Statisztikai Szemle 100 éves

Közzététel: 2022. május 19.

A tanulmány címe:
Valószínűségszámítás és statisztika

Szerző:
HUNYADI LÁSZLÓ
professor emeritus
E-mail: hunyadi44@gmail.com

DOI: https://doi.org/10.20311/stat2022.5.hu0514

Hunyadi László

Valószínűségszámítás és statisztika*

Probability theory and statistics

HUNYADI LÁSZLÓ
professor emeritus
E-mail: hunyadi44@gmail.com

Tudomásom szerint már jó ideje nem jelent meg a piacon egyetemi alapszintű statisztikakönyv. A Budapesti Corvinus Egyetem, valamint a Pécsi Tudományegyetem még a 2000-es évek elején adott ki tankönyveket, jegyzeteket, és a többi egyetem jórészt ezek felhasználásával, illetve ezekre alapozott segédanyagokkal oldotta meg az oktatást. Kiemelkedő újdonságnak számít *Kehl Dániel* könyve, amely tavaly év végén jelent meg Pécsett, és kifejezetten az ott folyó oktatást (alapoktatás) szolgálja.

A kiadvány tükrözi azt az oktatási felfogást, amelyet Pécsett talán a szükség (a közgazdasági karon lévő matematika tanszék elsorvadása) hozott létre, és itthon nem általános. Amerikában (és sok más helyen) viszont szokásos az, hogy a valószínűségszámítást és a statisztikát szorosan együtt oktassák. Ez a szemlélet persze az Amerikában (is) tanult szerzőtől nem idegen.

Könyvet említek (szerintem több mint egyetemi jegyzet), pedig valójában elektronikus tananyagról van szó, amely alapvetően R környezetben készült, és a bevezetőben olvasottak szerint ennek csak egyik oka a könnyebb és olcsóbb hozzáférhetőség, a másik viszont a rugalmasság: ilyen formában könnyen javítható és bővíthető a tartalma. Az online ismeretanyag 10 fejezetet tartalmaz 150 oldalon. Témái, azaz fő fejezetei:

1. A statisztika tárgya, alapismeretek
2. A sokaság leírása egy ismérv alapján
3. A sokasági eloszlás alakja
4. Bevezetés a valószínűségszámításba
5. Diszkrét valószínűségi változó

* KEHL DÁNIEL [2021]: *Valószínűségszámítás és statisztika*. Pécsi Tudományegyetem Közgazdasági Kar. Pécs.

6. Folytonos valószínűségi változó
7. Valószínűségi vektorváltozó
8. Mintavétel, mintavételi eloszlás
9. Az intervallumbecslés alapjai
10. Összetett becslések

Az első megjegyzés e tartalomra vonatkozik, arra, hogy mit tartalmaz (de főleg azt, hogy mit nem) a többé-kevésbé standardnak tekinthető korábbi magyar egyetemi oktatás anyagából. Sok minden hiányzik, hogy csak néhányat emeljek ki, nagyon hiányzik a statisztikai próbák tárgyalása, ami sokak számára (talán nem is indokolatlanul) a statisztika talán legfontosabb eleme; fejezetszinten a regressziószámítás, a teljes idősorelemzés és talán valami kevés a statisztika szervezetéről, a hivatalos, valamint a tudományos statisztika működéséről. Végül hiányzik – legalábbis egy közgazdasági karon oktatott alapstatisztikából – az indexszámítás és „környéke", kellékei (alighanem ez a témakör a pécsi egyetemen valamely gazdaságstatisztika jellegű tárgyban jelenik meg), valamint természetesen a „haladó szintű" statisztika fejezetei (többváltozós statisztika, statisztikai számítások, bayes-i statisztika, statisztikai modellezés stb.)

Nem kétlem, hogy a tematika ilyen lehatárolása tudatos, bizonyára hosszas viták eredménye, és alaposan indokolható is. Valószínűnek tartom, mindez csak első része az egyetemen oktatott és oktatandó tananyagnak, és a további anyagrészek tartalmazzák az általam hiányolt fejezeteket vagy legalább azok egy részét. Egyébként a lehatárolás kérdése nem lehet bírálat tárgya, hiszen egyrészt a valószínűségszámítás eleve sok helyet, időt, energiát emészt fel, másrészt viszont az, hogy milyen struktúrát, rendszert, szakmai felfogást képvisel egyik vagy másik műhely, aligha lehet vita tárgya. Mindenesetre talán jó lett volna e könyv egyébként meglehetősen szűkszavú bevezetőjében erről említést tenni.

