EXCEL FÜGGVÉNYEK FEJLESZTÉSE ARÁNYRA VONATKOZÓ STATISZTIKAI PRÓBÁHOZ

Fabulya Zoltán – Hampel György – Kiss Anita

Absztrakt: Kutatásunk céljaként tűztük ki, hogy egyszerűen, széles körben alkalmazható eszközt alakítsunk ki, mellyel egy statisztikai sokaságban tesztelhetjük egy tulajdonságnak megfelelő egyedek arányára vonatkozó hipotézist nagy minták esetén. Több szempont szólt az Excel táblázatkezelő program használata mellett, melyben a Microsoft Excel Visual Basic for Applications szolgáltatással fejlesztett függvényekkel végezhető el a kiértékelés. A statisztikai programcsomagokkal szemben, az Excel számolótáblákon automatizálható a kiértékelés, így gyorsabb és kényelmesebb technikát biztosít. A függvényeink különböző adattípusok mellett képesek akár relációkkal megadható feltételeknek megfelelő egyedek arányának tesztelésére, eredményként a próba szignifikancia értékét képezve. Hibaüzenetet kapunk, ha a nagy mintára vonatkozó feltétel nem teljesül.

Abstract: Our research objective was to develop a simple and widely applicable tool which is able to test the hypothesis regarding the proportion of individuals corresponding to a characteristic in a statistical population in the case of large samples. There are several benefits of using the Excel spreadsheet program, in which the evaluation can be performed with functions developed with the Microsoft Excel Visual Basic for Applications service. Compared to statistical software packages, the test can be automated on Excel spreadsheets, thus providing a faster and more convenient technique. In addition to different data types, our functions are capable of testing the proportion of individuals that meet conditions specified by relations, forming the significance value of the test as a result. An error message is displayed if the criterion of the big sample size is not satisfied.

Kulcsszavak: próba arányra, Excel, statisztika, VBA programozás

Keywords: test for proportion, Excel, statistics, VBA programming

1. Bevezetés

A kutatások során gyakran adódó feladat, hogy egy statisztikai minta adatai alapján olyan hipotézist vizsgáljunk, mely a populációban egy adott tulajdonságú egyedek arányára vonatkozik. Jellemzően erre nagy minták esetén van szükség, amikor közelítőleg normális eloszlásúnak tekinthető a vizsgált arány. Viszont a minta elemszámán túl a hipotézisben vizsgált arányt is figyelembe kell venni a nagy minta feltételének teljesítéséhez. Míg 50%-os arány vizsgálatához 10 elemű minta elegendő, addig 1% esetén már 500 elem szükséges (Thode, 2002).

A statisztikai számítások végrehajtása, a hipotézisek kiértékelése az Excel táblázatkezelő programmal több szempontból is célszerű:

- Sokszor a kiértékelendő adatok is Excel táblázatokban szerepelnek, így nincs szükség programok közötti adatmozgatásra, az adatok konvertálására.
- A szükséges számítások könnyen, függvényekkel támogatott módon végezhetők el.
- A számítások automatizálhatók. A formulák másolatával megkaphatóak további adategységek kiértékelése. Az adatváltozások esetén aktualizálódik a számított eredmény. A Visual Basic for Applications (VBA) szolgáltatás programozási lehetőséget nyújt az ismétlődő tevékenységek újbóli

ISSN 1788-7593 (Nyomtatott) DOI: https://doi.org/10.14232/jtgf.2023.3-4.31-39
ISSN 2676-9867 (Online)

végrehajtásának algoritmusát végrehajtó programok megírására (Mansfield, 2019).
- Saját függvények készíthetők igényeinknek megfelelően (Alexander–Kusleika, 2016).

A cikkben bemutatásra kerül a sokasági arányra vonatkozó hipotézis kiértékelését nagy minták esetére kifejlesztett függvények tervezésének és elkészítésének menete a függvényekkel szemben támasztott követelmények megfogalmazásától eljutva az elkészített függvények VBA kódjáig.

