SZÁMÍTÁS AUTOMATIZÁLÁSA SHAPIRO-WILK PRÓBÁHOZ ROYSTON ALGORITMUSA SZERINT EXCEL SZÁMOLÓTÁBLÁVAL

Fabulya Zoltán

Absztrakt: Egy statisztikai sokaság normális eloszlásának tesztelésére számolótáblát alakítottunk ki. A legmegbízhatóbb Shapiro-Wilk próbának Royston szerinti kiterjesztése tette lehetővé, hogy 50 elemszám feletti minta esetén is elvégezhető legyen a vizsgálat. Az Excel számolótábla biztosítja a számítás automatizálását egyrészt azzal, hogy az adat megváltozásakor újraértékelést végez, másrészt programozási lehetőséget biztosít az adatok automatizált lecseréléséhez és az eredmények kigyűjtéséhez a Visual Basic for Applications szolgáltatással. Az elkészített felhasználói felület több minta adatának feltöltését támogatja. A kiértékelő program minden mintán elvégzi a számítást, és egyetlen táblázatba rendezi az eredményeket.

Abstract: We created a spreadsheet to test the normal distribution of a statistical population. The extension of the most reliable Shapiro-Wilk test according to Royston made it possible to perform the test even for samples with more than 50 elements. The Excel spreadsheet ensures the automation of the calculation, on the one hand, by performing a recalculation when the data changes, and on the other hand, it provides a programming option for the automated replacement of the data and the collection of the results with the Visual Basic for Applications service. The created user interface supports the uploading of the data of several samples. The evaluation program performs the calculation on each samples and arranges the results in a single table.

Kulcsszavak: Shapiro-Wilk próba, Royston algoritmus, Excel, statisztika, VBA programozás

Keywords: Shapiro-Wilk test, Royston algorithm, Excel, statistics, VBA programming

1. Bevezetés

A matematikai statisztikában sok próba alkalmazhatóságának feltétele, hogy normális eloszlású sokaságból származzon a kiértékelendő minta (Obádovics, 2020). Ezért a normalitás ellenőrzésére számos módszert fejlesztettek ki. Ezek között a legmegbízhatóbb a Shapiro-Wilk próba, mely kisebb elemszámú minta esetén is alkalmazható (Thode, 2002). A végrehajtása viszont táblázatok adatain alapul, és csak 3 és 50 közötti elemszámra készültek táblázatok (Shapiro–Wilk, 1965). Royston dolgozott ki algoritmust arra, hogy akár 5000 méretű minta esetén, táblázatok adatai nélkül elvégezhessük a számításokat a normalitás ellenőrzésére (Royston, 1995).

A statisztikai programok a Shapiro-Wilk próbát jellemzően Royston algoritmusa alapján hajtják végre, de mivel nem biztosítanak programozással automatizálható technikát, így a több mintán alapuló kiértékelési sorozat monoton ismétlődő tevékenységgel jár. Excel számolótáblát kialakítva elegendő csak a minta adatait feltölteni a kiértékelés elvégzéséhez az adatváltozások utáni automatikus újraszámítások miatt. Amennyiben több mintával rendelkezünk, akkor a Visual Basic for Applications (VBA) szolgáltatással készíthetünk programot a minták váltásához és az eredmények táblázatba szervezett kigyűjtéséhez (Zimmerman, 1996).

A cikk bemutatja Royston algoritmusának Microsoft Excel környezetben, számolótáblán munkalapfüggvényekkel, formulákkal automatikusan kiszámítódó

ISSN 1788-7593 (Nyomtatott) DOI: https://doi.org/10.14232/jtgf.2022.3-4.67-74
ISSN 2676-9867 (Online)

megvalósítását a Shapiro-Wilk próba végrehajtásához 3 és 5000 közötti elemszámú mintához, valamint több minta esetén a próba ismételt végrehajtására és az eredmények táblázatos megjelenítésére kifejlesztett VBA programot.

