PÉCSI TUDOMÁNYEGYETEM
KÖZGAZDASÁGTUDOMÁNYI KAR

GAZDÁLKODÁSTANI DOKTORI ISKOLA

Kehl Dániel

Mintaelemszám tervezés Likert-skálás lekérdezések esetén klasszikus és bayesi keretek között

DOKTORI ÉRTEKEZÉS

Témavezető: dr. Rappai Gábor

Pécs, 2012

Tartalomjegyzék

Táblázatok jegyzéke

Ábrák jegyzéke

Köszönetnyilvánítás

Disszertációm elkészítése során sokan sokféle módon segítették munkámat, az alapgondolat is egy közös tanszéki beszélgetésen fogalmazódott meg bennem évekkel ezelőtt. Köszönetet szeretnék mondani témavezetőmnek, Rappai Gábornak, aki több alkalommal terelt vissza a megkezdett útra, türelme, javaslatai és támogatása nélkül a dolgozat nem készülhetett volna el.

Hálás vagyok tanszéki kollégáimnak, Sipos Bélának, Herman Sándornak, Tiszberger Mónikának és a tragikusan korán elhunyt Pintér Józsefnek az alkotó légkör megteremtéséért és biztosításáért, akikhez bármikor fordulhattam kérdéseimmel, problémáimmal, valamint a Módszertani Intézet oktatóinak sokrétű segítségükért.

Köszönettel tartozom az eddigi pályafutásom során társszerzőként velem együtt dolgozó kollégáimnak, hiszem, hogy valamennyi közös munka hozzásegített a dolgozat végső formájának kialakításához akár közvetlenül, vagy közvetett módon. Szintén szeretném megköszönni a fenti és a disszertáció részét is képező cikkek lektorainak, opponenseinek az áldozatos munkát.

A dolgozat elkészítésében döntő jelentőségű volt az egy éves Amerikai Egyesült Államokban eltöltött időszak, mellyen kapcsolatban köszönetet mondok a Rosztoczy Foundationnek, amely a kiutazáshoz szükséges anyagi források nagy részét biztosította, valamint Szidarovszky Ferencnek és Vörös Józsefnek a támogatásukért.

A disszertáció végleges formájának kialakításában nagy segítségemre volt két előopponensem, Hunyadi László és Szidarovszky Ferenc professzorok, akik javaslataikkal nagyban hozzájárultak az esetleges pontatlanságok kiküszöbölésében. A dolgozatban maradó esetleges hibák természetesen kizárólag a szerzőt terhelik.

1. Bevezetés

Equation Chapter 1 Section 1A „Mintaelemszám tervezés Likert skálás lekérdezések esetén klasszikus és bayesi keretek között” című dolgozatomban a mintatervezés egyik szűk, mégis nagy jelentőséggel bíró részterületét járom körül és próbálom újszerű eredményekkel kiegészíteni.

Hagyományosan a mintaelemszám tervezés a becslés standard hibájából, pontosabban a hibahatárból indul ki. Az aránybecslés esetén a mintaelemszámra rendezett képletben a z-érték és a kívánt hibahatár mellett az ismeretlen sokasági arány szerepel a független változóként. A leggyakoribb feltevés, hogy erről nem tudunk semmit a vizsgálat előtt, így a lehető legrosszabb esettel számolunk, ami gyakran túlbecsüli a szükséges mintaelemszámot, mégis a gyakorlatban hatékony segítséget jelent ez a megközelítés. A várható érték becslése esetén a sokasági arány helyett a sokasági variancia szerepel a megfelelő képletben. Az irodalomban javaslatként korábbi kutatási eredmények alapján meghatározott variancia szerepeltetése, illetve előzetes minta alapján történő becslés merül fel általánosan.

Dolgozatomban érintem a kétkimenetelű, aránybecslésre vezető esetet is, a fő hangsúly azonban azon van, hogy a sokasági variancia, illetve a sokasági szórás bizonyos, kevés kimenetellel rendelkező esetekben (pl. Likert-skála) hogyan becsülhető hatékonyan, amennyiben a kutatás végzőjének van előzetes elképzelése az eloszlás várható alakjáról. Olyan eloszlásokat definiálok, melyek a gyakorlati esetek nagyobb részét lefedik, majd ezekre meghatározom a varianciát, illetve az ebből származó becsült szükséges mintaelemszámot. Foglalkozom azzal a kérdéssel, hogy a lehetséges összes eloszlás figyelembe vétele esetén mekkora a szükséges mintaelemszám várható értéke.