2. Anyag és módszer

2.1. Egymintás z-próba arányra

Az egymintás z-próba (u-próba) minden adattípus esetén alkalmazható egy statisztikai sokaságban a vizsgált tulajdonságú egyedek arányára (p_0) vonatkozó hipotézis ellenőrzésére. A próba statisztikai függvényével kiszámított érték (z, *1-2 képlet*) csak nagy minták esetén tekinthető standard normális eloszlásúnak, azaz, ha a (*3*) és a (*4*) képletekkel adott feltételek egyszerre teljesülnek. Ezek a próba végrehajthatóságának feltételei (Obádovics, 2020).

$$z = \frac{r - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}} \quad (1)$$

$$r = \frac{m}{n} \quad (2)$$

$$n \cdot p_0 \geq 5 \quad (3)$$

$$n \cdot (1 - p_0) \geq 5 \quad (4)$$

ahol:
z – a próba statisztikai függvényének értéke
n – a minta elemszáma
m – a vizsgált tulajdonságú elemek száma, gyakorisága a mintában
r – a vizsgált tulajdonságú elemek relatív gyakorisága a mintában
p_0 – a vizsgált tulajdonságú elemek feltételezett aránya a sokaságban

A próba végrehajtásakor arról kell dönteni, hogy a nullhipotézist (*5*) fogadjuk el, vagy az ellenhipotézist. Az ellenhipotézis lehet kétszélű (*6*), egyszélű baloldali (*7*), vagy egyszélű jobboldali (*8*).

$$H_0: p = p_0 \quad (5)$$

$$H_1: p \neq p_0 \quad (6)$$

$$H_1: p < p_0 \quad (7)$$

$$H_1: p > p_0 \quad (8)$$

ahol:
p – a vizsgált tulajdonságú elemek ismeretlen aránya a sokaságban
p_0 – a vizsgált tulajdonságú elemek feltételezett aránya a sokaságban

A döntéshez a kritikus tartományt (K) kell meghatározni, melyet az elsőfajú hibavalószínűség (α) alapján kapunk a (*9*) képletet felhasználva.

$$P(z \in K | H_0) = \alpha \qquad (9)$$

ahol:
z – a statisztikai függvény értéke
K – a kritikus tartomány
H_0 – a nullhipotézis
α – az elsőfajú hibavalószínűség

A (9) képlet jelentése, hogy a nullhipotézis teljesülése esetén α valószínűséggel adódhat olyan értéke a statisztikai függvénynek, mely a kritikus tartománynak eleme. Mivel ezt a valószínűséget kicsinek választjuk, alapértelmezetten 0,05-nak, ezért csak ennyi a valószínűsége, hogy elsőfajú hibát kövessünk el a H_0 igaz volta esetén azzal, hogy hamisnak tekintjük, mert a kritikus tartományba tartozó statisztikai függvényérték adódott a mintaelemek kiértékelésével. A kritikus tartomány határát az adott valószínűséghez tartozó kvantilisként kapjuk meg a standard normális eloszlás mellett a (*10*) képletet felhasználva.

$$\Phi(z_p) = p \qquad (10)$$

ahol:
Φ – a standard normális eloszlás eloszlásfüggvénye
z_p – a p valószínűséghez tartozó kvantilis
p – egy valószínűség

Az *1. táblázatban* látható, hogy az egyes ellenhipotézisek (H_1) esetén mi a kritikus tartomány (K), és milyen feltétel ($z \in K$) teljesülésekor kell elfogadni az ellenhipotézist.

1. táblázat: **A kritikus tartomány az egyes ellenhipotéziseknél**

H_1	K	$z \in K$
$p \neq p_0$	$\left]-\infty; z_{\frac{\alpha}{2}}\right[\cup \left]z_{1-\frac{\alpha}{2}}; \infty\right[$	$\lvert z\rvert > z_{1-\frac{\alpha}{2}}$
$p < p_0$	$]-\infty; z_{\alpha}[$	$z < z_{\alpha}$
$p > p_0$	$]z_{1-\alpha}; \infty[$	$z > z_{1-\alpha}$

Forrás: Hunyadi–Vita (2008), Tóthné (2008), Domán et al. (2009), Vita (2011) alapján a szerzők szerkesztése.

