Mintaelemszám tervezése Likert-skálát alkalmazó lekérdezésekben

Kehl Dániel,
a Pécsi Tudományegyetem Közgazdaságtudományi Karának PhD-hallgatója
E-mail: keda05pg@ktk.pte.hu

Dr. Rappai Gábor,
a Pécsi Tudományegyetem Közgazdaságtudományi Karának egyetemi docense
E-mail: rappai@ktk.pte.hu

Jelen tanulmányukban a szerzők a mintavétel tervezésének egyik sarkalatos pontjával, a mintaelemszám meghatározásával foglalkoznak. A hagyományos – az aránybecslés standard hibájából kiinduló – metódust kiegészítve olyan módszert mutatnak be, mely a Likert-skálát tartalmazó lekérdezések esetén alkalmazható. A tanulmányban elsőként olyan tipikus eloszlásokat határoznak meg – a teljesség igénye nélkül – melyek véleményük szerint jól reprezentálják a gyakorlatban előforduló eseteket. Majd különböző hibahatárok mellett közlik az ezekhez kiszámított mintaelemszámokat, illetve a meghatározásukhoz szükséges formulákat. A szükséges mintanagyságok meghatározása után a szerzők kísérletet tesznek a kapott, és a hagyományos módszer segítségével meghatározott eredmények összehasonlítására az ún. relatív hibahatár bevezetésével.

TÁRGYSZÓ:
Mintavétel. Klasszikus módszertan.

A gyakorlati statisztikai munka egyik legfontosabb részét kétségtelenül a kutatások előkészítése, illetve ennek egyik központi eleme, a mintavétel megtervezése jelenti. A gyakorlatban dolgozók (közvélemény-kutatók, egyéb megrendelők) gyakran fordulnak az elméleti statisztikushoz azzal a nehezen (vagy egyáltalán nem) megválaszolható kérdéssel, hogy mekkora mintát kell venni ahhoz, hogy egy felmérés eredménye megbízható és pontos legyen. Nem kívánunk fejtegetésekbe bocsátkozni arról, hogy megbízhatóság és pontosság – adott mintanagyság és mintavételi mód mellett – csak egymás rovására javítható mértékek, ugyanakkor nem akarjuk kicsinyíteni sem e valóban fontos kérdést. Amikor a mintavétel tervezője és megrendelője egy gyakorlati probléma megoldása során egymással „szembekerül", gyakorlatilag ellentétes „érdekeik" vannak: a megbízhatóság és pontosság együttes növelése érdekében. A minta tervezője minél nagyobb elemszámú részsokaság kiválasztásra törekszik, a lekérdezés költségeit minimalizálni kívánó megrendelő – általában – a lehető legkisebb minta mellett érvel. A probléma megoldását az elméleti statisztikától várják (várjuk), ám e tekintetben a módszertudomány is elég kevés kézzelfogható választ kínál.

Jelen tanulmányunkban annak bemutatására törekszünk, hogy az általánosan alkalmazott (*igen-nem típusú* feleletválasztásos kérdésből kiinduló) mintanagyság-meghatározásnál – bizonyos esetekben és feltételek mellett – hatékonyabb (vagyis azonos eredményeket kisebb elemszámmal garantáló) megoldások is léteznek.

1. A mintanagyság tervezésének általános módja

A reprezentatív mintavétel alapján történő kutatások tervezésének egyik legfontosabb problémája a minta nagyságának (mintaelemszám) meghatározása. A közismert statisztikai gyakorlat a minta nagyságának meghatározása során az aránybecslés standard hibájából indul ki, ennek során ugyanis különböző – előre adott – hibahatárok esetén meghatározható a szükséges mintaelemszám. Az egyszerűség kedvéért független azonos eloszlású (FAE-) mintát feltételezve, a hibahatár[1]

$$\Delta = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}},$$

[1] A z-vel – a szokásoknak megfelelőn – a standard normális eloszlás megfelelő kvantilisét jelöljük (lásd például *Hunyadi–Vita* [2004]).

ahol a leggyakrabban alkalmazott 95,5 százalékos megbízhatósági szint $(1-\alpha)$ és a „legrosszabb eset”[2] feltételezése mellett a mintaelemszám felírható:

$$\Delta = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 2\times\sqrt{\frac{0,5(1-0,5)}{n}} = \frac{1}{\sqrt{n}} \qquad n = \frac{1}{\Delta^2}.$$

Néhány „kitüntetett” hibahatár mellett szükséges elemszámokat mutatja az 1. táblázat.

1. táblázat

Szükséges mintaelemszámok, 95,5 százalékos megbízhatósági szint és különböző hibahatárok esetén

Δ (százalékpont)	*n*
0,5	40 000
1,0	10 000
2,5	1 600
5,0	400

Az 1. táblázat értelmezése szerint, ha egy eldöntendő kérdésre adott válasz esetén az *igen* válasz aránya 100*p* százalék, és a felmérés végzője törekszik arra, hogy 95,5 százalékos megbízhatósággal azt állíthassa: az alapsokaság 100*p* ± 1 százaléka válaszolna igennel, akkor 10 000 elemű FAE-mintát kell vennie. Ha ugyanezen a megbízhatósági szinten de kisebb pontossággal kívánja állítását megfogalmazni, például a becsült érték 2,5 százalékpontos környezetében kíván maradni, akkor 1600 elemű mintára van szükség stb.[3]

Az igen-nem típusú feleletválasztós kérdések vonatkozásában a legfontosabb alapstatisztika a korábban már vizsgált arány. Ugyanakkor gyakran fordul elő, hogy egy kétkimenetelű kérdésre adható feleletet 1-gyel, illetve 2-vel jelöljük, és ezt követően nem az arányra, hanem a válaszok várható értékére vagyunk kíváncsiak. Ekkor az 1. táblázatban feltüntetett, százalékpontban felírt hibahatárok helyett használhatunk „pontértékben” mért Δ-t is, vagyis a szükséges mintaelemszám a következők szerint alakul:[4]

[2] Beláthatóan a $p(1-p)$ kifejezés maximuma, vagyis a szórás szempontjából „legrosszabb eset” p=0,5-nél áll elő, akkor a kifejezés értéke 0,25.

[3] A képet némiképpen árnyalja az egyszerű véletlen (EV-) mintavétel, illetve a maximálisnál kisebb variancia feltételezése, ám mindez a továbbiak megértését nem érinti.

[4] Vegyük észre, hogy a mintanagyságot meghatározó képlet „látszólag” azonos, ám tartalmában gyakorlatilag egészen más. A továbbiakban ezt a mintanagyságot tekintjük viszonyítási alapnak, ezért a megkülönböztető jelzés.

$$\Delta = 2\sqrt{\frac{0,5(1-1,5)^2 + 0,5(2-1,5)^2}{\tilde{n}}} = \frac{1}{\sqrt{\tilde{n}}} \qquad \tilde{n} = \frac{1}{\Delta^2}.$$

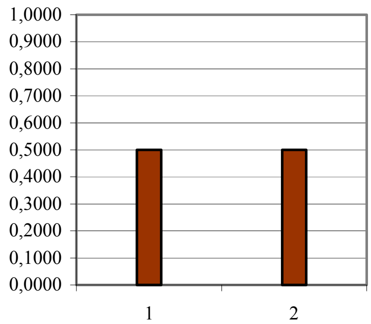
2. táblázat

Szükséges mintaelemszámok, 95,5 százalékos megbízhatósági szint és különböző hibahatárok esetén

Δ (pontérték)	*n*
0,005	40 000
0,010	10 000
0,050	400
0,100	100

A 2. táblázat eredményei – az előzők analógiájára – tehát úgy interpretálhatók, hogy ha egy konkrét válaszra adott feleletek átlaga esetében a második tizedes helyi értékben „biztos” akarok lenni, akkor 40 ezer elemű; ha csak az első tizedes „fontos” a számomra, akkor 400 elemű mintára van szükség. Összességében tehát kijelenthetjük, hogy az előző „hüvelykujj-szabállyal”, viszonylag kevés statisztikai előképzettséggel rendelkező felhasználó számára is egyszerűen meghatározható a szükséges mintanagyság; a problémát inkább az jelenti, hogy a felhasználó által elvárt (még értelmezhető) hibahatár általában olyan kicsi, hogy az túlságosan „drágává” teszi a közvélemény-kutatást.

