	**1,28**	**0,80**
Relatív szórás	**40%**	**25%**

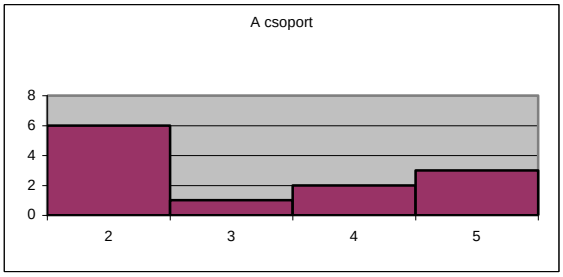
Feladat

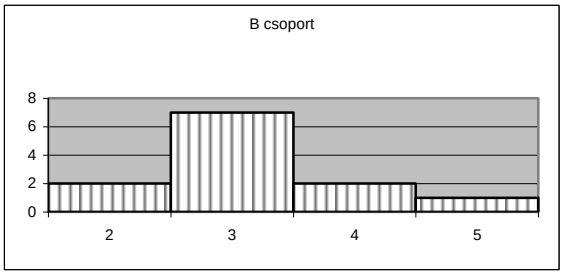
Két tanulócsoport matematika tárgyból elért eredményét szeretnénk összehasonlítani. Az összehasonlításhoz számítsuk ki mindkét csoportra az átlagot, móduszt, mediánt, szórást és a variációs együtthatót.

Az átlagok alapján azt mondhatjuk, hogy mindkét csoport eredménye matematikából közepes. A szórás és a relatív szórás kiszámítása után azonban ennél többet is mondhatunk. Az A csoportban az átlaghoz képest 40% a jegyek szórása, ami arra utal, hogy a csoport a matematikai tehetség dolgában enyhén szélsőséges. Ezzel szemben a B csoport teljesítménye egyenletesen közepes. (Kérdés, hogy melyik csoportban ki tanított?)

Gyakoriság

Az A és B csoport összehasonlítását megkönnyíti egy grafikon, amelyen a jegyek előfordulásának gyakoriságát ábrázoljuk:





Nagyobb méretű adathalmaz esetén a gyakorisági grafikon elkészítését eléggé megnehezítené, ha kézzel kellene megszámolnunk a jegyeket. A számlálást az Excel GYAKORISÁG tömb-függvényével automatizálhatjuk:

GYAKORISÁG(adat tömb, csoport tömb)

A függvény paraméterei:
adat tömb: az adatokat tartalmazó tartomány
csoport tömb: az előforduló értékeket tartalmazó tömb.

Jegy	A	B
2	6	2
3	1	7
4	2	2
5	3	1
Az adatok száma	**12**	**12**

Jegy	A	B
2	=GYAKORISÁG(A;A16:A19)	=GYAKORISÁG(B;A16:A19)
3	=GYAKORISÁG(A;A16:A19)	=GYAKORISÁG(B;A16:A19)
4	=GYAKORISÁG(A;A16:A19)	=GYAKORISÁG(B;A16:A19)
5	=GYAKORISÁG(A;A16:A19)	=GYAKORISÁG(B;A16:A19)
Az adatok száma	**=SZUM(B16:B19)**	**=SZUM(C16:C19)**

Megjegyzés:

1. A tömbfüggvény egy tartományt ad vissza. A függvény beillesztése előtt a teljes tartományt ki kell jelölni, és a műveletet a CTRL+SHIFT+ENTER billentyű kombinációval kell befejezni.
2. Nagyobb adathalmazzal könnyebb számolni, ha előzetesen elnevezzük a taromány. A példában a csoportok adatait tartalmazó blokkoknak az **A** és **B** nevet adtuk.

Konfidencia intervallum

A statisztikai módszerek jellegéből adódik, hogy a minta alapján kiszámított eredmény minden esetben csak becsült, közelítő értéke a teljes statisztikai sokaságra vonatkozó elméleti értéknek. Azt mondhatjuk, hogy az eredmény „egy bizonyos valószínűséggel" tekinthető igaznak. Ha egy minta alapján azt kaptuk, hogy a magyar lakosság naponta átlagosan 3 órát tölt el TV-nézéssel, akkor azt is hozzá kell tennünk, hogy ez az eredmény csak becslés. Abban az esetben, ha a minta normális eloszlású statisztikai sokaságból származik, a becsült átlagra a

MEGBIZHATÓSÁG(alfa, szórás, minta_elemszám)

Excel függvénnyel egy általunk előre megadott valószínűséggel meghatározhatjuk a mintából számított átlag „hibáját". A függvény paraméterei:

alfa: a tévedés valószínűsége, szignifikancia szint
szórás: a sokaság ismert szórása

Tegyük fel, hogy egy 25 elemű mintát vettünk egy olyan sokaságból, amelynek szórása ismert.

	G	H	
9	**A mintából számolt átlag**	2309	
10	**Ismert szórás**	578	
11	**Minta elemszám**	25	
12	**Megbízhatóság**	226,5717	**=MEGBÍZHATÓSÁG(0,05;H9;H10)**

A kapott eredmény jelentése:

$$\mathbf{P(|m-2309|>226,5717)=0,05}$$

ahol **m** az ismeretlen várható érték. A **P** valószínűség az Excel függvény **alfa** paramétere, a szignifikancia szint. Az ellentett esemény valószínűsége: 1-alfa (esetünkben 95%) a biztonság valószínűsége, vagy más néven megbízhatósági szint. Az eredményt a következőképpen is értelmezhetjük: annak valószínűsége, hogy az ismeretlen várható érték 2309-226,57 és 2309+226,57 közé esik, 0,95.

$$\mathbf{P(2082,43<m<2535,57)=0,95}$$

Az 5%-os szignifikancia szinthez tartozó a konfidencia intervallum: (2082,43 , 2535,57).