EXCEL FÜGGVÉNYEK FEJLESZTÉSE SHAPIRO-WILK PRÓBÁHOZ ROYSTON ALGORITMUSA ALAPJÁN

Fabulya Zoltán – Hampel György – Kiss Anita – Béresné Mártha Bernadett

Absztrakt: Kutatásunk céljaként Excel számolótáblán alkalmazható függvényeket fejlesztettünk ki, melyek egy statisztikai sokaság normális eloszlásának tesztelésére alkalmasak. Függvényeink Royston algoritmusát alkalmazzák, mely a normalitás ellenőrzésére legerősebb Shapiro-Wilk próbának a kiterjesztése. Így 4 és 2000 közötti elemszámú minta kiértékelése közelítő számításokkal valósítható meg úgy, hogy szignifikanciaszint kiszámításával dönthessünk a normalitásról, elkerülve a Shapiro-Wilk próba táblázatban adott kritikus értékeinek használatát. Az Excel számolótáblán a kiértékelések automatizálhatók függvények használatával, ezért gyorsabb és kényelmesebb technikát nyújtva, mint a statisztikai programcsomagok. A függvények programozását a Microsoft Excel Visual Basic for Applications szolgáltatása biztosította. Royston képleteit átalakítva olyan függvény is készült, mellyel a próba kritikus értéke adódik tetszőleges elsőfajú hibavalószínűséghez.

Abstract: As the goal of our research, we developed functions that can be used on an Excel spreadsheet, which are suitable for testing the normal distribution of a statistical population. Our functions use Royston's algorithm, which is an extension of the Shapiro-Wilk test, the strongest test for normality. Thus, the evaluation of a sample with between 4 and 2,000 elements can be carried out with approximate calculations so that we can decide on normality by calculating the significance level, avoiding the use of the critical values of the Shapiro-Wilk test given in the table. Evaluations on the Excel spreadsheet can be automated using functions, therefore providing a faster and more convenient technique than statistical program packages. The programming of the functions was provided by the Microsoft Excel Visual Basic for Applications service. By transforming the Royston formulas, a function was also created that gives the critical value of the test for any first-order error probability.

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

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

1. Bevezetés

Számos próba alkalmazhatóságának feltétele a matematikai statisztikában, hogy a kiértékelendő minta normális eloszlású sokaságból származzon (Obádovics, 2020), ezért több módszert fejlesztettek ki a normalitás ellenőrzésére. A Shapiro-Wilk próba ezek között a legmegbízhatóbb, mely alkalmazható kisebb elemszámú minta esetén is (Thode, 2002). A végrehajtása viszont táblázatok adatain alapul, melyek korlátozottan, csak 3 és 50 közötti elemszámra készültek (Shapiro–Wilk, 1965). Royston algoritmusával viszont akár 5000 méretű minta esetén, táblázatok adatai nélkül elvégezhetőek a számítások a normalitás ellenőrzésére (Royston, 1995).

Könnyen és egyszerűen végrehajtható normalitás vizsgálatokra a kutatásokban azért is nagy szükség van, mert sokszor nem a változó tekinthető normális eloszlásúnak, hanem a változó logaritmusa. Ekkor nevezzük a változót log-normális eloszlásúnak. Pénzügyi adatok esetén log-normális eloszlású az üzleti érték, az újrabefektetés és a befektetett tőke értéke. Statisztikai torzítások vezethetnek hibás következtetésekre, mely torzítások abból is származhatnak, ha a hipotézis kiértékelését nem előzi meg a változók normalitás vizsgálata, amikor ez a hipotézis végrehajthatóságának feltétele. Vállalatok pénzügyi elemzésekor a jövedelem

ISSN 1788-7593 (Nyomtatott) DOI: https://doi.org/10.14232/jtgf.2023.1-2.165-172
ISSN 2676-9867 (Online)

adatokat (hozam, kamat, árfolyamnyereség) szintén fontos ellenőrizni normális eloszlás szempontjából a további hipotézisvizsgálatok előtt.

A statisztikai programok jellemzően a Shapiro-Wilk próbát Royston algoritmusa alapján hajtják végre, de mivel ezek nem biztosítanak automatizálható technikát, ezért a több mintán alapuló kiértékelési sorozat egyhangú, ismétlődő tevékenységgel jár. Excel számolótáblán használható függvényekkel egyszerű és gyors kiértékelés valósítható meg több minta esetén is, s még az Excel azon tulajdonsága is segíti a kényelmes munkát, hogy adatváltozáskor aktualizálódnak az eredmények. Nagyfokú további automatizálásra nyílik mód a Visual Basic for Applications (VBA) szolgáltatással, mellyel programot készíthetünk igényeinknek megfelelően a számítások ismételt végrehajtásának segítésére (Zimmerman, 1996).