1. ábra. Teljesen megosztott válaszadók, kétkimenetelű kérdés esetén



Az alternatív ismérv esetén történő mintanagyság bemutatása során ki kell térnünk arra a tényre is, hogy az általunk vizsgált „legrosszabb eset” $\left(p = (1-p) = 0,5\right)$ tulajdonképpen – a későbbi szóhasználattal élve – *szimmetrikus* megítélésű kérdés, vagyis a válaszadók fele az egyik, másik fele a másik alternatívát fogadja el. A későbbiekben alkalmazandó jelöléseket használva a válaszok empirikus eloszlását az 1. ábra szemlélteti.

Ugyanakkor szintén nem elhanyagolható probléma, hogy egy felmérés kérdéseinek jelentős része (zöme) nem eldöntendő, hanem többkimenetelű feleletválasztós (diszkrét), illetve mért adat (folytonos). Az előbbi kérdéstípus esetén a társadalomtudományokban elterjedt az ún. *Likert-skála*, amely 5-7-9 stb. fokozatú ordinális skálának felel meg. Tanulmányunk további részében a mintanagyság tervezésének kérdéseivel foglalkozunk Likert-skálán mért válaszokat tartalmazó kérdőívek esetén.

2. A szükséges mintaelemszám meghatározása Likert-skálán vizsgált kérdések esetén

A Likert-skálát első alkalmazójáról, *Rensis Likert*-ről nevezték el.[5] Létrehozásának célja adott egyén adott tevékenységekkel, illetve fogalommal kapcsolatos attitűdjének vizsgálata volt. Szerkezetét tekintve ezen attitűdskála két végpontján kijelölünk két „extrém" értéket, ezek testesítik meg a kérdőíven megfogalmazott állítással kapcsolatos totális ellenkezést (minimum érték), illetve teljes azonosulást (maximum érték); a skálát úgy kalibrálják, hogy középpontjában (a medián értéknél) az állítással kapcsolatos semleges érzület fejeződik ki. A skálát általában az 1–5, illetve 1–7 intervallumban szokás felállítani (vegyük észre, hogy a páratlan számú kimenetel választása lehetővé teszi, hogy a neutrális válasz is megfeleltethető legyen egy konkrét értéknek); bizonyos extrém esetekben használnak 9 fokozatú, illetve páros kimenetelű skálát is. Manapság a Likert-skálás megkérdezések nagy népszerűségnek örvendenek. A skála előnye, hogy elkészítése gyors és könnyű, valamint az, hogy akár telefonos, elektronikus úton is egyszerűen kitöltethető. (A skálát manapság nagyon gyakran alkalmazzák kérdéscsoportok formájában is, vagyis egy-egy vizsgálandó területre vonatkozóan nem egy, hanem több – estenként 20, sőt 100 – állítást fogalmaznak meg, és az ezen állításokra adott összegzett válaszértékkel dolgoznak tovább. Ez az eset távol esik tanulmányunk tárgyától, így ezzel a továbbiakban nem foglalkozunk.)

A Likert-skálás lekérdezések, vagyis a kettőnél több, páratlan[6] számú válaszlehetőséget tartalmazó kérdések esetén, a mintanagyság meghatározásának problémája azonos a korábban tárgyalttal: meg kívánjuk határozni a szükséges mintaelemszámot,

[5] *Rensis Likert* (1903–1981), a róla elnevezett skála első kifejtését tartalmazza *Likert* [1932].

[6] Tanulmányunkban csak a páratlan kimenetelű skálákat elemeztük alaposabban. A páros számú válaszlehetőséget tartalmazó lekérdezésekhez hasonló képletek határozhatók meg, de ez nem képzi dolgozatunk témáját.

előre adott hibahatár és rögzített megbízhatósági szint mellett. Ebben az esetben a hibahatár általános képlete a következőképpen módosul:

$$\Delta = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},$$

ahol σ a kérdésre adott válaszok elméleti (alapsokasági) szórása.[7] A későbbiek során látni fogjuk, hogy az eljárás eredményeképpen keletkező mintaelemszámok elégségesen nagyok ahhoz, hogy az átlagbecslés standard hibája esetén a normális eloszlás kielégítően alkalmazható legyen. Ebből kifejezhető a szükséges mintaelemszám (a korábban már említett, leggyakrabban alkalmazott feltevések mellett):

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{\Delta}\right)^2 = \left(\frac{2\sigma}{\Delta}\right)^2.$$

Láthatjuk, hogy a minta nagysága az előre adott feltételektől, valamint az alapsokasági varianciától függ. Ez utóbbi Likert-skála esetén nyilvánvalóan a – viszonylag kevés számú – válaszlehetőségekből tulajdonképpen könnyen kifejezhető abban az esetben, ha az alapsokasági eloszlás bizonyos feltételeknek megfelel. Tanulmányunkban éppen azzal foglalkozunk, hogy milyen típusú alapsokasági eloszlások feltételezése lehet reális, illetve melyik eloszlástípus, milyen alapsokasági varianciát eredményez, áttételesen mekkora mintaelemszámot tesz szükségessé. Gondolatmenetünk tehát a következő: különböző eloszlástípusokat definiálunk, majd ezek esetében meghatározzuk az elméleti (adott típusú eloszlást követő alapsokaság esetén az alapsokasági) szórást, majd ennek felhasználásával felírjuk a standard hibát, és ebből kiszámítjuk a szükséges mintaelemszámot.

Annak érdekében, hogy szórás nagyságát könnyebben meg tudjuk határozni, a továbbiakban kétféle alapsokasági eloszlástípust különítünk el:

1. szimmetrikus eloszlások, vagyis amikor

$$p_1 = p_k\,; p_2 = p_{k-1}\,; \ldots ; p_{\frac{k-1}{2}} = p_{\frac{k+3}{2}}\,; p_{\frac{k+1}{2}} = 1 - 2 \times \sum_{j=1}^{\frac{k-1}{2}} p_j\ ;$$

2. aszimmetrikus megítélésű kérdések.

[7] Rendkívül érdekes, ám általunk jelen tanulmányban nem tárgyalandó kérdés, hogy hány kimenetellel kell rendelkezni egy válasznak ahhoz, hogy a diszkrét kimenetelek szórásának legyen tárgyi értelme. Jelen írásban úgy gondoljuk, hogy akár egy ötfokozatú skála, vagyis 1, 2, 3, 4, 5 kimenetel esetén a szórás a szokásos módon értelmezhető.

2.1. Szimmetrikus eloszlású válaszadások

Könnyen belátható, hogy a *szimmetrikus eloszlások* esetén a kérdésekre adott válaszok átlaga megegyezik a neutrális értékkel (mediánnal), vagyis – páratlan kimenetelt feltételezve – meghatározása a következő képlettel történik:

$$\bar{x} = \frac{k+1}{2}.$$

Hasonlóan többször fogjuk használni a későbbiekben az első k szám (ahol k páratlan) átlagtól való eltérésének négyzetösszegét, ezért vezessük be a következő jelöléseket:[8]

$$SS^{(k)} = \left(1-\frac{k+1}{2}\right)^2 + \left(2-\frac{k+1}{2}\right)^2 + \ldots + \left(k-\frac{k+1}{2}\right)^2 = \frac{(k-1)k(k+1)}{12},$$

illetve az egyes válaszlehetőségekre adott válaszok relatív gyakoriságaival súlyozva, az átlagos eltérés-négyzetösszeg (vagyis a variancia):

$$MSS^{(k)} = p_1\left(1-\frac{k+1}{2}\right)^2 + p_2\left(2-\frac{k+1}{2}\right)^2 + \ldots + p_k\left(k-\frac{k+1}{2}\right)^2 = 2\times\sum_{j=1}^{\frac{k-1}{2}} p_j\left(j-\frac{k+1}{2}\right)^2.$$

Vegyük észre, hogy $MSS^{(k)}$ értéke maximális, ha

$$p_1 = p_k = \frac{1}{2} \;\; \textit{és} \;\; p_2 = p_3 = \ldots = p_{k\text{-}1} = 0,$$

vagyis az eloszlás extrém kétmóduszú. Ekkor a variancia:

$$0{,}5\times\left(1-\bar{x}\right)^2 + 0{,}5\times\left(k-\bar{x}\right)^2,$$

ami a következő szórást eredményezi:

$$s = \sqrt{0{,}5\times\left(1-\frac{k+1}{2}\right)^2 + 0{,}5\times\left(k-\frac{k+1}{2}\right)^2} = \frac{k-1}{2}.$$

[8] Bizonyítását lásd a Függelékben.