Az érdemi bírálat előtt tekintsük át röviden a könyv szerkezetét, legfeljebb a lefedettségre figyelve. Az 1. fejezet az alapfogalmakat veszi sorra röviden, igen tömören, de célszerű megfogalmazásokkal. Ez a fejezet – akár a többi is – jól szerkesztett, átgondolt, lényegre törő, érthető. Jóval kevesebb időt, helyet és figyelmet szentel részletkérdéseknek, mint a korábbi hasonló célú tankönyvek, de úgy gondolom, ez a mai idők követelményeivel tökéletesen összhangban van. A statisztika *adatokkal* való „összenövéséről", a statisztikai adatok természetéről, pontosságáról, egyáltalán az adatok kicsit alaposabb vizsgálatáról, körbejárásáról azért talán lehetett volna többet írni. A *viszonyszámok* tárgyalása világos, logikus, a bíráló legfeljebb azt említheti meg, hogy kimaradt a két- vagy többdimenziós adatábrázolás, a kontingenciatáblák és az ezekkel kapcsolatos kérdések ismertetése. De ez még nem nagy probléma, hiszen, ha némiképp más formában is, más kontextusban a vektorváltozók témakörnél ez bejön. Ami viszont szerintem fontos lenne és hiányosság, hogy az *ábrák*,

valamint a statisztikai ábrázolás részletesebb bemutatására nem kerül sor. Az ábrázolás – manapság már inkább vizualizációnak (?) nevezik – egyre kiemeltebb szerepet kap a statisztikai elemzések során, és a rendelkezésre álló szoftverek is erősen támogatják ezt. A könyvben persze az új iránt elkötelezett szerző is él az adatvizualizáció lehetőségével (például a 3. fejezetben a box-plot ábra részletes bemutatásával), de – szerintem – az ábrázolás tárgyalásának ennél lényegesen nagyobb teret lehetett volna szentelni, és sok példa bemutatására adódott volna (vagy adódna) lehetőség. Már csak azért is, mert a szerzőnek – felkészültségét és egyéb munkáit ismerve – ez nem jelentett volna nehézséget.

A 2. fejezet a *sokaságok leírását* adja meg, alapos, de nem részletekbe menő módon. A többdimenziós (legalább kétdimenziós) esetek leírása itt nem szerepel, így az ismérvek közötti kapcsolat egyszerű elemzésére sincs mód. Ugyancsak nem szerepelnek itt a kvantilisek – ezek későbbi fejezetben kapnak helyet. Ez egy kicsit furcsa, de lehet, hogy így észszerűbb. Érdekes viszont a sokasági elhelyezkedés itteni tárgyalása: bevezeti egyrészt a *z*-transzformációt, másrészt az ún. minimax normalizálást, ami őszintén szólva számomra teljesen új, és nem is tudom, a későbbiekben hol, mikor és milyen körülmények között alkalmazható. Úgy vélem, a 2. fejezethez szorosan kapcsolódik a 3. fejezet. Ez az eloszlás *alakjával* foglalkozik, és itt vezeti be a szerző a korábban hiányolt terjedelemmutatókat és a kvantiliseket. A fejezet fajsúlyos része az, ahol 5 különböző típusú (formájú?) eloszlást mutat be, és ezeket elsősorban box-plot ábra segítségével elemzi. Itt jelennek meg az osztályközös gyakorisági sorok is, bár eléggé visszafogottan. (Az az eddigiek alapján nem meglepő, hogy szűkszavúan, de alaposan tárgyal fontos témákat, ám például a nem egyenlő osztályközök ennél talán kicsit több figyelmet érdemelnének.) Az eloszlás alakját jellemző mutatók közül csak a momentumokon alapulót említi, ami persze érthető, és egybevág a nemzetközi szokásokkal is. Ugyancsak ennek a felfogásnak tulajdonítható, hogy a könyv a ferdeséget (aszimmetriát) fordítva értelmezi, mint ahogy az a mi, Corvinus Egyetemen készült tananyagainkban és oktatásunkban szokásos. A ferdeségmutató negatív értéke ugyanis Kehl anyagában (és általában a nyugati szakirodalomban) jobb oldali ferdeséget jelent. A mi rendszerünkben – a Köves–Párniczky-koncepcióval összhangban – ez a bal oldali aszimmetria jele. Talán apróság, de olykor zavaró lehet.