2.2. Függvények kialakítása az Excel VBA környezetben

A Visual Basic for Applications (VBA) szolgáltatás programozási lehetőséget biztosít, hogy igényeinknek megfelelő programokat fejlesztve könnyítsük meg a táblázatkezelő program használatát. Így saját függvényeket alakíthatunk ki, melyek

ugyanúgy használhatók, mint a többi függvény (Walkenbach, 2015). Bár az Excel rendelkezik olyan függvényekkel is, melyek statisztikai próbák számításait végzik el, de ezek között nem szerepel az arányra vonatkozó z-próba függvénye. Saját függvény fejlesztésekor célszerű figyelembe venni a meglévő függvények tulajdonságait, hogy ezekhez hasonló függvényt fejlesztve, a felhasználó számára megszokott módon vehesse használatba. Emiatt a z-próbához fejlesztett függvény eredménye a minta alapján kiszámított szignifikancia érték (p), mely alapján az ellenhipotézis elfogadása mellett akkor döntünk, ha a szignifikancia érték kisebb, mint a választott elsőfajú hibavalószínűség ($p < \alpha$). A függvény argumentumaként kell szerepeltetni minden olyan értéket, mely szükséges az eredmény kiszámításához.

3. Eredmények és értékelésük

A függvények elkészítéséhez elsőként arra kell választ adni, hogy milyen elvárásoknak kell megfelelniük, melyek a következők:

- A függvények értéke a próba szignifikancia értéke (p-érték) legyen. Ez azért fontos, mert az Excelben elérhető statisztikai próbát végző függvények is ezt az értéket (p) eredményezik. Ennek ismeretében a döntés egyszerűen meghozható, hiszen ezt már csak az elsőfajú hibavalószínűséggel (α) kell összehasonlítani az ellenhipotézis elfogadásához ($p < \alpha$) vagy elutasításához ($p \geq \alpha$).
- A próba kiértékeléséhez szükséges mintáról elérhető legkülönbözőbb információk ismeretében legyen kiszámítható az eredmény. A (*2*) képlet kiértékeléséhez szükséges a minta elemszáma (n) és a vizsgált tulajdonságú elemek száma a mintában (m). Legegyszerűbben a két értéket a függvény argumentumaként adhatjuk meg, de arra is legyen lehetőség, hogy ezeket az értékeket a függvények határozzák meg a minta adatsorának és a vizsgált feltételnek az ismeretében. Ezért két függvény szükséges az eltérő jellegű argumentumok miatt. Az (*1*) képlethez szükséges a függvények további argumentumaként a populációban feltételezett arány (p_0). Végül a próbánál alkalmazott ellenhipotézist kell ismerni, mert eltérő képlettel határozható meg a szignifikancia érték (p) a statisztikai függvény értékéből (z) az egyes ellenhipotézisek mellett.
- A kiértékelés legyen elvégezhető numerikus és kategorikus adattípusú minta esetén is. Ennek megfelelően relációs jellel leírható ("< 30") feltételt alkalmazhassunk numerikus adattípusnál, vagy szövegként tárolt kategorikus adatok esetén szöveges feltételt ("$férfi$").

3.1. A szignifikancia érték kiszámítása

A statisztikai függvény értékéből (z) határozható meg, hogy a minta mekkora szignifikancia értéket (p) képvisel. A számítás alapja, hogy a z értékek a standard normális eloszlás megfelelő kvantilis értékei, melyek más-más valószínűségekhez tartoznak attól függően, hogy melyik ellenhipotézist alkalmazzuk a próbánál. A

kétoldali ellenhipotézis esetén a (*11-15*) levezetésből adódik a kritikus tartományba tartozó z érték feltétele alapján a próba szignifikancia értéke (p).

$$|z| > z_{1-\frac{\alpha}{2}} \tag{11}$$

$$|z| = z_{1-\frac{p}{2}} > z_{1-\frac{\alpha}{2}} \tag{12}$$

$$\Phi(|z|) = \Phi\left(z_{1-\frac{p}{2}}\right) \tag{13}$$

$$\Phi(|z|) = 1 - \frac{p}{2} \tag{14}$$

$$p = 2 \cdot \left(1 - \Phi(|z|)\right) \tag{15}$$

A *2. táblázat* tartalmazza mindhárom ellenhipotézis esetén a szignifikancia érték kiszámításának képletét.

2. táblázat: **A szignifikancia érték kiszámítása**

H_1	$z \in K$	Szignifikancia érték
kétoldali	$\|z\| > z_{1-\frac{\alpha}{2}}$	$p = 2 \cdot \left(1 - \Phi(\|z\|)\right)$
baloldali	$z < z_{\alpha}$	$p = \Phi(\|z\|)$
jobboldali	$z > z_{1-\alpha}$	$p = 1 - \Phi(\|z\|)$

Forrás: saját kutatás alapján a szerzők szerkesztése.