A hibahatár ezután a korábbi megkötésekkel (FAE-minta, és $1-\alpha=0,955$):

$$\Delta=\frac{k-1}{\sqrt{n}},$$

ebből

$$n^{EKM}=\frac{(k-1)^2}{\Delta^2}.$$

Vagyis képezhető a 2. táblázat „analógiája”, különböző méretű Likert-skálák esetére.[9] (Lásd a 3. táblázatot.)

3. táblázat

Szükséges mintaelemszámok extrém kétmóduszú sokaságok esetén

Δ	Válaszlehetőségek száma (*k*)			Általánosan
	5	7	9	
0,005	640 000	1 440 000	2 560 000	$n=\left(\frac{k-1}{0,005}\right)^2=40000\times(k-1)^2$
0,010	160 000	360 000	640 000	$n=\left(\frac{k-1}{0,010}\right)^2=10\,000\times(k-1)^2$
0,050	6 400	14 400	25 600	$n=\left(\frac{k-1}{0,025}\right)^2=400\times(k-1)^2$
0,100	1 600	3 600	6 400	$n=\left(\frac{k-1}{0,050}\right)^2=100\times(k-1)^2$

Megjegyzés. Itt és a következő táblázatokban 95,5 százalékos megbízhatósági szint és különböző hibahatárok mellett.

Láthatjuk, hogy a 3. táblázat alapján, Likert-skála alkalmazása során mindig lényegesen nagyobb mintára van szükségünk, mint a korábban feltételezett. Ne feledjük azonban, hogy az előző értékek extrém eloszlású válaszadást feltételeznek, vagyis vélelmezhetően túlbecsülik a szükséges mintaelemszámot.

A tanulmány további részében néhány könnyen beazonosítható empirikus eloszlás feltételezésével határozzuk meg a kívánatos mintaelemszámokat, majd megkísér-

[9] Vegyük észre, hogy a korábban tárgyalt alternatív (kétkimenetelű) ismérv a következő eset speciális esete.

lünk felírni néhány összefüggést, melyek a szükséges mintanagyságok, illetve a kérdésekre adott válaszok eloszlása között mutathatók ki. A tárgyalt empirikus eloszlások nem fedik le az összes elképzelhető megítéléstípust, ám az alapeseteket bemutatjuk.

Az előzőkben tárgyalt maximális variancia mellett, nyilvánvalóan felírható a minimális $MSS^{(k)}$ is, ami a következő esetben áll elő:

$$p_1 = p_2 = \ldots = p_{\frac{k-1}{2}} = p_{\frac{k+3}{2}} = \ldots = p_{k-1} = p_k = 0$$

$$p_{\frac{k+1}{2}} = 1 - 2 \times \sum_{j=1}^{\frac{k-1}{2}} p_j = 1,$$

ilyenkor $MSS^{(k)}$ értéke 0.

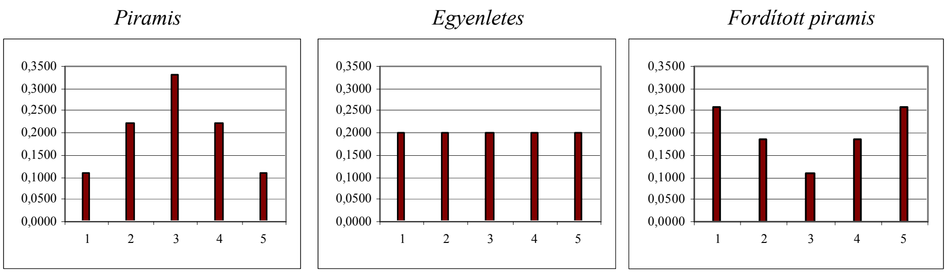
Mindez tehát azt jelenti, hogy Likert-skálás lekérdezés esetén a szükséges mintaelemszám 0 és $\left(\frac{k-1}{\Delta}\right)^2$ intervallumban mozog. Célunk, hogy ennél a tág intervallumnál szűkebb intervallumot határozzunk meg a szükséges mintanagyság tervezésénél, annál is inkább, hiszen egyik eset sem túl valószerű. Az extrém kétmóduszú esetben nehezen érthető, hogy miért van szükség 5, vagy 7 fokozatú skálára, hiszen a válaszadók csak két kimenetelt használnak; az extrém egymóduszú esetben pedig mintavételre sincs szükség, hiszen feltételeztük, hogy mindenki semlegesen viseltetik a megfogalmazott állítással szemben. Ebből következően a továbbiakban olyan válaszadási megoszlásokkal foglalkozunk, melyeknek statisztikai szempontból jó tulajdonságaik vannak, és emellett az extrém eseteknél életszerűbbek. A következőkben a szimmetrikus eloszlástípusok két csoportját mutatjuk be, az ún. *lépcsős* és a *normálison alapuló* eloszlásokat.

Lépcsős eloszlások

A lépcsős eloszlások jellemzője, hogy alapvetően a piramis típusú eloszlásra épülnek, mely úgy képződik, hogy a különböző lehetőségekre adott válaszok gyakoriságai egymás többszörösei[10] egészen a móduszig, majd ezt követően a gyakoriságok folyamatosan csökkennek. A különböző lépcsős eloszlásokat jól szemlélteti a 2. ábra.

[10] Vegyük észre, hogy relatív gyakoriságok ilyen elven képzése meglehetősen önkényes feltételezés, alkalmazását az indokolja, hogy ilyenkor viszonylag egyszerű a variancia meghatározása.

2. ábra. Lépcsős eloszlások k=5 esetén



A lépcsős eloszlásokat alapvetően az ún. piramiseloszlás (lásd a 2. ábrát) segítségével határoztuk meg, melyet a következő módon képeztünk. Legyenek a válaszadások relatív gyakoriságai rendre:

$$p;\, 2p;\, 3p;\ldots\left(\frac{k-1}{2}\right)p;\left(\frac{k+1}{2}\right)p;\left(\frac{k-1}{2}\right)p;3p;\, 2p;\, p\,.$$

Mivel a súlyok összege 1, ezért adódik:

$$p=\frac{1}{\left(\frac{k+1}{2}\right)^2}\,.$$

Ismert, hogy az átlag:

$$\bar{x}=\frac{k+1}{2},\ \text{ ekkor }\ p=\frac{1}{\bar{x}^2},\text{ és }\frac{k-1}{2}=\bar{x}-1\,.$$

Ilyen esetben a szórásnégyzet a következőképpen adódik:[11]

$$MSS^{(k)}=2\times\sum_{j=1}^{\bar{x}-1}\frac{j}{\bar{x}^2}\left(j-\bar{x}\right)^2=\frac{(k-1)(k+3)}{24}$$

$$\Delta=\sqrt{\frac{(k-1)(k+3)}{6n}}$$

$$n^{PIR}=\frac{\frac{(k-1)(k+3)}{6}}{\Delta^2}\,.$$

[11] Bizonyítását lásd a Függelékben.