A cikk bemutatja Royston algoritmusának számításait elvégző Microsoft Excel függvények kialakítását, a függvények VBA kódját, melyekkel a Shapiro-Wilk próba kiterjesztett módon 4 és 2000 közötti elemszámú mintákon alkalmazható.

2. Anyag és módszer

2.1. A Shapiro-Wilk próba Royston szerinti kiterjesztése

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 (α) esetén, csak a leggyakoribb alkalmazott értékeknél (Shapiro–Wilk, 1965).

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

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

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*):

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

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*) (Royston 1993).

$$a_n = -2{,}706056u^5 + 4{,}434685u^4 - 2{,}07119u^3 - 0{,}147981u^2 + 0{,}221157u + c_n \tag{4}$$

$$a_{n-1} = -3{,}582633u^5 + 5{,}682633u^4 - 1{,}752461u^3 - 0{,}293762u^2 - 0{,}042981u + c_{n-1} \tag{5}$$

$$a_1 = -a_n \tag{6}$$

$$a_2 = -a_{n-1} \tag{7}$$

$$a_i = \begin{matrix} \frac{m_i}{\sqrt{f}} \\ \end{matrix} \quad \begin{matrix} i = 2, \dots, n-1 \quad i \leq 5 \\ i = 3, \dots, n-2 \quad i > 5 \end{matrix} \tag{8}$$

ahol:

$$u = \frac{1}{\sqrt{n}} \tag{9}$$

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

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

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

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

$$f = \begin{cases} \dfrac{m^2 - 2{m_n}^2}{1 - 2{a_n}^2} & ha\ n \leq 5 \\ \dfrac{m^2 - 2{m_n}^2 - 2{m_{n-1}}^2}{1 - 2{a_n}^2 - 2{a_{n-1}}^2} & ha\ n > 5 \end{cases} \tag{13}$$

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 (*14-20*) (Royston 1993).

$$p = 1 - \Phi(z) \tag{14}$$

ahol

$$z = \frac{g(W) - \mu}{\sigma} \tag{15}$$

$$g(W) = \begin{cases} -\ln(\gamma - \ln(1-W)) & 4 \le n \le 11 \\ \ln(1-W) & 12 \le n \le 2000 \end{cases} \quad (16)$$

$$\mu = \begin{cases} 0{,}544 - 0{,}39978n + 0{,}025054n^2 - 0{,}0006714n^3 & n \le 11 \\ -1{,}5861 - 0{,}31082u - 0{,}083751u^2 + 0{,}0038915u^3 & 12 \le n \end{cases} \quad (17)$$

$$\sigma = \begin{cases} \exp(1{,}3822 - 0{,}77857n + 0{,}062757n^2 - 0{,}0020322n^3) & n \le 11 \\ \exp(-0{,}4803 - 0{,}082676u + 0{,}0030302u^2) & 12 \le n \end{cases} \quad (18)$$

$$\gamma = 0{,}459n - 2{,}273 \quad (19)$$

$$u = \ln(n) \quad (20)$$

n – a minta elemszáma

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

2.2. Az Excel VBA lehetőségei

A számítások végrehajtását a Microsoft Excel program már azzal is nagymértékben megkönnyíti, hogy a számolótáblán elhelyezett formulák ismételten kiértékelődnek, amikor egy hivatkozott cella értéke megváltozik. Programokat is kialakíthatunk a Visual Basic for Applications (VBA) szolgáltatással a számítási feladatok további automatizálásához (Matteson, 1995). A VBA programozási nyelvén saját függvények készíthetők, melyek a megszokott módon használhatók. Így a függvény argumentumaként cellákra, tartományokra hivatkozhatunk, s még opcionális argumentumot is kialakíthatunk, mely alapértelmezett értéket vesz fel, amikor nem kap értéket. A program írásakor kialakíthatunk szelekciós és iterációs szerkezetet a feltételtől függő és az ismételten végrehajtandó utasítások számára. Különböző típusú adatok tárolására alkalmas változókat használhatunk, akár több dimenziós változóként is, melyek egyes értékeit indexük segítségével kezelhetjük.

3. Eredmények és értékelésük

Royston algoritmusának végrehajtásakor két fontos eredmény alapján ellenőrizhető a normalitás. Elsőként a minta adatain alapuló statisztikai függvény értékét (W) kell kiszámítanunk az (*1*)-(*13*) képletekkel, majd ezt felhasználva a (*14*)-(*20*) képletek eredményezik a szignifikanciaszint értékét (p). Így a normalitásról egy adott elsőfajú hibavalószínűség (α) mellett meghozható a döntés. A W és p érték kiszámítására VBA programmal kialakított `Sh_W` és `Sh_P` függvényt Excel számolótáblán használhatjuk.