3.2. A függvények tervezése, fejlesztése

Két függvényre van szükség ahhoz, hogy egyrészt a minta adatsora és a vizsgált tulajdonság, másrészt a minta elemszáma és a tulajdonságnak megfelelő gyakoriság alapján is elvégezhessük a próba kiértékelését. Viszont az első esetben is képeznünk kell a második adatait részeredményként a számításhoz, ezért elegendő csak a második eset adataira támaszkodó függvénynél elvégezni az eredmény kiszámítását, és ezt felhasználni az első esetnél. Ennek megfelelően készültek el a `Z_test_prop` és a `Z_test_prop2` nevű függvények.

3.2.1. A `Z_test_prop` függvény

A függvény a minta adatsora alapján képez részeredményeket, melyekkel meghívja a másik függvényt a végeredmény, a szignifikancia érték kiszámításához. A függvény VBA kódja:

```
Public Function Z_test_prop(sample, criteria, _
                            P0, Optional hip_type = 0)
  Dim n, m As Integer
  n = WorksheetFunction.CountA(sample)
  m = WorksheetFunction.CountIf(sample, criteria)
  Z_test_prop = Z_test_prop2(n, m, P0, hip_type)
End Function
```

A függvény argumentumai:

- a minta adatsora (`sample`),
- a vizsgált tulajdonság (`criteria`)
- a feltételezett sokasági arány (`P0`),
- az ellenhipotézis típusa (`hip_type`).

Az utolsó argumentumot nem kötelező megadni. Elhagyásakor 0 lesz az értéke, melynek jelentése, hogy kétoldali ellenhipotézist alkalmazunk. Megadásakor a negatív érték baloldali, míg a pozitív érték jobboldali ellenhipotézist ír elő a számításhoz. Meghatározva részeredményként a minta elemszámát (`n`) és a vizsgált tulajdonságú elemek gyakoriságát (`m`), az eredmény a `Z_test_prop2` függvénnyel számítódik ki.

3.2.2. A `Z_test_prop2` függvény

A függvény az elemszám és a vizsgált tulajdonságú elemek gyakorisága alapján számítja ki végeredményként a próba szignifikancia értékét. Ha nem teljesül a nagy minta feltétele, akkor eredményként hibát jelez. A függvény VBA kódja:

```
Public Function Z_test_prop2(sample_size, _
        criterion_frequency, P0, Optional hip_type = 0)
  Dim n, m As Integer
  Dim r, z, p As Double
  n = sample_size
  m = criterion_frequency
  r = m / n
  If n * P0 < 5 Or n * (1 - P0) < 5 Then 'kis minta?
    Z_test_prop2 = "Too small sample!"
    Exit Function
  End If
  z = (r - P0) / (P0 * (1 - P0) / n) ^ (1 / 2)
  If hip_type < 0 Then          'baloldali
    p = WorksheetFunction.Norm_S_Dist(z, True)
  ElseIf hip_type = 0 Then      'kétoldali
    p = 2 * (1 - WorksheetFunction.Norm_S_Dist(Abs(z),_
                                          True))
  Else                          'jobboldali
    p = 1 - WorksheetFunction.Norm_S_Dist(z, True)
  End If
  Z_test_prop2 = p
End Function
```

A függvény argumentumai:

- a minta elemszáma (`sample_size`),
- a vizsgált tulajdonságú elemek gyakorisága (`criterion_frequency`)
- a feltételezett sokasági arány (`P0`),
- az ellenhipotézis típusa (`hip_type`).