4. táblázat

Szükséges mintaelemszám piramis típusú eloszlások esetén

Δ	Válaszlehetőségek száma (*k*)			Általánosan
	5	7	9	
0,005	213 333	400 000	640 000	$n=\dfrac{\frac{(k-1)(k+3)}{6}}{(0{,}005)^2}=40\,000\times\dfrac{(k-1)(k+3)}{6}$
0,010	53 333	100 000	160 000	$n=\dfrac{\frac{(k-1)(k+3)}{6}}{(0{,}01)^2}=10\,000\times\dfrac{(k-1)(k+3)}{6}$
0,050	2 133	4 000	6 400	$n=\dfrac{\frac{(k-1)(k+3)}{6}}{(0{,}05)^2}=400\times\dfrac{(k-1)(k+3)}{6}$
0,100	533	1 000	1 600	$n=\dfrac{\frac{(k-1)(k+3)}{6}}{(0{,}1)^2}=100\times\dfrac{(k-1)(k+3)}{6}$

Amennyiben az eloszlás az egyenletes eloszlás felé közelít, úgy a szórás egyre nagyobb lesz változatlan átlag mellett. Az egyenletes eloszlás esetén a variancia a következő módon határozható meg:

$$MSS^{(k)}=\frac{1}{k}\times\left(1-\frac{k+1}{2}\right)^2+\frac{1}{k}\times\left(2-\frac{k+1}{2}\right)^2+\ldots+\frac{1}{k}\times\left(k-\frac{k+1}{2}\right)^2=\frac{SS^{(k)}}{k}=\frac{(k-1)(k+1)}{12},$$

ebből felírható a hibahatár:

$$\Delta=\sqrt{\frac{k^2-1}{3n^E}},$$

majd a szükséges mintaelemszám:

$$n^E=\frac{\frac{k^2-1}{3}}{\Delta^2}.$$

Ebből következően a szükséges mintaelemszámok az 5. táblázatba rendezhetők.

5. táblázat

Szükséges mintaelemszámok egyenletes eloszlású sokaságok esetén

Δ	Válaszlehetőségek száma (k)			Általánosan
	5	7	9	
0,005	320 000	640 000	1 066 667	$n=\frac{k^2-1}{3(0{,}025)^2}=40\,000\left[\frac{k^2-1}{3}\right]$
0,010	80 000	160 000	266 667	$n=\frac{k^2-1}{3(0{,}05)^2}=10\,000\left[\frac{k^2-1}{3}\right]$
0,050	3 200	6 400	10 667	$n=\frac{k^2-1}{3(0{,}1)^2}=400\left[\frac{k^2-1}{3}\right]$
0,100	800	1 600	2 667	$n=\frac{k^2-1}{3(0{,}25)^2}=100\left[\frac{k^2-1}{3}\right]$

Amennyiben a szélsőséges válaszok felé történő átrendeződés folytatódik, egyre nagyobb lesz a szórás. A gondolatmenetünkben a következő sarkalatos eloszlás az ún. fordított piramis eloszlás. Az eloszlást a következő képlet alapján határoztuk meg:

$$p_j=\frac{1-2\times p_j^{(k)}}{k-2},$$

ahol $p_j^{(k)}$ a megfelelő tagszámú piramis típusú eloszláshoz tartozó valószínűség.

A képlet biztosítja, hogy

$$p_1>p_2>\ldots>p_{\frac{k+1}{2}}<\ldots<p_{k-1}<p_k$$

$$p_1=p_k;\quad p_2=p_{k-1};\quad \ldots$$

vagyis az eloszlás két azonos valószínűséggel előforduló, különböző nagyságú maximummal rendelkezzen, mégpedig a két szélső, extrém értéknél, valamint azt is, hogy a súlyok összege 1 legyen. (Lásd a 6. táblázatot.)

Ekkor a szórásnégyzet a következőképpen adódik:[12]

[12] Bizonyítását lásd a Függelékben.

$$MSS^{(k)} = 2 \times \sum_{j=1}^{\frac{k-1}{2}} p_j \left(j - \frac{k+1}{2} \right)^2 = 2 \times \sum_{j=1}^{\bar{x}-1} \frac{1 - \frac{2j}{\bar{x}^2}}{k-2} \left(j - \bar{x} \right)^2 = \frac{(k-1)(k^2-3)}{12(k-2)}$$

$$\Delta = \sqrt{\frac{(k-1)(k^2-3)}{3(k-2) \times n}}$$

$$n^{FPIR} = \frac{\frac{(k-1)(k^2-3)}{3(k-2)}}{\Delta^2}.$$

6. táblázat

Szükséges mintaelemszámok fordított piramis eloszlású sokaságok esetén

Δ	Válaszlehetőségek száma (k)			Általánosan
	5	7	9	
0,005	391 111	736 000	1 188 571	$n = 40\,000 \times \frac{(k-1)(k^2-3)}{3(k-2)}$
0,010	97 778	184 000	297 143	$n = 10\,000 \times \frac{(k-1)(k^2-3)}{3(k-2)}$
0,050	3 911	7 360	11 886	$n = 400 \times \frac{(k-1)(k^2-3)}{3(k-2)}$
0,100	978	1 840	2 971	$n = 100 \times \frac{(k-1)(k^2-3)}{3(k-2)}$

Vegyük észre, hogy az előző három lépcsős eloszlástípus felírható a következő módszer segítségével:

$$p_j = \frac{1 - a \times p_j^{(k)}}{k - a},$$

mely $a \to \pm\infty$ esetén a piramis-, $a = 0$ esetén az egyenletes, míg $a = 2$ esetben a fordított piramis eloszlást mutatja. Vegyük észre azt is, hogy különböző a értékek esetén eltérő lesz az eloszlások „lapultsága”. Ennek megfelelően különböző értékei segítségével is kifejezhető lenne a szórás. Mivel azonban a lekérdezés tervezésekor még nem állnak rendelkezésünkre ezen információk, az előzetes mintaelemszám-tervezés esetére megelégszünk az előzőkben részletesebben bemutatott esetek tárgyalásával. Úgy gondoljuk, hogy ezek az esetek jó támpontot nyújthatnak a mintatervezés folyamán. Ráadásul az *a* paraméter nem minden értéke esetén értelmezhető ez az eloszlás, hisz némely értékek esetén negatív relatív valószínűségeket eredményez.

Normalitáson alapuló eloszlások

A társadalmi, gazdasági élet sok jelenségét írja le közelítőleg a széles körben ismert normális eloszlás. Emiatt, valamint a némileg eltérő szórás és mintaelemszámok miatt vezetjük be a következő eloszlásokat:[13]

– fordított normális (U-alakú) (FNORM),
– „kvázi” normális (NORM),
– normális eloszláson alapuló extrém egymóduszú („nagyon csúcsos”) eloszlás (EEM).

Mivel ezen csoport összes tárgyalt alesete a „kvázi normális” eloszláson alapul, ez utóbbi eloszlástípushoz némi magyarázat tartozik. A tömegjelenségek esetén sokszor feltételezhető, és a mintavétel megrendelői körében is viszonylag széles körben ismert normális eloszlás – mint tudjuk – folytonos. Tanulmányunkban, a továbbiakban „kvázi-normálisnak” nevezzük azt a *k* darab diszkrét kimenetelhez tartozó eloszlást, amely a legjobban illeszkedik a normális eloszláshoz. Ezen empirikus eloszlás tulajdonképpen *k* darab valószínűségből álló sorozat, mely sorozat *j*-edik elemét a következő elven képezzük:

$$p_j = \varphi_j^{(k)} = \frac{\Phi\left(-z + j\frac{2z}{k}\right) - \Phi\left(-z + (j-1)\frac{2z}{k}\right)}{1 - 2\times\Phi(-z)},$$

ahol $\Phi(x)$ a standard normális eloszlás eloszlásfüggvény értéke az x helyen; és $[-z; z]$ az az intervallum, ahol a standard normális eloszlást értelmezzük.[14]

[13] Nem tárgyaljuk ismét az egyenletes eloszlást, hiszen ennek elemzése az előző alpontban megtörtént, ám – könnyen beláthatóan – az egyenletes eloszlás éppen úgy levezethető lenne a normálison alapuló eloszláscsaládból is.

[14] Természetesen a standard normális eloszlás a $(-\infty;\infty)$ intervallumon értelmezett, ám a kezelhetőség érdekében ezt az intervallumot szűkítenünk kell. Nyilvánvalóan olyan *z* értéket kell választanunk, hogy $\Phi(-z)$ minimális legyen, valamint – annak érdekében, hogy valószínűségek összege 1-et adjon – korrigálnunk kell. A továbbiakban mindvégig a [-3 ; 3] intervallummal számolunk, ekkor a nevezőben szereplő korrekciós faktor $1-2\times\Phi(-3)=0,9973$.