2. Anyag és módszer

2.1. Royston kiterjesztése a Shapiro-Wilk próbához

A Shapiro-Wilk próba numerikus adattípus esetén alkalmazható egy statisztikai sokaság normális eloszlásának ellenőrzésére. A legerősebb normalitást tesztelő próba, mely kis elemszám mellett is megbízható. Hátránya, hogy legfeljebb 50 elemű minta esetén alkalmazható, valamint nem számítható vele szignifikanciaszint, s így csak kritikus tartománnyal dönthetünk. A Shapiro-Wilk próba statisztikai függvényének (W, *1-2 képlet*) kiszámításához szükséges együtthatók (a_i) és a kritikus tartomány határa is táblázatban adottak. Emiatt a határok nem voltak ismertek tetszőleges elsőfajú hibavalószínűség (ε) esetén, csak a leggyakoribb alkalmazott értékeknél (Shapiro–Wilk, 1965).

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

$$\bar{x}=\frac{1}{n}\cdot\sum_{i=1}^{n} x_i \qquad (2)$$

ahol:
W – a Shapiro-Wilk próba statisztikai függvénye
n – a minta elemszáma
x_i – a minta elemei növekvő rendezettségben: $x_i \leq x_j \ (i<j)$
a_i – együtthatók

Az együtthatók értéke a minta elemszámától is függ, s mivel $a_{n+1-i}=-a_i$, ezért a kisebb táblázat érdekében csak a pozitív együtthatók szerepelnek a táblázatban, Emiatt a statisztikai függvény eredeti képlete (*3*):

$$W=\frac{\left(\sum_{i=1}^{n/2} a_i \cdot (x_{n+1-i}-x_i)\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (3)$$

Royston terjesztette ki a próba végrehajthatóságát először 2000, majd 5000 elemszámú mintára (Royston, 1982; Royston, 1995). Algoritmusa táblázatok nélkül teszi lehetővé a próba kiértékelését az együtthatók közelítő értékét eredményező képletekkel (*4-13*).

$$a_n=-2{,}706056\cdot y^5+4{,}434685\cdot y^4-2{,}07119\cdot y^3-0{,}147981\cdot y^2+0{,}221157\cdot y+c_n \qquad (4)$$

$$a_{n-1} = -3{,}582633 \cdot y^5 + 5{,}682633 \cdot y^4 - 1{,}752461 \cdot y^3 - 0{,}293762 \cdot y^2 - 0{,}042981 \cdot x + c_{n-1} \quad (5)$$

$$a_1 = -a_n \quad (6)$$

$$a_2 = -a_{n-1} \quad (7)$$

$$a_i = \frac{m_i}{\sqrt{\epsilon}} \quad (2 < i < n-1) \quad (8)$$

ahol:

$$y = \frac{1}{\sqrt{n}} \quad (9)$$

$$c_i = \frac{m_i}{m} \quad (i = 1, 2, \dots, n) \quad (10)$$

$$m_i = \Phi^{-1}\left(\frac{i - 0{,}375}{n + 0{,}25}\right) \quad (i = 1, 2, \dots, n) \quad (11)$$

$\Phi(x)$ – a standard normális eloszlás eloszlásfüggvénye

$$m^2 = \overline{m}^T \cdot \overline{m} = \sum_{i=1}^{n} {m_i}^2 \quad (12)$$

$$\epsilon = \begin{cases} \dfrac{m^2 - 2{m_n}^2}{1 - {a_n}^2} & ha\ n \leq 5 \\ \dfrac{m^2 - 2{m_n}^2 - 2{m_{n-1}}^2}{1 - 2{a_n}^2 - 2{a_{n-1}}^2} & ha\ n > 6 \end{cases} \quad (13)$$

n – a minta elemszáma

Royston algoritmusának további képletei teszik lehetővé, hogy táblázatok nélkül, tetszőleges elsőfajú hibavalószínűség mellett dönthessünk a normalitásról a próba szignifikanciaszintje (p) közelítő értékének kiszámításával (*14-17*).