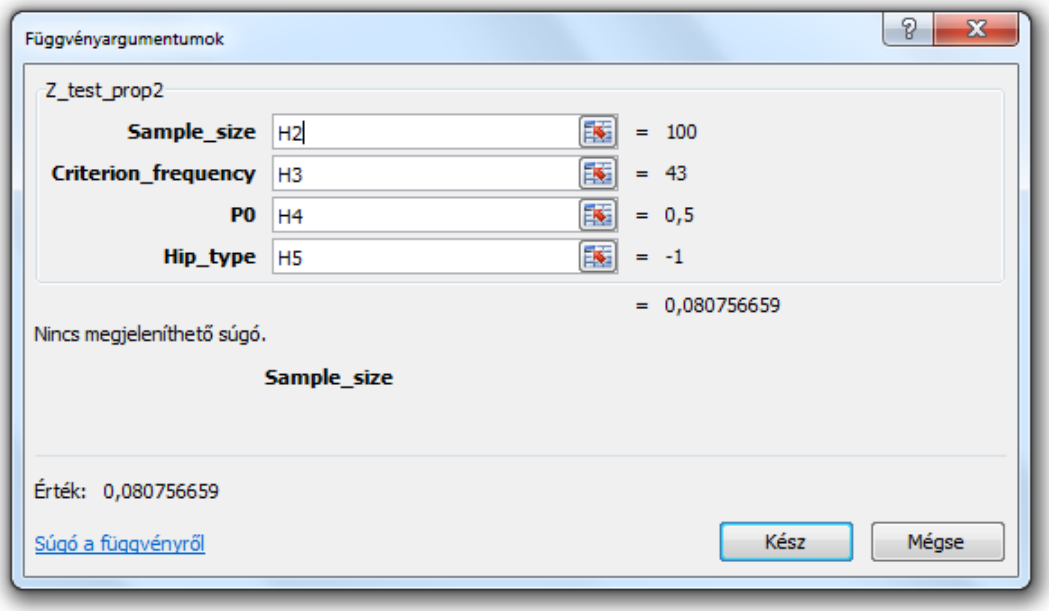
Először a nagy minta feltétele kerül ellenőrzésre. Ha ez nem teljesül, akkor hibajelzés lesz az eredmény, és befejeződik a függvény. Nagy minta esetén

kiszámítódik a statisztikai függvény értéke (`z`), majd az ellenhipotézisnek megfelelően a szignifikancia érték (`p`), mely a függvény eredménye lesz.

3.3. A függvények használata

Az Excel táblázatkezelő programban megszokott módon használhatók a kialakított függvények, így párbeszédablakkal támogatottan (*1. ábra*) is alkalmazhatók, Az argumentumaik lehetnek hivatkozott cellák, vagy begépelt adatok.

1. ábra: **A `Z_test_prop2` függvény párbeszédablaka**



Forrás: saját kutatás alapján a szerzők szerkesztése.

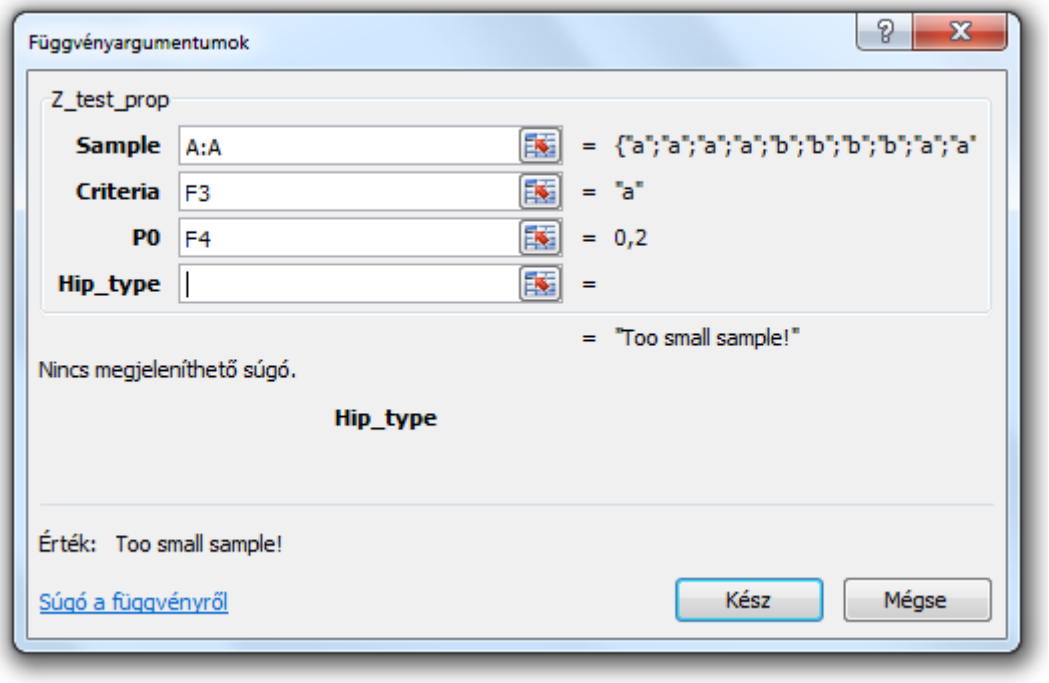
A szignifikancia érték kiszámításához szükséges programrészt tartalmazó `Z_test_prop2` függvényt az *1. ábra* szerinti argumentumokkal alkalmazva a (*16*) képletekben látható tartalmat eredményezi egy cellában, melynek értéke 0,08076.