Látható, hogy a bevezetésben említett alternatív ismérv esetén, ez

$$p_1 = \varphi_1^{(2)} = \frac{\Phi(0) - \Phi(-3)}{1 - 2 \times \Phi(-3)} = p_2 = \varphi_2^{(2)} = \frac{\Phi(3) - \Phi(0)}{1 - 2 \times \Phi(-3)} = 0,5$$

értékeket jelenti.

A fordított normális eloszlást a lépcsős kétmóduszú eloszláshoz hasonlóan határoztuk meg. Az U-alakú eloszlások esetén az egyes kimenetelekhez tartozó valószínűségeket (vélelmezett relatív gyakoriságokat) tehát a következő képlettel határoztuk meg:

$$p_j = \frac{1 - 2\varphi_j^{(k)}}{k - 2}.$$

A lépcsős eloszlásokhoz hasonlóan az előző képlet biztosítja, hogy

$$\begin{gathered} p_1 > p_2 > \ldots > p_{\frac{k+1}{2}} < \ldots < p_{k-1} < p_k \\ p_1 = p_k; \quad p_2 = p_{k-1}; \quad \ldots \end{gathered},$$

vagyis azt, hogy az eloszlás két azonos valószínűséggel előforduló, különböző nagyságú maximummal rendelkezzen.

Az extrém egymóduszú eloszlástípus igényel némi kifejtést. Előre kívánjuk bocsátani, hogy ez a típus sem definiálható úgy, hogy csak egy eloszlás legyen hozzárendelhető; ám törekedtünk arra, hogy olyan eloszlásokat határozzunk meg, melyek a korábban ismertetettek alapján általánosíthatók. Egymóduszú eloszlásokat a következő elven képeztünk: származzanak az egyes kimenetelekhez tartozó valószínűségek a következő formulából:

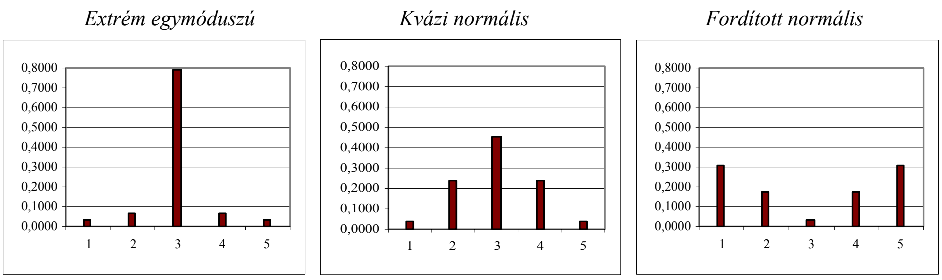
$$p_j = \begin{cases} j \times \varphi_1^{(k)}, \text{ha } j < \dfrac{k+1}{2} \\ 1 - 2\sum\limits_{i=1}^{j-1} p_j, \text{ha } j = \dfrac{k+1}{2} \\ (k + 1 - j) \times \varphi_1^{(k)}, \text{ha } j > \dfrac{k+1}{2} \end{cases}.$$

Az eljárás alapján keletkező valószínűségekre igaz, hogy:

$$\begin{gathered} p_1 < p_2 < \ldots < p_{\frac{k+1}{2}} > \ldots > p_{k-1} > p_k \\ p_1 = p_k; \quad p_2 = p_{k-1}; \quad \ldots \end{gathered},$$

és – könnyen beláthatóan – a kvázi normális eloszlásnál csúcsosabb empirikus sűrűségfüggvény keletkezik. (Természetesen érdemes megjegyeznünk, hogy az általunk előbb képzett „extrém csúcsos" eloszlás nem a maximális csúcsosságot jelenti.) Vegyük észre, hogy a lépcsős eloszlások esetén a hasonló elven képezhető eloszlás megegyezik a piramiseloszlással, így ott ezt az eloszlást nem emeltük ki külön. A normálison alapuló eloszlások sematikus képe a 3. ábrán látható.

3. ábra. Normálison alapuló eloszlások k=5 esetén



A korábban meghatározott „kvázi-normális" eloszlás esetén a mintaelemek varianciája a következő módon írható fel (kihasználva az átlagos eltérés-négyzetösszegről korábban írottakat):

$$MSS^{(k)} = 2 \times \sum_{j=1}^{\frac{k-1}{2}} p_j \left(j - \frac{k+1}{2} \right)^2 .$$

A hibahatár ebből következően:

$$\Delta = \sqrt{\frac{8 \times \sum_{j=1}^{\frac{k-1}{2}} \varphi_j^{(k)} \left(j - \frac{k+1}{2} \right)^2}{n^{NORM}}} ,$$

amiből kifejezhető a szükséges mintanagyság:

$$n^{NORM} = \frac{8 \times \sum_{j=1}^{\frac{k-1}{2}} \varphi_j^{(k)} \left(j - \frac{k+1}{2} \right)^2}{\Delta^2} .$$

Ismét képezhető ebből a 7. táblázat.

7. táblázat

Szükséges mintaelemszámok „kvázi normális” eloszlású válaszadás feltételezésével

Δ	Válaszlehetőségek száma (*k*)			Általánosan
	5	7	9	
0,005	120 852	224 636	363 049	$n = 40\,000\left(8 \times \sum_{j=1}^{\frac{k-1}{2}} \varphi_j^{(k)}\left(j - \frac{k+1}{2}\right)^2\right)$
0,010	30 213	56 159	90 762	$n = 10\,000\left(8 \times \sum_{j=1}^{\frac{k-1}{2}} \varphi_j^{(k)}\left(j - \frac{k+1}{2}\right)^2\right)$
0,050	1 209	2 246	3 630	$n = 400\left(8 \times \sum_{j=1}^{\frac{k-1}{2}} \varphi_j^{(k)}\left(j - \frac{k+1}{2}\right)^2\right)$
0,100	302	562	908	$n = 100\left(8 \times \sum_{j=1}^{\frac{k-1}{2}} \varphi_j^{(k)}\left(j - \frac{k+1}{2}\right)^2\right)$

Láthatjuk, hogy „kvázi normális” eloszlás feltételezése mellett, a mintanagyság – akárcsak korábban – függ a lehetséges kimenetelek számától, valamint a válaszlehetőségek számával párhuzamosan növekszik.

A korábban leírt fordított normális eloszlás esetén a szórás a következő képlettel határozható meg:

$$MSS^{(k)} = 2 \times \sum_{j=1}^{\frac{k-1}{2}} p_j \left(j - \frac{k+1}{2}\right)^2 = 2 \times \sum_{j=1}^{\frac{k-1}{2}} \frac{1 - 2\varphi_j^{(k)}}{k-2}\left(j - \frac{k+1}{2}\right)^2 .$$

A variancia alapján megállapítható a szükséges mintaelemszám (mivel a gondolatmenet azonos a korábbiakkal, ezért csak a mintanagyságokat közöljük).

8. táblázat

Szükséges mintaelemszámok fordított normális eloszlású sokaságok esetén

Δ	Válaszlehetőségek száma (*k*)		
	5	7	9
0,005	452 765	806 145	1 267 700
0,010	113 191	201 536	316 925
0,050	4 528	8 061	12 677
0,100	1 132	2 015	3 169

A korábban definiált extrém egymóduszú eloszlás (EEM) esetén a variancia a következő képlettel határozható meg:[15]

$$MSS^{(k)} = 2 \times \sum_{j=1}^{\frac{k-1}{2}} j \times \varphi_1^{(k)} \left(j - \frac{k+1}{2} \right)^2 = \varphi_1^{(k)} \frac{(k-1)(k+1)^2 (k+3)}{96}.$$

A szórás alapján megállapítható a szükséges mintaelemszám:

$$n^{EEM} = \frac{\varphi_1^{(k)} \frac{(k-1)(k+1)^2 (k+3)}{24}}{\Delta^2}.$$

Ebből felírható a szükséges mintanagyságok táblázata. (Lásd a 9. táblázatot.)

