Előszó

Az elmúlt évtizedekben a matematika – az informatikához, fizikához, biológiához és más természettudományokhoz hasonlóan önálló szakmává vált; a matematika hagyományos – elsősorban műszaki és orvosi jellegű – alkalmazási területei egyre nagyobb mértékben egészültek ki pénzügyi és közgazdasági alkalmazásokkal, amelyekben a matematikai statisztikának kulcsszerepe van, például a biztosítótársaságok és bankok kockázatelemzésének megalapozásában.

Ez az igény, továbbá a hallgatói létszám növekedése tette szükségessé egy szakmailag átfogó, ugyanakkor a gyakorlati alkalmazásokra közvetlenül rámutató egyetemi tankönyv megírását. Két, látszólag egymásnak ellentmondó célt állítottunk magunk elé: egyrészt érthető, alkalmazásorientált betekintést szeretnénk adni a matematikai statisztika ma ismeretes legfontosabb módszereibe, másrészt szeretnénk kielégíteni azokat a matematikai precizitást igénylő olvasókat is, akik nem szívesen fogadnak el bizonyítás nélküli állításokat. A könyv mindazok számára olvasható, akik alapvető felsőfokú matematikai ismeretekkel (lineáris algebra, valós analízis, valószínűségszámítás) rendelkeznek. A tételek mondanivalóját illusztráló példák és a gyakorlatban közvetlenül hasznosítható ismeretek jól elválnak a mélyebb matematikai megfontolást igénylő bizonyításoktól, így első olvasáskor az utóbbiak elhagyhatók. Ez a tárgyalási mód lehetővé teszi, hogy könyvünket különböző matematikai előképzettségű, a felsőoktatás különböző szintjein tanuló hallgatók tankönyvként, a gyakorló statisztikusok pedig kézikönyvként használják.

Könyvünk címe is a fenti kettősséget kívánja kifejezni: a hangsúly – megfelelő elméleti megalapozottsággal – a módszereken (következtetéseken) van, de egyben rámutatunk a matematikai statisztika lényegére, amely a véletlen jelenségek természetére vonatkozó minél mélyebb következtetések levonásában rejlik. Wald Ábrahám, magyar származású matematikus találó kifejezésével élve: a matematikai statisztika feladata az, hogy megmutassa, hogyan viselkedjünk véletlen körülmények között.

A statisztikus rendelkezésére csupán kísérleti megfigyelések állnak, és ezek alapján kell – a valószínűségszámítás eszközeivel – minél többet kiderítenie a háttérben álló véletlen jelenség természetéről (például, hogy milyen eloszlásból származnak a megfigyelések, becsülni a paramétereket, adott esetben döntéseket hozni, vagy egy többdimenziós adatrendszer struktúráját feltárni). Ténylegesen ugyan konkrét mérésekkel dolgozunk, de sohasem szabad elfelejtenünk, hogy a kezünkben tartott adatsor a véletlen műve. Ezért méréseinket úgy kell kezelni, mint a véletlentől függő mennyiségeket, azaz valószínűségi változókat. Ez a körülmény gyakran paradoxonhoz vezet: például, ha a tapasztalati korrelációs együttható értékére 0,6-et kapunk, az 10 megfigyelés esetén lehet pusztán a véletlen játéka, míg 100 megfigyelés esetén valódi kapcsolat meglétére utal.

Az ismertetett anyag az első szerzőnek a Budapesti Műszaki Egyetemen, a második szerzőnek a Szegedi Tudományegyetemen évek óta tartott előadásaira épül. A szerzők nosztalgiával gondolnak azokra az időkre, amikor Rényi Alfréd, Vincze István és Prékopa András ma már beszerezhetetlen tankönyvei és jegyzetei álltak a matematikus hallgatók rendelkezésére. E klasszikus művek jelentik számukra az anyag magasfokú, egzakt és mégis érthető átadásának ideálját. Ezeket szeretnék kiegészíteni a modern eredményekkel.