$$= Z\_test\_prop2(H2; H3; H4; H5) \qquad (16)$$

Ahogy az *1. ábrán* látható, mind a négy argumentum hivatkozás egy-egy cellára. Az utolsó argumentum negatív értéke miatt baloldali ellenhipotézis kerül kiértékelésre. A harmadik argumentum értéke miatt ez most az, hogy a populációban az adott tulajdonságú elemek aránya kisebb, mint 0,5 ($p < 0{,}5$). Ez most elutasításra kerül 0,05 elsőfajú hibavalószínűség (α), mint szignifikanciaszint mellet, mert a függvény értéke (0,08076) nem kisebb ennél, pedig az első két argumentum szerint 100 elemű mintában 43 elem volt adott tulajdonságú. Így a mintában 0,43 volt ez az arány.

A *2. ábra* a `Z_test_prop` függvény párbeszédablakát mutatja, mely alapján a cellában a (*17*) képlet szerinti formula adódik.

2. ábra: **A `Z_test_prop` függvény párbeszédablaka**



Forrás: saját kutatás alapján a szerzők szerkesztése.

$$= Z\_test\_prop(A{:}A; F3; F4) \qquad (17)$$

A *2. ábra* mutatja, hogy az *A* oszlopban ($A{:}A$) lévő minta adatokat értékeljük ki, melyek kategorikus adattípusúak, szövegként tárolódnak „a” vagy „b” értékkel. A második argumentum alapján az „a” értékű elemek arányát vizsgáljuk, hogy ez tekinthető-e 20%-nak (harmadik argumentum értéke 0,2) vagy sem. Most azért az alapértelmezés szerinti kétoldali az ellenhipotézis, mert nem lett megadva negyedik argumentumként a hipotézis típusa. Viszont a kiértékeléshez kevésnek bizonyult a minta elemszáma, mert a „Too small sample” (Túl kicsi mintaelemszám) eredmény adódott.

Numerikus adattípus esetén relációs jellel ($"<20"$), szövegként is megfogalmazható a második argumentumban a vizsgálandó tulajdonság, ahogy ez a (*18*) képletben látható. Tehát a 20-nál kisebb értékű egyedek arányát vizsgáljuk a B oszlopban tárolt minta adatai alapján, olyan jobboldali ellenhipotézissel (negyedik argumentum pozitív), hogy az arány nagyobb 50%-nál ($p > 0{,}5$).

$$= Z\_test\_prop(B{:}B; "<20"; 0{,}5; 1) \qquad (18)$$

4. Következtetések

Az elkészült függvényekkel a megszokott módon kiértékelhető az egymintás z-próba arányra nagy minta esetén. Az adatsort és a vizsgált tulajdonságot megadva is használhatók a függvények numerikus adattípus esetén, és szövegként tárolt kategorikus adatoknál is, de a minta elemszáma és a vizsgált tulajdonságú elemek gyakorisága is használható kiinduló adatként. Mivel az eredmény a szignifikancia érték (p-érték), ezért a döntés a hipotézisről csak egy összehasonlítást igényel az elsőfajú hibavalószínűséggel, mint szignifikanciaszinttel.

Irodalomjegyzék

Alexander, M., Kusleika, R. (2016): *Excel 2016 Power Programming with VBA*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Domán Cs., Szilágyi R., Varga B. (2009): *Statisztikai elemzések alapjai II.* Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, Miskolc.

Hunyadi L., Vita L. (2008): *Statisztika II.* Aula Kiadó, Budapest.

Mansfield, R. (2019): *Mastering VBA for Microsoft Office 365, 2019 Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey.

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

Tóthné Lőkös K. (2008): *Statisztika II.* Századvég Kiadó, Budapest.

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

Vita L. (2011): A statisztikai próbák gondolatvilága. *Statisztikkai szemle*, 89 (10-11): 1130–1149. <https://www.ksh.hu/statszemle_archive/2011/2011_10-11/2011_10-11_1130.pdf> (2023.10.05.)

Walkenbach, J. (2015): *Excel 2016 Bible (1st ed.)*. John Wiley & Sons, Inc., Hoboken, New Jersey.