9. táblázat

Szükséges mintaelemszámok extrém egymóduszú sokaságok esetén

Δ	Válaszlehetőségek száma (*k*)			Általánosan
	5	7	9	
0,005	66 574	94 413	135 812	$n = 40\,000 \left(\varphi_1^{(k)} \frac{(k-1)(k+1)^2 (k+3)}{24} \right)$
0,010	16 643	23 603	33 953	$n = 10\,000 \left(\varphi_1^{(k)} \frac{(k-1)(k+1)^2 (k+3)}{24} \right)$
0,050	666	944	1 358	$n = 400 \left(\varphi_1^{(k)} \frac{(k-1)(k+1)^2 (k+3)}{24} \right)$
0,100	166	236	340	$n = 100 \left(\varphi_1^{(k)} \frac{(k-1)(k+1)^2 (k+3)}{24} \right)$

Fel kell hívnunk a figyelmet arra tényre, hogy a normálison alapuló eloszlások esetén nem tudjuk a mintanagyságot csupán a válaszlehetőségek száma, valamint a hibahatár alapján kifejezni, ezen esetekben szükséges a standard normális eloszlás bizonyos kvantiliseinek ismerete. Ezek azonban ma már könnyen meghatározhatók,

[15] Bizonyítás lásd ismét a Függelékben.

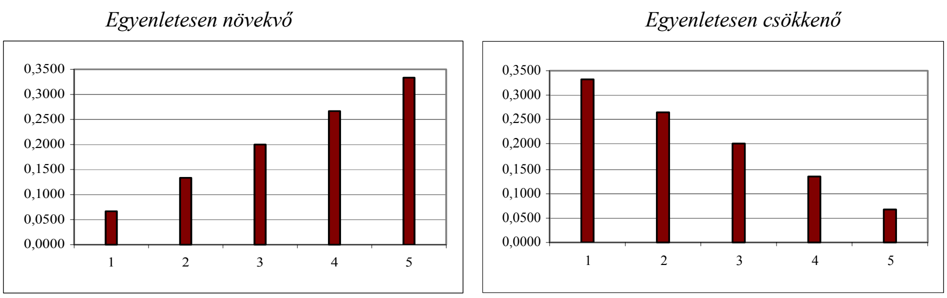
akár valamely kézikönyv táblázatainak, akár valamelyik statisztikai programcsomag használatával.

2.2 Aszimmetrikus eloszlású válaszok

Természetesen egy, a gyakorlatban végrehajtandó mintavétel esetén nem garantálható, hogy a válaszadók véleménye a semleges megfontolásra szimmetrikusan alakuljon ki. Éppen ezért célszerű megvizsgálni az aszimmetrikus vélemények esetén kialakuló eloszlások esetét is. A következőkben – a korábbinál nem kevésbé vitathatóan egyszerűsített – két esetet vizsgálunk meg:

1. az egyenletesen növekvő valószínűséggel adott válaszok esetét; valamint
2. az egyenletesen csökkenő eloszlások esetét.

4. ábra. Aszimmetrikus eloszlások



Elsőként vizsgáljuk meg azt az esetet, melyben a Likert-skála válaszlehetőségeinek előfordulási gyakorisága a teljes elutasítástól a teljes azonosulásig egyenletesen növekszik. (Az esetre a továbbiakban, mint aszimmetrikus egyenletesen növekvő eloszlásra, az AEN-kóddal hivatkozunk.) Ekkor az egyes osztályzatokra adott válaszok előfordulásának relatív gyakorisága:

$$p;\ 2p;\ \ldots;\ (k-1)p;\ kp\,.$$

Mivel

$$p+2p+\ldots+(k-1)p+kp=\frac{k(k+1)p}{2}=1\text{, ezért}$$

$$p=\frac{2}{k(k+1)}\,.$$

Az aszimmetrikus eloszlások esetén korábbi fejtegetéseinket az a tény is bonyolítja, miszerint ebben az esetben a válaszok átlagértéke nem a középső (semleges) válasz, hanem attól eltér. Egyenletesen növekvő arányban adott válaszok esetén a válaszértékek átlaga:

$$\bar{x} = 1 \times \frac{2}{k(k+1)} \times 1 + 2 \times \frac{2}{k(k+1)} \times 2 + \ldots + k \times \frac{2}{k(k+1)} \times k =$$
$$= \frac{2\sum_{j=1}^{k} j^2}{k(k+1)} = \frac{2k(k+1)(2k+1)}{6k(k+1)} = \frac{2k+1}{3}.$$

A variancia ebből a következőképen adódik:[16]

$$MSS^{(k)} = \sum_{j=1}^{k} \frac{2j}{k(k+1)} \left(j - \frac{2k+1}{3} \right)^2 = \frac{(k-1)(k+2)}{18}.$$

Amiből a szokásos módon

$$\Delta = \sqrt{\frac{4(k-1)(k+2)}{18 n^{AEN}}} \qquad n^{AEN} = \frac{\frac{2}{9}(k-1)(k+2)}{\Delta^2}.$$

A szükséges mintaelemszámokra lásd a 10. táblázatot.

10. táblázat

Szükséges mintaelemszámok egyenletesen növekvő valószínűségi válaszok esetén

Δ	Válaszlehetőségek száma (*k*)			Általánosan
	5	7	9	
0,005	248 889	480 000	782 222	$n = 40\,000\left[\frac{2}{9}(k-1)(k+2)\right]$
0,010	62 222	120 000	195 556	$n = 10\,000\left[\frac{2}{9}(k-1)(k+2)\right]$
0,050	2 489	4 800	7 822	$n = 400\left[\frac{2}{9}(k-1)(k+2)\right]$
0,100	622	1 200	1 956	$n = 100\left[\frac{2}{9}(k-1)(k+2)\right]$

[16] A variancia meghatározása a Függelékben megtalálható. Köszönetet mondunk a bizonyításért *Hunyadi Lászlónak*, aki az általunk eredetileg használt, meglehetősen nehézkes levezetés helyett az itt bemutatottat javasolta.

Az egyenletesen csökkenő arányban adott válaszok esetén sok, az előbbi (AEN-) esettel analóg megállapítást tehetünk. Az egyes válaszlehetőségek relatív gyakorisága:

$$kp;\ (k-1)p;\ \ldots;\ 2p;\ p\ ,$$

vagyis azonos számsor, csak fordított sorrendben. Ebből következően p értéke nem változik. Változik ugyan a mintaátlag:

$$\overline{x} = k \times \frac{2}{k(k+1)} \times 1 + (k-1) \times \frac{2}{k(k+1)} \times 2 + \ldots + 1 \times \frac{2}{k(k+1)} \times k =$$
$$= \frac{2}{k(k+1)} \sum_{j=1}^{k} (k+1-j)\, j = \frac{2}{k(k+1)} \left[(k+1) \sum_{j=1}^{k} j - \sum_{j=1}^{k} j^2 \right] = \frac{k+2}{3},$$

ám az átlagos eltérés-négyzetösszeg (variancia) triviálisan azonos az egyenletesen növekvő esettel. Ebből adódóan a hibahatár, illetve a szükséges mintaelemszámok megegyeznek az előbb bemutatottakkal.

3. A szükséges mintaelemszámok összehasonlítása

A folytatásban a korábban említett eseteket kíséreljük meg összevetni, ezáltal néhány gyakorlati tanácsot kívánunk adni a következtetéses statisztika alkalmazóinak. „Etalonnak” a bevett gyakorlat szerint az alternatív ismérven alapuló mintanagyság-meghatározást alkalmaztuk, ezzel vetjük össze a tanulmányban bemutatott további eloszlások feltételezésével nyert eredményeket. Ugyanakkor érdekes kérdést vet fel annak a vizsgálata, hogy milyen módon vethető össze a hagyományos, valamint a Likert-skálán mért adatok hibahatára. (A korábbiakban 8 esetet mutattunk be, ám ezek – mint bizonyítottuk – lényegileg csak 7 különböző típusnak tekinthetők, hiszen az egyenletesen növekvő, illetve csökkenő valószínűségek esete ekvivalens.)

