Jelenkori társadalmi és gazdasági folyamatok, (2018) XIII. évfolyam, 1–2. szám, pp. 77–82.

EXCEL ALKALMAZÁSA NORMÁLIS ELOSZLÁS TESZTELÉSÉRE SHAPIRO-WILK PRÓBÁVAL

Hampel György

Absztrakt: A Microsoft Excel Visual Basic for Application programok segítségével nagyméretű adathalmazok szűrésével kapott adatsorain képes programozottan megismételni a számításokat. Több statisztikai programmal tudnánk a Shapiro-Wilk próbát alkalmazni normális eloszlás ellenőrzésére, de programozható módon erre csak az Excelt használhatjuk. Az Excel táblázatkezelő program segítségével felhasználóbarát kezelőfelületet alakítunk ki a számítások automatizált elvégzéséhez.

Abstract: Microsoft Excel can repeatedly execute programmed calculations on filtered large data sets with the help of Visual Basic for Application. There are several statistical applications to check normality with Shapiro-Wilk test, but only Excel can be used to do this in a programmable way. We have created a user-friendly interface with the help of Excel spreadsheet to perform automatic calculations.

Kulcsszavak: Excel, VBA, statisztikai kiértékelés

Keywords: Excel, VBA, statistical evaluation

1. Bevezetés

Táblázatkezelő programokat olyan esetekben alkalmazunk, amikor a táblázatos megjelenítésen túl számított értékek képzése is szükséges. Az újabb verziók szolgáltatásai már támogatják a programozhatóságot is a Visual Basic for Application fejlesztői környezettel. Akár olyan fájlokat is készíthetünk, melyek csak saját fejlesztésű programokat, függvényeket tárolnak. Ezek akkor válnak hasznossá, amikor gyakori ismételt alkalmazásukra van szükségünk. Ezért célszerű úgy elkészítenünk egy bonyolult számítási folyamat eredményét biztosító eszközt, hogy az újból felhasználható legyen egy másik feladat során is. Ilyen lehet egy pénzügyi számításokat tartalmazó fájl (Zsótér, 2017), de egy komplexebb, folyamatok modellezését megvalósító alkalmazás (Fabulya, 2017) is.

Normális eloszlás ellenőrzősére gyakran szükségünk lehet adatok statisztikai kiértékelése során. Numerikus adattípus esetén erre az egyik leghatékonyabb módszer a Shapiro-Wilk próba. Mivel bonyolult számítást igényel a végrehajtása, célszerű olyan univerzális módú kialakítása, mely könnyen újból használható eltérő elemszámú adatsorok mellett is.

2. Anyag és módszer

A kutató, fejlesztő munka során az Excel 2010 verzióját alkalmaztuk. Így egyszerre biztosított volt a számításokhoz szükséges adatok tárolása munkalapokon, valamint azok az Excel függvények is, melyek segítették a kiértékeléshez szükséges számítások elvégzését. A legfontosabb indok az Excel alkalmazása mellett mégis az volt, hogy a Visual Basic for Application alkalmazás segítségével a felhasználó számára könnyen kezelhető, sőt, programozható felületet alakíthatunk ki, mellyel szükség esetén a Shapiro-Wilk próba újból elvégezhető csupán az adatok megadásával.

2.1. Az Excel VBA alkalmazása

A táblázatkezelő programnak azt az alapvető lehetőségét használjuk ki, hogy a számítások kiinduló adatait tartalmazó cellák módosításával az eredmények cellái automatikusan újraszámítódnak. Így kialakíthatunk egy olyan számolótáblát, melyben a számítások és a kiinduló adatok is szerepelnek. A cél az, hogy ezt olyan univerzális módon alakítsuk ki, hogy minden lehetséges körülmény mellett biztosítsa a számítások elvégzését. Ez a rész a legfontosabb eleme, magja minden további olyan munkánknak, mellyel a felhasználóbarát kezelési módot biztosítjuk. Ehhez a Visual Basic for Application (továbbiakban VBA) szolgáltatásait vesszük igénybe.