Alapvetően tehát abból indulok ki, hogy a felmérést végző bizonyos előzetes, ha úgy tetszik, a priori ismeretekkel rendelkezik a vizsgált sokaságról, ami megnyilvánulhat közvetlenül a variancia ismeretében is, azonban jellemzőbb, hogy magáról az eloszlásról vannak feltevéseink. A bayesi statisztika gondolati kerete (többek között) abban tér el a klasszikus statisztikáétól, hogy feltételezi előzetes információk létezését. A mintabeli információk és az a priori ismeretek ötvözésével előálló ún. poszterior eloszlás mindkét fajta tudást figyelembe veszi. Ennek alapján a bayesi statisztika megfelelő ke-

retet nyújt egy olyan szituáció kezeléséhez, ahol az előzetes ismeretek felhasználása kerül szóba.

A bayesi statisztika és ökonometria napjainkban reneszánszát éli a nemzetközi szakirodalomban, hazánkban azonban csupán néhány tanulmány, illetve műhely foglalkozik vele. A módszertan térnyerésének több oka van, többek között az, hogy a Bayes-tétel gyakorlatilag önmagában elegendő az elméleti keret megértéséhez. Sok esetben lehetőség van (sőt, szükséges) külső információk ellenőrzött, dokumentált módon modellbe való építésére. Szintén okként említhetjük, hogy az elemzés végeredményeként előálló poszterior eloszlás önmagában a paraméterek eloszlását írja le, így konfidencia intervallumok előállításához nincs szükségünk kiegészítő feltevések (pl. normalitás) használatára. A bayesi módszerek nehézségét az okozza, hogy a (jellemzően sokdimenziós) poszterior eloszlásban rejlő információk prezentálása, megértése nem triviális, leggyakrabban szimulációs technikákat kell segítségül hívnunk. Ezen szimulációs technikák matematikai alapjai hosszabb történetre tekintenek vissza, azonban sokáig hiányzott az eljárásokhoz szükséges számítási, számítástechnikai háttér. Amióta mindkét alapfeltétel adott, a terület fejlődése töretlen.

A bayesi megközelítésben lehetőségünk van tehát az előzetes ismeretek és a mintából származó információk egyesítésére. Éppen ez az a két módszer, amelyet a szakirodalom a várható érték becslés hibahatárából adódó mintanagyság képletben a szórás közelítésére javasol!

A dolgozatban a célom a mintaelemszám tervezésének hatékony módszerét kidolgozni. A klasszikus mellett a bayesi gondolatkört is felhasználom, valamint összevetem az általuk szolgált eredményeket. Azt várom, hogy a munkám olyan kézzel fogható eljárásokkal gazdagítja a szakirodalmat, amelyeket a gyakorlati statisztika képes felhasználni. A bayesi statisztika alapgondolatainak és fő módszereinek áttekintését jelen dolgozattól függetlenül is fontosnak tartom.

A fenti célokkal párhuzamosan megfogalmaztam azokat a hipotéziseket, melyek a kutatás fő motivációját adták. A hipotézisek szorosan kötődnek a disszertáció egyes fejezeteihez, a velük kapcsolatosan levont konklúzióimat az összefoglaló fejezetben tárgyalom.

A hipotézisek tehát a következők:

1. A Likert-skálás lekérdezések segítségével nyert változók esetén a módszerválasztás különös jelentőséggel bír.
2. Amennyiben rendelkezünk külső információval a sokasági eloszlásról, az hatékonyan alkalmazható az előzetesen kalkulált szükséges mintaelemszám csökkentésére Likert-skálás kérdések esetén is.
3. Adott kívánt hibahatár és megbízhatósági szint esetén meghatározható a szükséges mintaelemszám várható értéke.
4. Az előzetes információk és egy esetleges előzetes mintavétel adatainak összesítése a bayesi keretrendszerben probléma nélkül megoldható. A bayesi módszertannal kiszámított szükséges mintaelemszám értékek a klasszikus statisztikai módszerekkel számítható értékekkel összhangot mutatnak.
5. Az előzetes információk bizonytalansága és az előzetes minta mintavételi hibája figyelembe vehető a bayesi gondolatvilágban, ami egyértelműen előnyt jelent a klasszikus megközelítéshez képest.