Például az 1 százalékpontos hibahatár egészen más jelentéssel bír, amennyiben a 0–1 intervallumon belülre esik a pontbecslés, és akkor, ha az 1–k intervallumba. Emiatt a hibahatár megfelelő transzformációjára van szükség ahhoz, hogy a két érték összehasonlítható legyen. Amennyiben például a skála teljes terjedelmének 1 százaléka a „megcélzott” hibahatár, akkor – különböző fokszámú Likert-skálák esetén – felírhatjuk Δ általunk elvárt értékét:

$$\Delta_{0,01}^{(k)} = 0{,}01(k-1).$$

A tanulmány elején említett, bináris változó (k=2) esetén a terjedelem 1 százaléka értelemszerűen 0,01, így könnyen felírható az összefüggés, mellyel az eredeti, illetve a transzformált hibahatárok megfeleltethetők egymásnak:

$$\Delta^{(k)} = \frac{\Delta\left(k-1\right)}{100},$$

ahol Δ a kétkimenetelű kérdés esetén elvárt hibahatár százalékpontban kifejezett értéke. A transzformációval előállított relativizált hibahatár természetesen függ k értékétől. A 11. táblázatban a különböző Likert-skálák esetén alkalmazandó hibahatárok és az eredeti hibahatárok szerepelnek:

11. táblázat

Hibahatárok összehasonítása különböző terjedelmű Likert-skálák esetén

Δ (százalékpont)	$\Delta^{(5)}$	$\Delta^{(7)}$	$\Delta^{(9)}$
0,5	0,02	0,03	0,04
1,0	0,04	0,06	0,08
2,5	0,1	0,15	0,2
5,0	0,2	0,3	0,4

Az előző képlet alapján természetesen bármilyen fokszámú Likert-skálához meghatározható a relativizált hibahatár. Felvetődik a kérdés, hogy ezen új, „relatív” hibahatárok mellett milyen elemszámú minták szükségesek a különböző, feltételezett eloszlástípusok esetén?

A továbbiakban az ötfokozatú Likert-skálára kiszámított értékeket mutatjuk be; a nagyobb terjedelmű skálákhoz tartozó eredmények könnyedén számíthatók, és a bemutatotthoz hasonlóan értelmezhetők.

12. táblázat

Szükséges mintaelemszámok ötfokozatú Likert-skála, relatív hibahatár és különféle eloszlástípusok esetén, $1-\alpha=0{,}955$

$\Delta^{(5)}$	Extrém kétmóduszú	Fordított normális	Fordított piramis	Egyenletes	Piramis	Kvázi normális	Extrém egymóduszú
0,02	40 000	28 298	24 444	20 000	13 333	7 553	4 161
0,04	10 000	7 074	6 111	5 000	3 333	1 888	1 040
0,1	1 600	1 132	978	800	533	302	166
0,2	400	283	244	200	133	76	42

A 12. táblázat alapján jól látható, hogy az extrém kétmóduszú esetben a szükséges mintaelemszámok megegyeznek a bináris esetben számítottakkal (lásd az 1. táblázatot). Mindez természetesen nem lep meg minket, hisz a relatív hibahatár mellett teljesen mindegy, hogy a válaszadók csupán a 0, 1 lehetőségek valamelyikét választják (hisz csak ez megengedett), vagy az 1-es és 5-ös válaszlehetőségeket. A táblázat alapján jól láthatók a szükséges mintaelemszámokban mutatkozó különbségek is. A két extrém eloszlás esetén a különbség közel tízszeres a szükséges mintaelemszám tekintetében, de normális eloszlást feltételezve is a szükséges elemszám csak mintegy ötöde a maximálisnak. Dolgozatunk legfontosabb eredményeit tulajdonképpen a 12. táblázat értelmezésével nyerhetjük. Láthatjuk, hogy abban az esetben, ha valamilyen szakértői információ, vagy előzetes eredmény alapján jól meg tudjuk határozni az adott kérdésre adott válaszok eloszlásának típusát, akkor egy előre adott hibahatár, előre adott megbízhatósági szinten történő elérése az általánosan alkalmazottnál jóval kisebb mintával is megvalósítható.

Az aszimmetrikus megítélésű kérdések vizsgálata meglehetősen bonyolult, hisz míg a szimmetrikus megítélésű kérdések esetén – a feltételezett módusz és a szimmetria segítségével – könnyen tudtunk „tipikus" eseteket megnevezni, addig az aszimmetrikus eloszlások esetén a lehetséges változatok olyan nagy számával találkozunk, ami nem, vagy csak korlátozottan ad lehetőséget a kategorizálásra. Általánosságban azonban elmondható, hogy az esetek jelentős része a kvázi normális és az egyenletes eloszlás esetén szükséges mintaelemszámok közé esik, ami jó támpontot nyújt a továbbiakban. Az eddig is részletesebben bemutatott ötfokozatú skálán a különböző hibahatárokhoz tartozó elemszámok a 13. táblázatban szerepelnek. A 13. táblázatban feltüntettük az összehasonlíthatóság érdekében az egyenletes, valamint a kvázi normális eloszláshoz szükséges elemszámokat is. A szükséges minta nagysága jól láthatóan bármely hibahatárérték esetén az egyenletes és a kvázi normális eloszlás esetén szükséges közé esik.

13. táblázat

Az aszimmetrikus, illetve néhány szimmetrikus eloszlás esetén szükséges mintaelemszámok, $1–\alpha = 0,955$

$\Delta^{(5)}$	Egyenletes	Egyenletesen növekvő	Kvázi normális
0,02	20 000	15 556	7 553
0,04	5 000	3 889	1 888
0,1	800	622	302
0,2	200	156	76

Természetesen tisztában vagyunk azzal, hogy az egyenletesen növekvő/csökkenő eloszlások nem fedik le az aszimmetrikus eloszlások teljes körét, az azonban jól látható, hogy a szórás, és ezzel együtt a szükséges minta nagysága a kvázi normális és az egyenletes eloszlás közé esik. Az empirikus megfigyelések azt mutatják, hogy a valamilyen szempontból extrém eloszlásokon kívül valamennyi aszimmetrikus eloszlás ebbe az intervallumba esik, amely megállapítás jó támpontot nyújt a szükséges mintanagyság megtervezéséhez általános esetben.

*

A Likert-skálás felmérések egyre elterjedtebbek az attitűdvizsgálatokban, társadalomtudományi felmérések során. A minta nagyságának tervezése ezekben vizsgálatokban is kulcsfontosságú, hiszen az eredmények pontossága, illetve megbízhatósága jelentős részben ettől függ. Tanulmányunkban bemutattuk, hogy amennyiben létezik előzetes feltevésünk a vizsgált jelenség (válaszok) eloszlása tekintetében, akár tizedrészére tudjuk csökkenteni a szükséges mintanagyságot. Természetesen felmerülhet a kérdés, hogy mi értelme a mintavételnek, ha ilyen pontosan ismerjük a „leendő” válaszadók véleményét. Megítélésünk szerint a dolgozat fejtegetései egy kétfázisú mintavételi eljárást sugallnak: először egy kisebb elemszámú (például a kvázi normális eloszlás feltételezésével meghatározott) minta alapján meghatározzuk a válaszok várható eloszlását; majd ennek ismeretében – szükség szerint – kiegészítjük a korábbi mintavételt.

Természetesen a dolgozatban megfogalmazott eredmények semmiképpen sem tekinthetők a téma teljes „körüljárásának”. Az általunk felvázolt eloszlástípusok csak egy részét képezik a viszonylag jól meghatározható, könnyen számszerűsíthető varianciával kecsegtető eloszlásoknak. Kétségtelenül érdekes lenne megvizsgálni, hogy mi történne akkor, ha egy általunk vázolt eloszlás (például az egyenletes) nem teljesen pontosan, hanem csak sztochasztikus jelleggel alakul ki. Ugyanilyen érdekes lenne összefüggéseket keresni az eloszlást jelző aszimmetria együttható(k), és a variancia nagysága között. Tanulmányunkban – úgy érezzük – a mintavétel tervezése és a nemparaméteres (eloszlásmentes) statisztikai módszerek érdekes kapcsolódási pontját is bemutattuk.

Összességében úgy gondoljuk, hogy a mintavételi eljárás bonyolultabbá válását messzemenően ellensúlyozza az a költségcsökkenés, amit az elégséges méretű minta kiválasztásával elérhetünk, ezért a dolgozatban bemutatott eredmények gyakorlati alkalmazása célszerűnek látszik.