Az eredeti Shapiro-Wilk próba végrehajtásához elegendő csak a W érték kiszámítása, melyet egy táblázatból származó kritikus értékkel (W_{cr}) összehasonlítva hozható meg a döntés a normalitásról. Viszont Royston algoritmusának képleteit át kell alakítanunk a W_{cr} értékét eredményező `Sh_Wcr` függvény VBA programjának elkészítéséhez. A (*14*)-(*16*) képletek átrendezésével adódó (*21*)-(*23*) képletek eredményezik, hogy milyen W érték lenne az a kritikus W érték (W_{cr}), mely esetén a szignifikanciaszint (p) megegyezne az elsőfajú hibavalószínűséggel (α).

$$\alpha = p = 1 - \Phi(z) \Longrightarrow z = \Phi^{-1}(1-\alpha) \quad (21)$$

$$g(W) = \mu + z\sigma \quad (22)$$

$$W_{cr} = W = \begin{cases} 1 - \exp\big(\gamma - \exp\big(-g(W)\big)\big) & 4 \le n \le 11 \\ 1 - \exp\big(g(W)\big) & 12 \le n \le 2000 \end{cases} \quad (23)$$

ahol:

n – a minta elemszáma

A fenti képletekben szereplő részeredmények (μ, σ, γ) kiszámítása a (*17*)-(*20*) képletekkel történik. A minta adatokból adódó W értéket összehasonlítva az elsőfajú hibavalószínűségből számítható W_{cr}érték alapján akkor mondhatjuk, hogy a minta normális eloszlású sokaságból származhat adott elsőfajú hibavalószínűség mellett, ha W nagyobb, mint W_{cr} $(W > W_{cr})$.

Az eddigi függvényeink alapján könnyen programozható módon elkészült egy olyan `Sh_IsNorm` függvény, melynek logikai típusú (igaz/hamis) eredménye ad választ a normalitás kérdésére.

A négy kialakított függvényben közös, hogy csak 4 és 2000 közötti elemszám esetén végzi el a számításokat, különben a hiba szöveges kijelzése történik. Így az alábbi, az `Sh_W` függvényben lévő programrészhez hasonló szerepel a többi függvény elején:

```
If n>2000 Then
  Sh_W="Túl nagy minta"
  Exit Function
ElseIf n<4 Then
  Sh_W="Túl kicsi minta"
  Exit Function
End If
```

A továbbiakban az egyes függvények kerülnek bemutatásra.

3.1. Az `Sh_W` függvény

Feladata a Shapiro-Wilk próba W értékének kiszámítása Royston algoritmusával. Az eredményt egyetlen argumentuma, a minta adatai alapján határozza meg, s elsőként a minta nagyságát (`n`) számítja ki:

```
Public Function Sh_W(sample)
  n=WorksheetFunction.Count(sample)
```

A minta adataira növekvő rendezettségben van szükség, valamint a (*11*) képletet kell alkalmazni:

```
m2=0
For i=1 To n
  x(i)=WorksheetFunction.Small(sample, i)
  m(i)=WorksheetFunction.NormSInv((i-0.375)/(n+0.25))
  m2=m2+m(i)^2
Next i
```

Az együtthatók kiszámítása a (*4*)-(*10*) képletek alapján:

```
u=1/n^(1/2)
a(n)=-2.706056*u^5+4.434685*u^4-2.07119*u^3+ _
  -0.147981*u^2+0.221157*u+m(n)/m2^(1/2)
a(n-1)=-3.582633*u^5+5.682633*u^4-1.752461*u^3+ _
  -0.293762*u^2+0.042981*u+m(n-1)/m2^(1/2)
a(1)=-a(n)
a(2)=-a(n-1)
If n > 5 Then
  f=(m2-2*m(n)^2- 2*m(n-1)^2)/(1-2*a(n)^2-2*(n-1)^2)
  k=2
Else
  f=(m2-2*m(n)^2)/(1-2*a(n)^2)
  k=1
End If
For i=k+1 To n-k
  a(i)=m(i)/f^(1/2)
Next i
```

Végül a függvény eredménye, a W kiszámítása a (*3*) képlet helyett az Excelben könnyebben megvalósítható az együtthatók vektora és a növekvő rendezettségű minta adatok vektora közötti korreláció négyzeteként:

```
w=(WorksheetFunction.Correl(x, a))^2
Sh_W = w
```