A 4–7. fejezetek a *valószínűségszámítás* rövid bevezetőjét adják meg; mint a szerző is írja az elején, talán matematikus szemmel nézve nem teljesen pontosan, de szerintem nagyon jó, érthető, áttekinthető módon. Kétségtelen, hogy a szemlélet statisztikai orientáltságú, de éppen ezért akadálya lehet az esetleges továbblépésnek. Ez persze kérdés marad, de véleményem szerint, ami így és itt van, az nagyon jó, helyénvaló. Ez a rész is rövid, tömör, lényegre törő. Számomra újdonságot jelentett a „súlyfüggvény” kifejezés, amely a mi tananyagainkban (és általában) a *diszkrét valószínűségi változó eloszlása* kifejezést helyettesíti. Az alapfogalmak és a legfontosabb

összefüggések (halmazalgebra, teljes valószínűség, Bayes-tétel stb.) rövid, de jól összeszedett tárgyalásán túl a szerző bemutat egy sor, a statisztikában lényeges szerepet játszó eloszlást. Érdekes módon a diszkrét eloszlásokkal bőkezűbben bánik, hiszen a folytonos eloszlások közül például a későbbiekben fontos χ^2- vagy *F*-eloszlások kimaradnak. Ugyancsak meglehetősen rövid és szűkszavú a többdimenziós eloszlások tárgyalása, hisz lényegileg csak a változók közötti függetlenség, függés és *korreláció* fogalmának bevezetésére szolgál. Mindazonáltal ez a valószínűségszámítási bevezető, úgy vélem, nagyon hasznos, érthető, és nagyban hozzájárul a könyv önmagában is értékelhető jellegéhez.

A 8. fejezet és a folytatás visszatérés a statisztikához. A *mintavételi eloszlás* tárgyalása a szokásos utat követi, persze megint nagyon tömör formában. Középpontja a FAE- (független azonos eloszlású) mintavétel, és ebből vezeti le az ismert eredményeket. Szól a visszatevés nélküli mintavételről, annak következményeiről, valamint a mintanagyság kérdéseiről is. Talán ide lehetett volna néhány mondat erejéig beszúrni mindazt, ami a mai statisztikai alkalmazások egyik leglényegesebb pontja: a FAE-koncepció gyengülése, a sok, olykor ellenőrizetlen forrásból származó minta, a Big Data és igen rövid kritikája stb. Ebből a fejezetből az is látszik, hogy a korábban hiányolt eloszlások nélkül nehezen megy a következtetéses statisztika.

A 9. fejezet az intervallumbecslésé. Igaz, a szerző a pontbecsléssel és annak kritériumaival kezd, de a hangsúly kétségtelen az intervallumon van. Meglehet azért, mert ezzel akar megágyazni a hipotézisvizsgálatnak, ami sajnos nem fért bele ennek a könyvnek a kereteibe. Itt egy következetlenségre kell felhívnom a figyelmet: eddig a varianciát a valószínűségszámításban honos D^2-tel jelölte a szerző, de itt magyarázat nélkül áttér a *Var* jelölésre.

A 10. fejezet az összetett becsléseket ígéri, de gyakorlatilag csak a *különbségbecslést* tartalmazza. Persze talán ez a legfontosabb, hisz a statisztika lényege az összehasonlítás, amelyet legegyszerűbb módon különbséggel lehet megtenni. Fontos kiemelni, hogy ez alkalmat ad a szerzőnek a független és a páros mintavétel módjai és következményei összevetésére. Mégis sajnálatos, hogy például a hányadosbecslésnek már nem jutott hely, és ez nem szerencsés, mert egyrészt a hányados is fontos eleme a statisztikai összehasonlításnak, másrészt segítségével jó példákat lehet mutatni a becslések javítására, és nem utolsósorban átvezethet a regressziós becslésen keresztül a bonyolultabb statisztikai modellezés felé.

A nagy vonalakban ismertetett fejezetek után megkísérlem átfogóan értékelni a könyvet. A rövid terjedelem ellenére hatalmas anyagot ölel át. Lefedi az alapozó statisztika és a hozzá szükséges valószínűségszámítás jelentős részét. Jól szerkesztett, tömör, és amennyire meg tudom ítélni, pontos, hibáktól, következetlenségektől jórészt mentes. Ugyanakkor sajnos bizonyos eredményeket kénytelen kellő felvezetés, magyarázat, indoklás, levezetés nélkül tárgyalni. Alighanem ennek az az oka,

hogy a hallgatóság nem „vevő” a statisztika lényegi megértésére, „csak” melléktárgy marad, és a fontosabbaknak ítélt közgazdasági tárgyaknál kevesebb megbecsülést élvez. Az pedig, hogy a hallgatók egy tárgyat a szépségéért, mélységéért, beláthatatlan távlataiért és hasonló „ostobaságokért” megszeressenek, sajnos manapság illúzió. Nagyon szomorú.