A matematikai statisztika, sőt a valószínűségszámítás is, viszonylag új tudomány. Ez utóbbi születését Pascal és Fermat 1654-ben váltott leveleitől datálják, ezekben a levelekben oldották meg az első nemtriviális – szerencsejátékokkal kapcsolatos – valószínűségszámítási feladatot. Nincsenek arról feljegyzések, hogy az antik görögök akárcsak átlagot is számoltak volna. A matematikai statisztikai vizsgálódásokat a gyakorlati szükségszerűség hívta életre a XVIII. században, ugyanakkor John Arbuthnote-ot a társadalmi jelenségek iránti elvont érdeklődés késztette, hogy statisztikai próbával döntse el, vajon törvényszerű-e a nők és férfiak száma közötti eltérés. Thomas Bayest nagyhatású tételének megfogalmazásakor filozófiai kérdések motiválták. Biztosítótársaságok alakultak, az iparban minőségellenőr-

zésre, a mezőgazdaságban terméshozamok összehasonlítására volt szükség, a gyógyászatban kezelések hatásosságát kellett vizsgálni, kialakult az orvosi biometria (Galton, Watson). Laplace, Gauss és Csebisev koruk legkiemelkedőbb matematikusaiként statisztikai problémákkal is szembesültek (például Gauss a kisbolygók pályájának kiszámításakor), és azokat a feladatok által igényeltnél lényegesen általánosabban oldották meg. A közvetlen gyakorlatból származó feladatok megoldására kidolgozott alapvető módszerre példa a V. Gosset által a sörök minőségvizsgálatára bevezetett Student álnéven publikált *t*-próba. A statisztika rendkívüli társadalmi hatását bizonyítja Florence Nightingale munkássága, aki képzett matematikus volt, és többek között az ő jól rendszerezett adatai inspirálták Henri Dunant-t a Nemzetközi Vöröskereszt megalapítására. A XX. század első felében az angol iskola (K. Pearson, R. A. Fisher, J. Neyman) és az orosz iskola (Bernstein, Glivenko, Kolmogorov, Szmirnov) munkássága volt kiemelkedő, és természetesen a magyar származású Wald Ábrahámé, aki kidolgozta a játékelméletben is használatos döntésfüggvények elméletét, szekvenciális eljárását pedig lőszerek ellenőrzésére használták. Ő már az amerikai iskolához tartozott, akárcsak a szintén magyar származású Neumann János és Lukács Jenő. (További képviselők: E. L. Lehmann, H. Scheffé, C. Stein, a svéd H. Cramér, a francia L. Le Cam és D. Dugué). 1987-ig élt Kolmogorov, napjainkban is él még C. R. Rao. Számos ma élő és alkotó statisztikus nevét kellene még megemlítenünk.

Néhány szó a könyv egyes fejezeteinek tartalmáról.

Az 1. fejezetben a szükséges matematikai előismereteket foglaljuk össze. A lineáris algebra paragrafus a vektorok és mátrixok statisztikai alkalmazások szempontjából fontos tulajdonságait tárgyalja: Hilbert-térbeli vetítések, mátrixok spektrál- és szinguláris felbontásának optimumtulajdonságai és az optimum mibenlétére vonatkozó szeparációs tételek. Ezen eredmények nagy részét meglepő módon a XX. században fogalmazták meg. Feltételezzük, hogy az Olvasó ismeri a valószínűségszámítás mértékelméleti alapjait. Ezért a második paragrafusban csak a feltételes várható érték Kolmogorov-féle definíciójának és a tágabb értelemben vett feltételes eloszlás létezésének kérdéseit részletezzük. A valószínűségszámítási összefoglaló tartalmazza a statisztikában leggyakrabban előforduló – a normális eloszlásból származtatott – eloszlások (Γ, χ^2, t, F) sűrűségfüggvényeinek levezetését. Itt bizonyítjuk a mintaátlag és a tapasztalati szórásnégyzet függetlenségét, továbbá a Fisher–Cochran-tételt is.

A 2. fejezetben a statisztikai alapfogalmakat ismertetjük. Bizonyítjuk a matematikai statisztika egyik alaptételének számító Glivenko–Cantelli-tételt, betekintést adunk a rendezett minták elméletébe és a Kolmogorov–Szmirnov-tételkörbe. Az elégségesség, teljesség, exponenciális eloszláscsalád fogalmaira a következő fejezetekben támaszkodni fogunk.