3.2. Az `Sh_P` függvény

Feladata a próba szignifikanciaszintjének (p) kiszámítása. A függvény egyetlen argumentuma a minta adatai. Első lépésként a W értéket számítja ki az `Sh_W` függvény eredményeként:

```
Public Function Sh_P(sample)
  w=Sh_W(sample)
```

A további számítások a (*14*)-(*20*) képletekkel történnek:

```
If n>11 Then
  u=WorksheetFunction.Ln(n)
  mean=0.0038915*u^3-0.083751*u^2-0.31082*u-1.5861
  lnstd=0.0030302*u^2-0.082676*u-0.4803
  std=Exp(lnstd)
  gw=WorksheetFunction.Ln(1-w)
Else
  gamma=0.459*n-2.273
  mean=-0.0006714*n^3+0.025054*n^2-0.39978*n+0.544
  lnstd=-0.0020322*n^3+0.062767*n^2-0.77857*n+1.3822
  std=Exp(lnstd)
  gw1=WorksheetFunction.Ln(1-w)
  gw=-WorksheetFunction.Ln(gamma-gw1)
End If
Z=(gw-mean)/std
```

```
p=1-WorksheetFunction.NormSDist(Z)
Sh_P=p
```

Utolsó utasításként lesz a függvény eredménye a szignifikanciaszint.

3.3. Az Sh_Wcr függvény

Feladata a döntéshez szükséges kritikus érték (W_{cr}) meghatározása. Ehhez két argumentum szükséges, az elsőfajú hibavalószínűség és a minta elemszáma:

```
Public Function Sh_Wcr(alpha, n)
```

A függvény utasításai a (*21*)-(*23*) képleteknek felelnek meg, míg a részeredmények a (*17*)-(*20*) képletek szerint számítódnak:

```
  Z=WorksheetFunction.NormSInv(1-alpha)
  If n>11 Then
    u=WorksheetFunction.Ln(n)
    mean=0.0038915*u^3-0.083751*u^2-0.31082*u-1.5861
    lnstd=0.0030302*u^2-0.082676*u-0.4803
    std=Exp(lnstd)
    gw=mean+Z*std
    wcr=1-Exp(gw)
  Else
    gamma=0.459*n-2.273
    mean=-0.0006714*n^3+0.025054*n^2-0.39978*n+0.544
    lnstd=-0.0020322*n^3+0.062767*n^2-0.77857*n+1.3822
    std=Exp(lnstd)
    gw=mean+Z*std
    wcr=1-Exp(gamma-Exp(-gw))
  End If
  Sh_Wcr=wcr
```

A függvény utolsó utasítása állítja be a függvény eredményét.

3.4. Az Sh_IsNorm függvény

A függvény feladata, hogy két argumentuma, a minta adatok és az elsőfajú hibavalószínűség alapján igaz/hamis logikai értékű eredményként adjon választ a normalitás kérdésére:

```
Public Function Sh_IsNorm(sample, Optional alpha=0.05)
  .
  .
  .
  p=Sh_P(sample)
  Sh_IsNorm=(p>alpha)
End Function
```

Ahogy a többi függvényünk, először a korábban bemutatott módon ez is ellenőrzi, hogy csak 4 és 2000 közötti elemszám mellett számoljon, majd a szignifikanciaszint kiszámítása után a $p > \alpha$ reláció logikai értéke lesz a függvény eredménye. Külön érdekessége a függvénynek, hogy a második argumentum

megadása opcionális, s ha nem adjuk meg, akkor a függvény 0,05 értékűnek tekinti az elsőfajú hibavalószínűséget, ezzel is kényelmesebbé téve a függvény használatát.

4. Következtetések

A kialakított függvények egyszerűen használható lehetőséget nyújtanak a Shapiro-Wilk próba kiértékelésére Royston algoritmusát használva a Microsoft Excel táblázatkezelő programban. Royston kiterjesztésével a próba táblázatokból származó adatok (együtthatók, kritikus határ) nélkül, 50 elemszám feletti mintán is elvégezhető. A normalitás ellenőrzésére a próba kiértékelhető szignifikanciaszint alapján is, ahogy azt az eredeti képletek tartalmazzák, de olyan képletek is kialakításra kerültek, melyekkel a próba kritikus értéke határozható meg Excel függvénnyel. Így a próba végrehajtható az eredeti Shapiro-Wilk próbánál alkalmazott, kritikus értéken alapuló döntéssel a normalitás mellett vagy ellen.

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, 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. (1993). A Toolkit for Testing for Non-Normality in Complete and Censored Samples. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 42 (1): 37–43. https://doi.org/10.2307/2348109

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.