A felhasználó legegyszerűbben, további magyarázat nélkül akkor tud egy újabb szolgáltatást alkalmazni, ha az a számára megszokott módon viselkedik és jelenik meg. Ilyen egy Excel függvény paramétereinek megadásához hasonló felület, melynek kialakításához szükségünk van VBA programozói ismeretekre minimális szinten, melyek a következők:

- függvény létrehozása,
- cellák értékének kiolvasása,
- cellák adatainak módosítása

2.2. A Shapiro-Wilk próba

A Shapiro-Wilk próba (Shapiro–Wilk, 1965) a leghatékonyabb próba annak tesztelésére, hogy származhatott-e a minta adatsora egy normális eloszlású sokaságból, populációból. A többi próbával összevetve a Shapiro-Wilk próba még kis elemszámú minta esetén is erősnek tekinthető (Thode, 2002).

A próba alkalmazása során kiszámítjuk a minta adatsorából az alkalmazott statisztikai függvény (I) értékét (W), majd ezt összehasonlítjuk a minta elemszámától (n) és az elsőfajú hibavalószínűségtől (ε) függő kritikus értékkel (W_{kr}). Amennyiben $W > W_{kr}$, akkor feltételezhetjük a normalitást az adott hibavalószínűség mellett, különben pedig nem.

$$W = \frac{(\sum_{i=1}^{n} a_{ni} x_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (1)$$

ahol:

x_i – a nagyság szerint rendezett minta i. eleme (i=1,…, n),

$\bar{x}$ – a mintaelemek átlaga, (számtani közép),

a_{ni} – együtthatók (i=1,…, n).

Mivel az együtthatók kiszámítása igen nagy számításigényű, ezért ezeket nem szokták újból kiszámítani, hanem táblázatból vehetők. Ráadásul az együtthatók sorozata szimmetrikus olyan értelemben, hogy a sor elején és végén lévő elemek csak előjelben térnek el egymástól. Ezért szokásos a sornak csak a pozitív tagjait megadni. Páratlan elemszám esetén a középső együttható értéke nulla. A táblázat egy részlete látható az *1. táblázatban*. A teljes táblázat n=5 és n=50 közötti minta

elemszám mellett, valamint i=1 és i=25 közötti értékekre tartalmazza a pozitív együtthatókat.

1. táblázat: **Az együtthatók táblázatának részlete**

i	n		
	5	6	7
1	0,6646	0,6431	0,6233
2	0,2413	0,2806	0,3031
3	0	0,0875	0,1401
4			0

Forrás: A szerző saját szerkesztése.

3. Eredmények és értékelésük

Mivel célunk egy olyan felület kialakítása az Excelben, ahol a felhasználó csak a saját adatait tarthassa látókörében, ezért úgy kell kialakítanunk a szükséges objektumokat, hogy a számításokért felelős munkalap külön fájlban, rejtett maradhasson, míg a feldolgozandó adatokhoz a saját Excel munkalapjain férjen hozzá a felhasználó. A tervezési fázisban ennek megfelelően így a következő részfeladatok adódnak:

- munkalap kialakítása a számítások elvégzéséhez,
- kezelői felület elkészítése.

3.1. Számítások munkalapja

Mivel a számítások munkalapja a felhasználó számára rejtett, ezért itt nincs szükség olyan beállításokra, melyek az igényes megjelenést és használatot biztosítják. Így célszerű két munkalapot kialakítani. Az egyiken szerepeltetve a számításban résztvevő adatokat és számításokat, míg a másikon a felhasználótól érkező adatok tárolódnak. Ennek azért van jelentősége, mert így lehetőségünk van arra, hogy csak azokat az adatokat vonjuk be a számításokba, melyeknek van értelme. Így például a szöveges típusú adatoktól megtisztíthatjuk a feldolgozást. Ehhez elegendő csak az Excel munkalap függvényeit alkalmazni:

```
=HA(SZÁM(Adatok!A1);Adatok!A1;"")                    (2)
```

A (2) képlet mutatja, hogy az Excelben az *Adatok* munkalapról az *A1* cella tartalma csak akkor kerül át a számítások munkalapjára, ha ez a cella számot tartalmaz. E képlet másolatai biztosítják, hogy a rejtett munkalap B4..B53 tartományba helyezzük az adatokat. Ez a tartomány lesz a számítások alapját biztosító legfeljebb 50 adat tárhelye.