A 3., becsléselméleti fejezetben elsősorban pontbecslésekkel foglalkozunk. Ezek tulajdonságainak, és a tulajdonságok alapstatisztikákra való alkalmazásainak ismertetése után bizonyítjuk a Cramér–Rao- és a Rao–Blackwell–Kolmogorov-tételt, majd példákon illusztráljuk következményeiket. A pontbecslési módszerek közül a maximum-likelihood-, momentum- és Bayes-módszereket ismertetjük. Az intervallumbecslések (konfidenciaintervallumok szerkesztése) már átvezetnek bennünket a hipotézisvizsgálat fejezethez.

A 4. fejezetben a hipotézisvizsgálati alapfogalmakat egy gyakorlati példa segítségével vezetjük be. A Neyman–Pearson-féle likelihood-hányados alapú döntési eljárást kiterjesztjük összetett hipotézisek vizsgálatára és ismertetjük a likelihood-hányados próba fogalmát. Ezáltal az Olvasót képessé tesszük paraméteres próbák konstruálására, továbbá a konkrét paraméteres próbákat is ezekbe a konstrukciókba illesztjük be (u-, t-, F- és relatív gyakoriságok összehasonlítására, valamint a korrelációs együttható szignifikanciájára vonatkozó próbák). A nemparaméteres próbáknál már nem taglaljuk – az itt sokkal nehezebben átlátható – optimalitási kritériumokat (χ^2-, Kolmogorov–Szmirnov- és rangstatisztikákon alapuló próbák). Részletesen tárgyaljuk viszont a Wald-féle szekvenciális eljárást és bizonyítjuk a Wald–Wolfowitz-tételt.

A többváltozós statisztikai módszereket elméletileg megalapozó 5. fejezet szorosan kapcsolódik az előzőekhez. A többdimenziós normális eloszlás ekvivalens definícióival és a többdimenziós centrális határ-eloszlástétellel indítjuk a fejezetet, ezek alapján bebizonyítjuk a χ^2-statisztika határeloszlására vonatkozó, az előző fejezetben már felhasznált tételt. Kitérünk a korrelációs mátrixok geometriai tulajdonságaira és a többdimenziós normális eloszlásfüggvényt közelítő algoritmusokra. A paraméterek maximum likelihood becslése átvezet bennünket a Wishart-mátrix sűrűségének meghatározásához, majd a sajátértékek együttes sűrűségének ugyanezen logika alapján történő kiszámításához. Ezután a becsléselméleti és hipotézisvizsgálati fogalmakat általánosítjuk a többdimenziós esetre. A többitől eltérően ennek a fejezetnek a végén feladatok is találhatók, ugyanis számos olyan apró állítás

van, amelyekre a további fejezetekben hivatkozunk, bizonyításuk azonban egyszerű számolási gyakorlat, amit az Olvasóra bízunk.

A 6. fejezetben vezetjük be a klasszikus többváltozós statisztikai módszereket, amelyek többdimenziós normális háttéreloszlásból kiinduló ún. lineáris módszerek (a kovarianciamátrix alacsony rangú közelítésein alapulnak). Céljuk struktúrafeltárás (főkomponensanalízis), dimenziócsökkentés (faktoranalízis), szórásfelbontás (varianciaanalízis) és predikció (regresszióanalízis). Részletesen tárgyaljuk a legkisebb négyzetek módszerét és a Gauss–Markov-tételt. A kanonikus korrelációanalízis feladatát optimumkeresésként fogalmazzuk meg, és a faktorok számára vonatkozó hipotézisvizsgálatnál kihasználjuk, hogy egy több célváltozós regresszióanalízisről van szó.

A 7. fejezet az utóbbi 50 évben kidolgozott osztályozási módszerekről szól. A korrespondancianalízist tipikusan kategórikus változók, kontingenciatáblák vizsgálatára fejlesztették ki, ezt a kanonikus korrelációanalízis diszkrét analogonjaként tárgyaljuk mátrixok szinguláris felbontásának segítségével. A diszkriminanciaanalízis tárgyalásában – a könyvben általánosan követett elvtől eltérően – a bayes-i megközelítést alkalmazzuk. A klaszteranalízis szerteágazó módszereiből inkább csak ízelítőt adunk. A dimenziócsökkentéshez használatos többdimenziós skálázás módszerének lineáris algebrai hátterét viszont részletesen ismertetjük.