Az értekezés összesen nyolc fő fejezetet tartalmaz, amelyből négy alkotja a munka anyagának döntő részét. A disszertáció több éves munka eredménye, így az egyes fejezetek a korábbi években megjelent – jellemzően egyszerzős – cikkekben már publikálásra kerültek.

A jelen, bevezető fejezet után a 2. azzal foglalkozik, hogy a címben szereplő Likert-skálás kérdések esetén milyen statisztikai mutatók alkalmazhatók (Kehl, 2011). Az évtizedeken át zajló vita még a mai napig is tart, azt azonban megállapíthatjuk, hogy a legtöbb kutató egyetért abban, hogy átlag és szórás mutatókat csak olyan esetben számítunk, amikor a változó értékek ekvidisztansok, tisztán ordinális esetben ezek a mutatók nem alkalmazhatók. Az alkalmazók többsége Likert-skála használata esetén feltételezi ennek a jellemzőnek a teljesülését.

A 3. fejezetben a mintaelemszám tervezés klasszikus statisztikai szemléletben történő vizsgálatával foglalkozom (Kehl-Rappai, 2006). A fejezet elején a Likert-skála eredetét, fő felhasználási területeit, majd a mintaelemszám tervezés általános módszerét mutatom be. Előre definiált eloszlások esetén határozok meg szórásjellemzőket, amelyek segítségével az adott eloszlású véletlen változóhoz tartozó szükséges

mintaelemszámot becsülöm. Megtörténik egyfajta érzékenységvizsgálat és a különböző feltételezések mellett nyert elemszámok összehasonlítása, valamint a minden lehetséges mintára vonatkozó várható érték meghatározása (Kehl, 2007) is.

A 4. fejezetben a bayesi gondolkodás alapjait mutatom be, valamint kitérek azokra az alapvető és összetettebb módszerekre is, melyek a poszterior eloszlás összefoglalását segítik (Kehl, 2012a, 2012b). A szakasz gyakorlatilag egy általános bevezető a bayesi statisztikába, majd az azt kiegészítő szimulációs (MC és MCMC) technikákat 1-1 rövid példával kiegészítve mutatom be. A szimulációk futtatásához az R környezetet használom, a Függelék a szükséges programsorokat tartalmazza, így az anyag a későbbiekben egy magyar nyelvű, bayesi statisztikával, szimulációval foglalkozó kurzus alapja is lehet.

Az 5. fejezet a bayesi megközelítés gyakorlati megvalósítását mutatja be a dolgozat fő problémáján keresztül két alfejezetben. Elsőként definiálom a kétváltozós esetben könnyen alkalmazható konjugált prior eloszlást, amely rugalmasan képes a rendelkezésre álló előzetes információk leírására. A poszterior pedig tartalmazza az esetleges előzetes mintavételi eredményeket is, így a két tudás kombinációjaként értelmezhető. A második – jóval rövidebb – alpont a kettőnél több kimenetellel rendelkező változók esetét mutatja be. A fejezet rövidségét az teszi lehetővé, hogy a kétváltozós eset gyakorlatilag gond nélkül általánosítható többváltozósra. A binomiális eloszlás helyét a multinomiális, a bétáét a Dirichlet veszi át, így a folyamat gyakorlatilag teljes mértékben megegyezik a kétváltozós esetben leírttal.

A 6. fejezet az értekezés eredményeinek rövid összefoglalását adja és további potenciális kutatási területeket azonosít, míg az utolsó két pont a főszövegbe terjedelmüknél fogva nem illő levezetéseket, bizonyos eloszlások jellemzőit, illetve a már említett programokat tartalmazza. A dolgozatot részletes felhasznált irodalomjegyzék zárja.

2. Skálák és statisztikák: a méréselméletről és történetéről

Equation Section (Next)Adatok, ezen belül is statisztikai alapadatok jellemzően számlálás, illetve mérés útján keletkeznek. A számlálás útján előállított adatok esetén is találkozhatunk gyakorlati problémákkal, ebben a fejezetben azonban a mérés jellemzőivel foglalkozunk. Gondoljunk csak néhány példára: az infláció, a vásárlói attitűd, az életminőség, vagy az értelmi képesség számszerűsítésének esetére. Célom a mérési skálák elméletének kialakulását bemutatni, majd körvonalazni azt a tudományos vitát, amit az irodalom méréselméleti vitaként ismer. Természetesen ma már nem minden bemutatott megállapítással értünk egyet, a tanulmányok főbb tételeinek kiemelésére a vita folyamatának bemutatása miatt szükséges. A történeti áttekintés mellett fontosnak tartom felhívni a figyelmet a módszerválasztás, a mérési skála jelentőségére, valamint megkísérlek keretet adni a Likert-skála elhelyezéséhez a skálák és alkalmazható statisztikák rendszerében.