Szintén tárolni kell bemenő adatként az elsőfajú hibavalószínűség (ε) értékét is. Ez az E3 cellában lesz, viszont csak 0,01 és 0,05 lehet az értéke. Az ezektől eltérő értékeket úgy kerülhetjük el, ha a felhasználó egy legördülő listából csak ezeket az értékeket választhatja ki.

A számításokhoz szükséges, de már nem a felhasználótól érkező további adatok az együtthatók (a_{ni}), melyek a B97..AU121 tartományban érhetők el. Egy oszlopban szerepelnek egy adott elemszámú mintához a pozitív együtthatók. További adatként már csak a kritikus értékeket (W_{kr}) kell tárolnunk az E126..F171 tartományban, vagyis két oszlopban a 0,01 és 0,05 elsőfajú hibavalószínűséghez tartozóan.

A számítások elvégzéséhez és a részeredményekhez célszerű egy segéd táblázatot kialakítani a munkalapon (*2. táblázat*). A táblázatban egy 12 elemű adatsor esetén láthatjuk a *nagy* és *kicsi* oszlopokban a nagyság szerint rendezett adatsor azon párokba rendezett elemeit egy sorban, melyeket ellenkező előjelű a_i együtthatós szorzatával kell összegezni. Tehát a segédtáblázatunk az (*1*) képlet számlálójának kiszámítását segíti munkalap függvények alkalmazhatóságával a (*2*) képlet szerint.

2. táblázat: **Számítások segéd táblázata**

i	nagy	kicsi	di	ai	ai×di
1	3,71	3,42	0,29	0,5475	0,158775
2	3,69	3,45	0,24	0,3325	0,079800
3	3,64	3,47	0,17	0,2347	0,039899
4	3,63	3,49	0,14	0,1586	0,022204
5	3,61	3,51	0,10	0,0922	0,009200
6	3,59	3,54	0,05	0,0303	0,001515
7					
8					

Forrás: A szerző saját szerkesztése.

A táblázat első sorában a legnagyobb x_1 és legkisebb x_{12} adatokat ellenkező előjelű együtthatóval kell szorozni:

$$a_1x_1 + a_{12}x_{12} = a_1x_1 - a_1x_{12} = a_1(x_1 - x_{12}) = a_1d_1 \qquad (3)$$

Így munkalap függvényekkel képezhető egy cellában az (*1*) képlet számlálójának értéke a 2. *táblázat* utolsó oszlopában kapott értékei összegének négyzeteként.

A nagy és kicsi nevű oszlopokban a nagyság szerint rendezett adatsor a (*4*) és (*5*) képletekkel adódik, míg az együtthatók az (*6*) képlettel.

=HA(G4<=E$4/2;NAGY(B$4:B$53;G4);"") (4)
=HA(G4<=E$4/2;KICSI(B$4:B$53;G4);"") (5)
=HA(G4<=E$4/2;INDEX(B$97:AU$121;G4;E$4-4);"") (6)

Az *E4* cellában képeztük a minta elemszámát. A *HA* függvény mindhárom képletben azt biztosítja, hogy csak akkor jelenjen meg egy eredmény a segédtáblázatunk adott sorában, amíg ez értelmezhető. Mivel a segédtáblázatunk első adata a *G4* cellában a sorszám értéke, ezért ennek kisebb vagy egyenlőnek kell lennie az elemszám felénél, hiszen soronként két adatát dolgozzuk fel az adatsornak.

A *NAGY* és *KICSI* függvények képzik a *B4..B53* tartomány rendezetlen adatsorából a rendezett adatsor adott sorszámú elemét. Az együtthatókat az *INDEX* függvény eredményezi a *B97..AU121* tartományból két index, a sor száma és a minta elemszáma alapján.

Az (*1*) képlet nevezője szintén könnyen képezhető munkalap függvények alkalmazásával. Persze a számláló és a nevező ismeretében egy osztással megkapjuk a statisztikai függvény értékét (*W*). Már csak a kritikus értékre (W_{kr}) van szükségünk ahhoz, hogy a normalitásra választ adhassunk. Ezt szintén az *INDEX* függvénnyel kaphatjuk meg, ahol az indexek az elsőfajú hibavalószínűség és a minta elemszámából kaphatók meg. A végeredmény *HA* függvénnyel kapható meg akár logikai típusúan (normalitás feltételezhető vagy sem), de szövegesen is egy cellában a (*7*) képlettel.