$$p = 1 - \Phi(z) \quad (14)$$

ahol

$$z = \frac{\ln(1 - W) - \mu}{\sigma} \quad (15)$$

$$\mu = 0{,}0038915 \cdot \ln^3 n - 0{,}083751 \cdot \ln^2 n - 0{,}31082 \cdot \ln n - 1{,}5861 \quad (16)$$

$$\sigma = e^{0{,}0030302 \cdot \ln^2 n - 0{,}082676 \cdot \ln n - 0{,}4803} \quad (17)$$

n – a minta elemszáma

A szignifikanciaszint (p) ismeretében akkor dönthetünk úgy, hogy normális eloszlásúnak tekinthető a statisztikai sokaság az elsőfajú hibavalószínűség (ε) mellett, ha a $p > \varepsilon$ feltétel teljesül.

2.2. Az Excel programozási lehetősége

A Microsoft Excel program már azzal is nagymértékben megkönnyíti a számítások végrehajtását, hogy az egyes cellákban elhelyezett formulák ismételten kiértékelődnek, amikor egy hivatkozott cella értéke megváltozik. A Visual Basic for

Applications szolgáltatással még programokat is kialakíthatunk, melyekkel tovább automatizálhatjuk a számítási feladatot (Matteson, 1995). Így például a programunk változtathatja meg a cellák tartalmát a programunk utasításai szerint. A programok végrehajtását a felhasználó kezdeményezheti az eseményvezérelt programozási technikának köszönhetően. Legegyszerűbben a munkalapon lévő parancsgombra klikkeléssel indítható egy program, de más vezérlők is alkalmazhatók. Egy cella értékének beállítása léptető gombbal is megoldható úgy, hogy a beállításoknak megfelelően csak egy értelmezhető tartományban alakulhat az érték egy megfelelő lépésközzel módosulva. Ekkor a felhasználónak nincs lehetősége hibás adat beállítására. A munkalap celláinak zárolásával elérhető, hogy a felhasználó csak azokat a cellákat módosíthassa, melyeket nem zároltunk. Így a kialakított formulák és fix tartalmak még véletlenül sem írhatók felül.

3. Eredmények és értékelésük

A kitűzött célok megvalósítása két részfeladatra bontható:

- A Royston algoritmus számításait egy mintán elvégző számolótábla kialakítása egy Excel munkalapon.
- VBA program készítése, mely automatikusan felismeri, hogy hány minta érhető el a normalitás ellenőrzésére, és táblázatot készít minden mintához tartozóan a kiértékelés eredményéről.

3.1. Számolótábla kialakítása a Royston algoritmushoz

A számítások elvégzéséhez meg kell tervezni, hogy a szükséges adatok és a számított eredmények az Excel munkafüzetben hogyan helyezkedjenek el, s hogy a felhasználó hogyan tudja kényelmesen kezelni. Egyetlen munkalapot kialakítva, az alábbi szempontok érvényesülnek:

- A munkalap baloldalán az adatsorokhoz szükséges, oszlopos elrendezést igénylő részek találhatók (*1. ábra*). Az A oszlopban lehet elvégezni az adatok megadását begépeléssel vagy más módon, míg a C-F oszlopokban számított értékek adódnak:
  - Adat sorszáma (i).
  - Adatok növekvő rendezettségben (x_i).
  - Részeredmények az együtthatókhoz (m_i).
  - Együtthatók (a_i).

1. ábra: **Számolótábla az oszlopban megjelenő adatokkal**

	A	B	C	D	E	F
1	Adatok		i	x_i	m_i	a_i
2	96		1	80	-1,6729	-0,5358
3	81		2	81	-1,1619	-0,3327
4	111		3	81	-0,8484	-0,2408
5	103		4	83	-0,602	-0,1709

Forrás: kutatás adatai alapján a szerző saját szerkesztése.