Ennélfogva természetes, hogy a könyv szerzője nem ereszthetti szabadjára fantáziáját, kreativitását, hanem be kell állnia a sorba: megpróbálni a lehető legtöbbet kihozni a helyzetből. Ezért lehet az, hogy bár komolyan gondolkodik a statisztika jövőjéről, a könyvben erről meglehetősen kevés szó esik, inkább a hagyományos formákhoz és tartalomhoz ragaszkodik. Épphogy *csak említés szintjén* találhatók meg olyan fogalmak, mint a Big Data, az adatelemzés, a mesterséges intelligencia stb. A könyv szerkezetében, az érdemi tárgyalásban mindez nem jelenik meg.

Az egyik fontos pont, ahol komolyan előre lehetne (lehetett volna) lépni, az a statisztikai módszerek számítástechnikai, informatikai megtámogatása. És ez részben meg is történt azzal, hogy a könyv R környezetben, elektronikusan jött létre, így tárolható, másolható, sokszorosítható, ha kell, nyomtatható, továbbá ilyen formában könnyen javítható, szerkeszthető és fejleszthető. Sajnos ezek a kétségtelen előnyök a tartalomban nem jelennek meg. Számomra tökélesen érthetetlen, hogy a szerző, aki az R nyelv egyik legjobban felkészült hazai művelője, aki R kurzusokat tart (egyebek között Nyugat-Európában is!), a könyv példáit miért Excel-ben készítette el, és miért ez a program lett a tartalom alapvető támogatója. Igaz, nem ismerem annyira az Excel-t, hogy azokat a fejlettebb, alapokon túlmutató hivatkozásokat, eljárásokat igazán megértsem, de abban így is biztos vagyok, hogy még a kezdő szinten is, de kiváltképp azért, hogy haladó szinten a statisztika folytonossága megmaradjon, jobb lenne R-ben készült példákkal oktatni a hallgatóságot. Mivel mindkét program ingyen hozzáférhető, csak két dolgot tudok elképzelni ennek indoklására. Az egyik az, hogy az Excel-t a hallgatóság már más tárgyak keretében valamelyest megismerte, ezért erre (mármint a csomag használatának megismerésére) a statisztikában már kevesebb időt és energiát kell fordítani. A másik, ismét egy kicsit elszomorító ok lehet az, hogy az R alkalmazása nem (vagy kevésbé) menüvezérelt, szabadabb, ennél fogva nehezebben kiismerhető, ugyanakkor persze összehasonlíthatatlanul rugalmasabb, nyitottabb eszköz, és persze olyan, amivel korábban, más tárgyak kereteiben nem találkozhattak a hallgatók. Ezért ennek elsajátítása többletmunkát igényel(ne), amire aligha van idő, és ez a munka nem is mindig feltételezhető az átlagos hallgatótól. Mindez ismét rossz érzéseket indukál a szakmáját szerető oktatóban és bírálóban.

Pedig az jól látszik, hogy a szerző a példákat, azok ábráit R-rel készítette. Ezek szemléletesek, jól szerkesztettek, talán egyedül a 3.3. ábra az, amelyik érthetetlen módon szerencsétlen arányokkal készült. Nem hiszem, hogy nehéz feladat lenne kijavítani. A könyvön érezni lehet az amerikai hatást (tematikája, terjedelme, tárgya-

lásmódja stb.), ugyanakkor nem tartom szerencsésnek, hogy a példák témái sok esetben Amerikához fűződnek. Gondolok itt főként az elején bemutatott, és több elemzés illusztrációján továbbvezetett amerikai TOP100 technológiai vállalat példájára. Ugyanakkor meggondolandónak tartom azt, hogy egy alkalmas helyen a fontosabb szakkifejezések (amilyen például a sokaság, minta, becslés stb.) angol nyelvű megfelelője is szerepeljen. Akárhogy is nézzük, a statisztika nyelve mára angol lett, és gondolnunk kell arra is, hogy a hallgatók jó része előbb vagy utóbb ilyen nyelvi környezetben folytatja tanulmányait, és találkozik ezekkel a fogalmakkal. Ettől eltekintve a könyv *nyelvezete világos, szabatos, jól érthető.*

Mindent összefoglalva úgy vélem, a szerző az évek során összegyűjtött tapasztalatait nagyszerűen felhasználva készített hosszú idő után végre egy *jól szerkesztett és gondosan kivitelezett, modern*, amerikai stílusú *statisztika könyvet*, amely remélhetőleg egy sorozat bevezetője, első darabja. Mind technikailag, mind tartalmilag bővíthető, változtatható, és jó alapozásánál fogva tetszés szerint folytatható. Jó lenne, ha rövidesen mindez valósággá válna. A szerző – aki kérésre a könyv anyagát bárkinek rendelkezésre bocsátja – ehhez várja a *Statisztikai Szemle* olvasóinak megjegyzéseit, véleményét a kehld@ktk.pte.hu címen.