=HA(E9<E10;"Nem normális eloszlású";"Normális eloszlású") (7)

A (*7*) képletben az E9 és E10 cellákban a statisztikai függvény értéke, valamint a kritikus érték találhatók. Az így kapott végeredményt kell az *Adatok* munkalapon is szerepeltetni, hiszen a felhasználó csak ezt a munkalapot látja.

3.2. Felhasználói felület kialakítása

Mivel az eddig elkészített munkalapon csak Excel munkalap függvényeket alkalmaztunk, ezért az *Adatok* munkalapon bekövetkező minden adatváltozásra a számítások aktualizálódnak automatikusan. Ez azt jelenti, hogy elegendő csak ezt a munkalapot olyan felhasználói felületté alakítani, ahol az adatbevitel és az eredmény megjelenítése valósul meg. A kész felület látható az *1. ábrán.*

1. ábra: **A próba felhasználói felülete**

Shapiro-Wilk próba

Adatok		
3,42	Elsőfajú hibavalószínűség (ε)	0,05
3,49		
3,54	A próba eredménye:	
3,71	Normális eloszlású	
3,47		
3,64		
3,69		
3,45		
3,59		
3,63		
3,61		
3,51		

Forrás: A szerző saját szerkesztése.

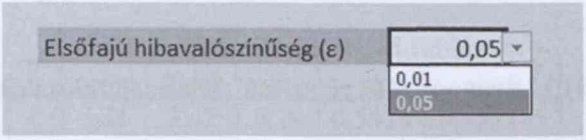
A színek kialakításán túl a felület biztonságos működéséhez a következő beállítások szükségesek:

- munkalap védelme,
- adat érvényesítés legördülő listával.

A munkalap védelmével tudjuk elérni, hogy az adatok celláinak kivételével tartalommódosítást ne végezhessen a felhasználó. A színek is jelzik, hogy mely cellák nem zároltak.

A legördülő listás adatérvényesítés (*2. ábra*) biztosítja, hogy az elsőfajú hibavalószínűség értéke csak 0,01 vagy 0,05 lehessen.

2. ábra: **Legördülő listás adatérvényesítés**



Forrás: A szerző saját szerkesztése.

4. Összegzés

A Shapiro-Wilk próba számításait és a végrehajtáshoz szükséges adatokat táblázatosan tartalmazó fájlt korlátlanul újból felhasználhatjuk. Ezt akár a Visual Basic for Application szolgáltatással programozott módon is megtehetjük. Ehhez csak arra van szükségünk, hogy a feldolgozandó adatokat tartalmazó Excel fájlon túl megnyitott legyen a számítások Excel táblája is, valamint az esetleges automatizált többszöri újraszámításokhoz egy minimális szintű VBA proramozói ismeretre (Fabulya, 2017).

A lehetőségeinket tovább bővíthetjük, ha a Shapiro-Wilk próba számításainak Excel fájljából bővítményt alakítunk ki. Ez azt eredményezné, hogy a bővítmény telepítésével az adott számítógépen már a fájl megnyitására sem lesz szükségünk a használat érdekében, mert a bővítmények automatikusan megnyílnak az Excel indításakor. Így az Excellel végezhető lehetőségeinket gazdagíthatjuk.

Irodalomjegyzék

Fabulya Z. (2017): Hőkezelési folyamatok összehangolása Excel VBA szolgáltatásokkal. *Jelenkori társadalmi és gazdasági folyamatok* 12 (4): 19–25.

Shapiro S. S., Wilk M. B. (1965): An analysis of variance test for normality (complete samples). *Biometrika*, 52 (3/4): 591–611.

Thode H. C. (2002): *Testing for normality*. Statistics , a series of textbooks and monographs (Book 164). CRC press, Bosa Roca USA.

Zsótér B. (2017): Financial planning in connection with accomodation development in a sport centre. *Quaestus multidisciplinary research journal* 4 (11): 172–177.