- A munkalap jobboldalán egyetlen számmal leírható értékek szerepelnek (*2. ábra*). Itt jelennek meg a számítást segítő részeredmények (μ, σ, Z) és a végeredmények (W, p, Normalitás), valamint az elsőfajú hibavalószínűség (ε) adható meg begépeléssel.

2. ábra: **Számolótábla jobboldalának egy részlete**

W	0,79477
μ	-2,8687
σ	0,51047
Z	2,51733
p	0,00591
ε	0,05
Normalitás	HAMIS

Forrás: kutatás adatai alapján a szerző saját szerkesztése.

- Színek segítik az adatok értelmezését:
  - Sárga alapú a begépeléssel változtatható adat.
  - Sötétkék háttéren jelennek meg a fontosabb végeredmények.
- A munkalapon a felhasználó csak azokat a cellákat módosíthatja, melyek begépelendő adatok. Ezek a sárga háttérről ismerhetők fel könnyen. A többi cella zárolt.
- A minta elemszáma legfeljebb 5000 lehet. Emiatt ennyi sorban kellett formulákat is elhelyezni a munkalapon. Viszont eredmények csak akkor jelennek meg, amikor az elemszámot még nem haladtuk meg. Ez oldható meg a HAHIBA() és a HA() munkalapfüggvényekkel.

A használat során elegendő megadni a minta adatsorát és az elsőfajú hibavalószínűséget, míg az eredmények automatikusan megjelennek. A számítások első fázisában az adatsorok (*1. ábra*) kiértékelése történik meg a számítást segítő részeredményekkel (*3. ábra*).

3. ábra: **Részeredmények az adatsorok kiszámításhoz**

n	13
y	0,27735
m^2	10,835
ϵ	12,4085

Forrás: kutatás adatai alapján a szerző saját szerkesztése.

A számítás menete:

- Az elemszám (n) meghatározása: `J2=DARAB(A:A)`. Tehát a `DARAB()` munkalapfüggvény eredményezi az `A` oszlopban lévő adatok számát.

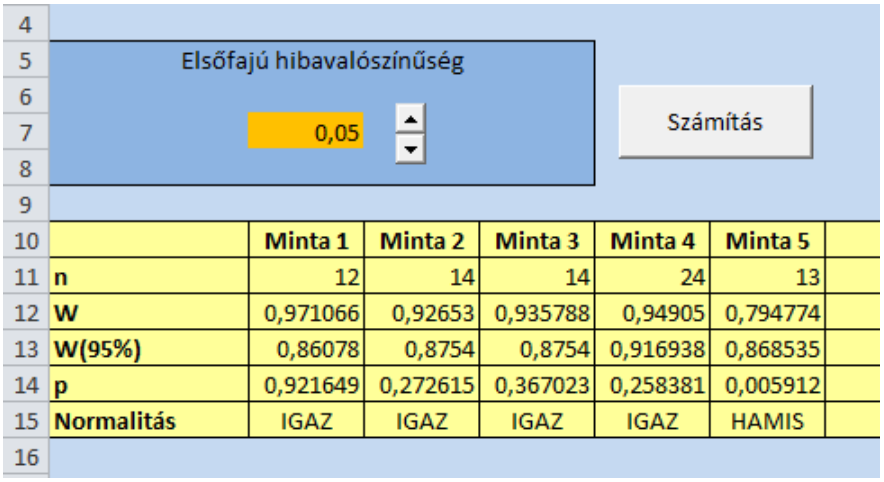
- A sorszámok kialakítása a `C` oszlopban.
- A minta adatai növekvően a `D` oszlopban. Itt a `KICSI()` munkalapfüggvény képzi az adat sorszáma alapján az `A` oszlopból a megfelelő értéket.
- Az E oszlop kiszámítása a (*11*) képlet szerint, melyben a standard normális eloszlás eloszlásfüggvényének inverzét munkalapfüggvénnyel kapjuk: `=HAHIBA(NORM.S.INVERZ((C2-0,375)/($J$2+0,25));"")`
- Az együtthatók (a_i) kiszámítása a (*4-8*) képletekkel történt, részeredmények (y, m^2, ϵ) segítségével a (*9-13*) képletek szerint.
- A statisztikai függvény (W) értéke a (*3*) képlettől eltérően, az Excelben egyszerűbben meghatározható a `KORREL()` munkalapfüggvény értékének négyzeteként adódott, ahol a korrelációt a növekvő adatsor (x_i) és az együtthatók adatsora (a_i) esetén számítjuk.
- A szignifikanciaszint (p) kiszámításához a *2. ábrán* megjelenített részeredmények (μ, σ, Z) számítása is szükséges a (*14-17*) képletekkel.
- Végül a normalitás a $p > \varepsilon$ feltétel teljesülésekor fogadható el. Ez `IGAZ` vagy `HAMIS` kijelzéssel látható a `J14` cellában utolsó eredményként a *2. ábrán*. A cellában kialakított formula: `=J12>J13`.