A 8. fejezetben algoritmikus modellek címszó alatt azokat az újonnan bevezetett eljárásokat tárgyaljuk, amelyek nem a klasszikus modelleken alapulnak, hanem éppen ezek hiányosságait és a szokásostól eltérő (hiányos vagy cenzorált) adatstruktúra által felvetett numerikus nehézségeket szeretnék kiküszöbölni. Leo Breiman [Br] cikkében arról számol be, hogy gyakorlati statisztikai problémákkal szembesülve a klasszikus apparátus néha csődöt mondott. Foglalkozunk a Kaplan–Meier becslésekkel, a többdimenziós küszöbmodellekkel (ezen belül probit- és logitanalízissel), a hiányos adatokra kidolgozott EM-algoritmussal, továbbá az általános regressziós problémára alkalmazható ACE-algoritmussal. Az újramintavételezési eljárásokat (jackknife, bootstrap) egy-egy jellemző példán illusztráljuk. A fejezet végén röviden ismertetünk három, Magyarországon jelenleg elterjedt statisztikai programcsomagot. Az utolsó paragrafusban betekintést adunk a nagyon nagy minták kezelését lehetővé tevő, napjainkban kidolgozott, véletlenített lineáris algebrai módszerekbe.

Az irodalomjegyzék a klasszikus könyveken és a feldolgozott cikkeken kívül a történeti hűség kedvéért az egyes tételek eredményeinek eredeti köz-

léseit is tartalmazza, ahol ez fellelhető. Feltétlenül ki kell emelnünk H. Cramér [Cr], E. L. Lehmann [Le], C. R. Rao [Rao] és S. Zacks [Za] könyveit; ezekből vettük át a statisztikai törzsanyag már klasszikussá vált bizonyításait. A lineáris algebra statisztikában alkalmazott tételeinek bizonyításában Bhatia [Bh] könyvére támaszkodtunk. Az utóbbi 50 évben kidolgozott módszerek ismertetéséhez az eredeti dolgozatokon kívül az első szerző [BM1] [BM2] és [BM3] cikkeit is felhasználtuk. Kiemeljük még, hogy a Wishart sűrűségfüggvényt Olkin szellemes, 2002-ben publikált [Ol] dolgozata alapján számítjuk ki.

A jelölések a szokásosak, ha attól eltérő jelölést használunk, azt magyarázzuk. A tételeket és állításokat az egyes fejezeteken belül paragrafusonként számozzuk (a definíciókat, példákat nem). Az azonos fejezeten belüli hivatkozásnál csak a 3.2. Tétel jelölést használjuk, ha például az 5. fejezet 4. paragrafusában hivatkozunk ugyanezen fejezet 3. paragrafusának 2. tételére. Ha viszont ugyanerre a tételre egy másik fejezetben hivatkozunk, akkor az 5.3.2. Tétel jelölést használjuk. Az olvasó tájékozódását megkönnyítendő könyvünk végén közreadjuk a 4. fejezet hipotézisvizsgálataihoz szükséges táblázatokat, míg a nevezetes eloszlásokat és paramétereiket függelékben foglaljuk össze.

Végül köszönetet mondunk Arató Mátyásnak és Tusnády Gábornak, akik felkeltették bennünk a matematikai statisztika iránti vonzalmat, és akiknek 1960–70-es években tartott előadásait könyvünk részben követi. Köszönettel tartozunk továbbá kollégáinknak: Friedl Katalinnak, Major Péternek, Michaletzky Györgynek, Móri F. Tamásnak és a könyv lektorának, Székely J. Gábornak értékes megjegyzéseikért.

Köszönettel tartozunk tanítványainknak, Felszeghy Bálintnak, Rácz Balázsnak és Ráth Balázsnak, a Budapesti Műszaki Egyetem, valamint Ambrus Gergelynek, Balogh Ferencnek, Krauczi Évának, Szűcs Gábornak és Varjú Péternek, a Szegedi Tudományegyetem hallgatóinak számos hasznos észrevételükért.

Budapest – Szeged, 2005. február Bolla Marianna – Krámli András