Függelék

1. Az első k szám (k páratlan) átlagtól mért „súlyozatlan” eltérés-négyzetösszege:

$$SS^{(k)}=\left(1-\frac{k+1}{2}\right)^2+\left(2-\frac{k+1}{2}\right)^2+\ldots+\left(k-\frac{k+1}{2}\right)^2=$$
$$=\left(1^2+2^2+\ldots+k^2\right)-\left(1+2+\ldots+k\right)\left(k+1\right)+k\left(\frac{k+1}{2}\right)^2=$$
$$=\frac{k\left(k+1\right)\left(2k+1\right)}{6}-\frac{k\left(k+1\right)^2}{2}+\frac{k\left(k+1\right)^2}{4}=$$
$$=\frac{\left(k-1\right)k\left(k+1\right)}{12}.$$

2. Az átlagos (relatív gyakoriságokkal súlyozott) eltérés-négyzetösszeg:

$$MSS^{(k)}=p_1\left(1-\frac{k+1}{2}\right)^2+p_2\left(2-\frac{k+1}{2}\right)^2+\ldots+p_{\frac{k+1}{2}}\left(\frac{k+1}{2}-\frac{k+1}{2}\right)^2+\ldots+p_k\left(k-\frac{k+1}{2}\right)^2,$$

mivel

$$p_1=p_k\,;p_2=p_{k-1}\,;\ldots$$
$$\left(1-\frac{k+1}{2}\right)=\left(k-\frac{k+1}{2}\right);\left(2-\frac{k+1}{2}\right)=\left(\left(k-1\right)-\frac{k+1}{2}\right);\ldots$$

és a „középső” tag 0, ezért

$$MSS^{(k)}=2\times\sum_{j=1}^{\frac{k-1}{2}}p_j\left(j-\frac{k+1}{2}\right)^2.$$

3. Variancia piramiseloszlás esetén

$$MSS=2\times\sum_{j=1}^{\bar{x}-1}\frac{j}{\bar{x}^2}\left(j-\bar{x}\right)^2=2\times\sum_{j=1}^{\bar{x}-1}\left(\frac{j^3}{\bar{x}^2}-\frac{2j^2}{\bar{x}}+j\right)=$$
$$=\frac{2}{\bar{x}^2}\sum_{j=1}^{\bar{x}-1}j^3-\frac{4}{\bar{x}}\sum_{j=1}^{\bar{x}-1}j^2+2\sum_{j=1}^{\bar{x}-1}j.$$

Ismeretes, hogy

$$\sum_{j=1}^{n} j^3 = \left(\frac{n(n+1)}{2}\right)^2,$$

ebből

$$MSS = \frac{2}{\bar{x}^2}\left(\frac{(\bar{x}-1)^2\bar{x}^2}{2^2}\right) - \frac{4}{\bar{x}}\frac{(\bar{x}-1)\bar{x}(2\bar{x}-1)}{6} + 2\frac{(\bar{x}-1)\bar{x}}{2} =$$

$$= \frac{(\bar{x}-1)^2}{2} - \frac{2(\bar{x}-1)(2\bar{x}-1)}{3} + (\bar{x}-1)\bar{x} =$$

$$= \frac{(\bar{x}-1)}{6}\left[3(\bar{x}-1) - 4(2\bar{x}-1) + 6\bar{x}\right] = \frac{(\bar{x}-1)(\bar{x}+1)}{6},$$

az átlagot visszahelyettesítve

$$MSS = \frac{(\bar{x}-1)(\bar{x}+1)}{6} = \frac{\left(\frac{k+1}{2}-1\right)\left(\frac{k+1}{2}+1\right)}{6} = \frac{(k-1)(k+3)}{24}.$$

4. Variancia fordított piramis eloszlás esetén

$$MSS^{(k)} = 2\times\sum_{j=1}^{\frac{k-1}{2}} p_j\left(j-\frac{k+1}{2}\right)^2 = 2\times\sum_{j=1}^{\bar{x}-1}\frac{1-2p^{(k)}\times j}{k-2}\left(j-\bar{x}\right)^2 =$$

$$= \frac{2}{k-2}\times\left[\sum_{j=1}^{\bar{x}-1}\left(j-\bar{x}\right)^2 - 2\times\sum_{j=1}^{\bar{x}-1}p^{(k)}\times j\times\left(j-\bar{x}\right)^2\right] =$$

$$= \frac{2}{k-2}\times\left[\frac{(\bar{x}-1)\bar{x}(2\bar{x}-1)}{6} - \frac{(k-1)(k+3)}{24}\right] = \frac{(k-1)(k^2-3)}{12(k-2)}.$$

5. Variancia extrém egymóduszú eloszlás esetén

$$MSS = 2 \times \sum_{j=1}^{\frac{k-1}{2}} j \times \varphi_1^{(k)} \left(j - \frac{k+1}{2} \right)^2 = 2\varphi_1^{(k)} \times \sum_{j=1}^{\frac{k-1}{2}} j \left(j - \frac{k+1}{2} \right)^2 =$$

$$= 2\varphi_1^{(k)} \left[\sum_{j=1}^{\frac{k-1}{2}} j^3 - (k+1) \sum_{j=1}^{\frac{k-1}{2}} j^2 + \left(\frac{k+1}{2} \right)^2 \sum_{j=1}^{\frac{k-1}{2}} j \right] =$$

$$= 2\varphi_1^{(k)} \left[\left(\frac{\left(\frac{k-1}{2}\right)\left(\frac{k+1}{2}\right)}{2} \right)^2 - (k+1) \frac{\left(\frac{k-1}{2}\right)\left(\frac{k+1}{2}\right) k}{6} + \left(\frac{k+1}{2} \right)^2 \frac{\left(\frac{k-1}{2}\right)\left(\frac{k+1}{2}\right)}{2} \right] =$$

$$= 2\varphi_1^{(k)} \left[\frac{(k-1)^2 (k+1)^2}{64} - \frac{(k-1) k (k+1)^2}{24} + \frac{(k-1)(k+1)^3}{32} \right] =$$

$$= 2\varphi_1^{(k)} \frac{(k-1)(k+1)^2}{192} \left[3(k-1) - 8k + 6(k+1) \right] = \varphi_1^{(k)} \frac{(k-1)(k+1)^2 (k+3)}{96}.$$

6. Variancia egyenletesen növekvő valószínűségek esetén

$$MSS = \sum_{j=1}^{k} \frac{2j}{k(k+1)} \left(j - \frac{2k+1}{3} \right)^2 ,$$

ismeretes, hogy

$$\sigma_x^2 = \frac{\sum x^2}{n} - \bar{x}^2 ,$$

a tanulmányban alkalmazott jelölésekkel

$$MSS = \frac{2\sum_{j=1}^{k} j^3}{k(k+1)} - \left(\frac{2k+1}{3} \right)^2 = \frac{k(k+1)}{2} - \frac{4k^2 + 4k + 1}{9} =$$

$$= \frac{k^2 + k - 2}{18} = \frac{(k-1)(k+2)}{18}.$$

Irodalom

HUNYADI L. [2001]: *A mintavétel alapjai*. Számalk Kiadó. Budapest.

HUNYADI L. – VITA L. [2004]: *Statisztika közgazdászoknak*. Központi Statisztikai Hivatal. Budapest.

KISH, L. [1989]: *Kutatások statisztikai tervezése*. Statisztikai Kiadó. Budapest.

LIKERT, R. [1932]: *A technique for the measurement of attitudes*. McGraw-Hill. New York.

PINTÉR J. – RAPPAI G. [2001]: A mintavételi tervek készítésének néhány gyakorlati megfontolása. *Marketing & Menedzsment* 35. évf. 4.sz. 4–10. old

Summary

In this paper the authors deal with one of the fundamental questions of sampling, the determination of the sampling size. Besides the conventional method (based on the standard deviation of the proportion estimation) the paper introduces an other technique, which can be used in surveys containing Likert scales. In this essay typical distributions are defined that – in the authors' opinion – give a good representation of the practical cases professionals may encounter. The requested sample sizes and the necessary formulas to determine them at various levels of errors are also introduced. After determining the requested sample sizes the paper attempts to compare their own results with the results given by the conventional method using the introduction of the so called "relative level of error".