3.2. Számítások automatizálása a VBA programmal

Több minta kiértékeléséhez arra van szükség, hogy az eredményeket kiszámító számolótáblán lecseréljük a minta adatait. Ezt úgy automatizáljuk, hogy további két munkalap egyike (`Adatok` munkalap) tartalmazza a minták adatait, míg a másik (`Eredmények` munkalap) táblázatosan megjeleníti az eredményeket. A minták adatainak feltöltése tetszőleges technikával megoldható az Excel szolgáltatásait használva úgy, hogy a szomszédos oszlopokba egy-egy minta adatsora kerüljön a bal szélső (`A`) oszloptól kezdve. Az `Eredmények` munkalapon (*4. ábra*) csak az elsőfajú hibavalószínűséget kell megadni léptetőgombokkal, majd a `Számítás` gombra klikkelve kialakul a táblázat az eredményekkel.

Csak a *4. ábrán* látható két vezérlő használható a lapon, a cellák zároltak. Így adat begépelésére nincs lehetőség, csak a léptetőgomb beállításával megvalósítható, hogy az elsőfajú hibavalószínűség 0,01 és 0,2 között legyen 0,01 lépésközzel. Ezzel elkerültük az értelmetlen adatból származó problémákat, nem kell ellenőriznie az adatot a programunknak. A `Számítás` parancsgomb indítja el azt a VBA eljárást (`Public Sub szamitas()`), mely elvégzi a táblázat kialakítását az eredményekkel.

4. ábra: **Az Eredmények munkalap**



	Minta 1	Minta 2	Minta 3	Minta 4	Minta 5
n	12	14	14	24	13
W	0,971066	0,92653	0,935788	0,94905	0,794774
W(95%)	0,86078	0,8754	0,8754	0,916938	0,868535
p	0,921649	0,272615	0,367023	0,258381	0,005912
Normalitás	IGAZ	IGAZ	IGAZ	IGAZ	HAMIS

Forrás: kutatás adatai alapján a szerző saját szerkesztése.

A program a következő részfeladatokból áll:

- Törli a korábbi számítások táblázatát:

```
m = 1
While Cells(10, m + 1) <> ""
  m = m + 1
  For s = 10 To 15
    Cells(s, m) = ""
  Next s
Wend
```

- Áttölti az elsőfajú hibavalószínűség értékét az `Eredmények` lapról a számolótáblára:

```
Sheets("Royston").Cells(13, 10) = Cells(7, 2)
```

- Az `Adatok` lapon lévő minden minta (`m=1, 2, …`) esetén a következőket végzi:
  - Törli a számolótáblán (`Royston` munkalap) lévő korábbi mintát:

```
s = 2
While Sheets("Royston").Cells(s, 1) <> ""
  Sheets("Royston").Cells(s, 1) = ""
  s = s + 1
Wend
```

  - Áttölti a soron következő mintát az `Adatok` munkalapról a számolótáblára (`Royston`):

```
s = 1
Cells(10, m + 1) = "Minta " & m
While Sheets("Adatok").Cells(s, m) <> ""
  Sheets("Royston").Cells(s + 1, 1) =_
                Sheets("Adatok").Cells(s, m)
  s = s + 1
```

```
    Wend
```

- Kigyűjti az eredményeket a számolótábláról az `Eredmények` munkalap táblázatának következő oszlopába:

```
Cells(11, m+1)=Sheets("Royston").Cells(2, 10)
Cells(12, m+1)=Sheets("Royston").Cells(8, 10)
Cells(13, m+1)=Sheets("Royston").Cells(18, 10)
Cells(14, m+1)=Sheets("Royston").Cells(12, 10)
Cells(15, m+1)=Sheets("Royston").Cells(14, 10)
```

4. Következtetések

A Microsoft Excel táblázatkezelő programmal kialakított számolótábla egyszerűen használható lehetőséget nyújt a Shapiro-Wilk próba kiértékelésére a Royston algoritmust használva. Royston kiterjesztésével 50 elemszám feletti mintán is elvégezhető a próba, táblázatokból származó adatok (együtthatók, kritikus határ) nélkül. Több minta esetén a számolótáblát használva automatizálhatjuk a kiértékeléseket, az összes mintára adódó eredmény egyetlen táblázatba gyűjthető VBA program segítségével. Vezérlőkkel felhasználóbarát felületet alakíthatunk ki a számítások elvégzésére, amely alkalmazható lenne piackutatási adatok (Zsótér, 2006, Zsótér–Kaliczka, 2014) és számos egyéb pl. helyi gazdaság fejlesztési kutatás eredményeinek (Zsótér et al., 2020) feldolgozására.

Irodalomjegyzék

Matteson, B. L. (1995): *Microsoft Excel Visual Basic Programmer's Guide*. MicrosoftPress, Washington.

Obádovics J. Gy. (2020): *Valószínűségszámítás és matematikai statisztika*, Scolar Kiadó Kft., Budapest.

Royston, J. P. (1982): Algorithm AS 181: The W Test for Normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31 (2): 176–180. https://doi.org/10.2307/2347986

Royston, P. (1995): Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44 (4): 547–551. https://doi.org/10.2307/2986146

Shapiro, S. S., Wilk, M. B. (1965): An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52 (3/4): 591–611. https://doi.org/10.2307/2333709

Thode, H. C. (2002): *Testing For Normality (1st ed.)*. CRC Press. https://doi.org/10.1201/9780203910894

Zimmerman, M. W. (1996): *Microsoft Office 97 Visual Basic Programmer's Guide*, MicrosoftPress, Washington.

Zsótér B. (2006): Turizmus Mezőhegyesen: a Hotel Nonius bemutatása. In: Gál József (szerk.): *Európai Uniós Kutatási és Oktatási Projektek Napja és Leonardo da Vinci Learn at Work Projekt-találkozó* [European Union Research and Educational Projects Day and Leonardo da Vinci Learn at Work Project Meeting]. Konferencia helye, ideje: Hódmezővásárhely, Magyarország, 2006.10.06 Hódmezővásárhely: Delfin Computer Informatikai Zrt., 2006. Paper CD. 6 p.

Zsótér B., Illés S., Simonyi P. (2020): Model of Local Economic Development in Hungarian Countryside. *European Countryside* 12 (1): 85–98. DOI: 10.2478/euco-2020-0005.

Zsótér B., Kaliczka Renáta (2014): Examinations carried out in relation to the shopping habits and satisfaction of costumers in the shops of Coop Szeged Ltd. *Review of Faculty of Engineering Analecta Technica Szegedinensia,* 8 (1): 38–41.