2.1. A méréselméleti vita története

A mérési skálák napjainkban is alkalmazott típusai egy, a Harvardon tevékenykedő pszichológus, Stanley Smith Stevens (1946, 1955) klasszikus, sokat hivatkozott tanulmányaihoz kötődnek. Stevens évekig egy tudományos bizottság élén a mérés problematikájával foglalkozott, a csoport célja annak a kérdésnek a megválaszolása volt, hogy mérhető-e az emberi érzékelés, és ha igen, milyen módon. A legfőbb és talán legfontosabb vita a körül alakult ki, hogy mit is nevezhetünk mérésnek, amiben a bizottság tagjai markánsan eltérő véleményt alakítottak ki. Stevens szerint fontos felismernünk, hogy a mérésnek számos formája létezik, ennek megfelelően különböző mérési skálák definiálhatók. A mérési skála típusát egyaránt meghatározzák a mérés folyamán alkalmazott konkrét eljárások és a skála matematikai tulajdonságai. Stevens kiemeli, hogy a mérési skálától függ, hogy az adott empirikus adatok esetén mely statisztikai módszerek, eljárások alkalmazhatóak, és melyek nem. Ez a megállapítása volt a tanulmány legnagyobb visszhangot kiváltó kijelentése. A mérésre vonatkozó definíciója szerint a mérés nem más, mint számértékek hozzárendelése különböző objektumokhoz vagy esemé-

nyekhez, mégpedig meghatározott szabályok szerint. Amennyiben a mérés definíciója helytálló, a skálák problematikája visszavezethető az alábbiakra.

Meg kell határoznunk:

- a számértékek hozzárendelésének szabályait;
- az eredményként előálló skálák matematikai tulajdonságait és
- az egyes mérési skálák esetén alkalmazható statisztikai műveletek, eljárások körét.

A skálák jellemzője, hogy bizonyos hasonlóság van a megfigyelt objektumok tulajdonságai és a számsorok között. A megfigyelt egységek tulajdonságainak vizsgálatakor alapvetően a következőket vizsgálhatjuk: az egyedek tulajdonságainak egyezőséget; az egyedek jellemzőinek nagyságrendi sorrendjét; különbségeket és a különbségek egyezőségét; valamint arányokat, arányok egyezőségét. A fenti tulajdonságok leírására a (pozitív) valós számok tökéletesen megfelelhetnek, azaz – Stevens szavaival élve – a számok a valós világ jelenségeinek megfelelő modelljét adhatják.

Az elérhető skála minősége természetesen függ a mérni kívánt jelenség jellemzőitől, és a mérés konkrét folyamatától, de a szerző szerint az eredmény az alábbi táblázatban szereplő – az alapszintű statisztika tankönyvekből jól ismert – skálák egyike lesz.

2-1. táblázat: Mérési skálák és tulajdonságaik

Skála	Alapvető művelet	Matematikai csoport tulajdonság	Megengedhető statisztikai műveletek
NOMINÁLIS	Egyenlőség meghatározása	Permutációs csoport $x' = f(x)$, ahol f tetszőleges, kölcsönösen egyértelmű hozzárendelés	Esetek száma Módusz
ORDINÁLIS	Sorrendiség meghatározása	Isotonikus csoport $x' = f(x)$, ahol f tetszőleges, monoton növekvő függvény	Medián Percentilisek
INTERVALLUM	Intervallumok/különbségek egyezőségének vizsgálata	Általános lineáris csoport $x' = ax + b,\ a > 0$	Számtani átlag Szórás Rang korreláció Szorzat momentum korreláció
ARÁNY	Hányadosok egyezőségének vizsgálata	Hasonlósági csoport $x' = ax,\ a > 0$	Mértani átlag Harmonikus átlag Relatív szórás

Forrás: Stevens (1946, 678) és Stevens (1955, 113)

A skálákon megengedett, azaz elvégezhető műveleteket a fenti táblázat utolsó oszlopában soroltam fel, mely lista kumulatívan értelmezendő: az alacsonyabb rendű skálák megengedett műveletei a magasabb rendűeken is elvégezhetők. A matematikai

csoport tulajdonság oszlopban azon matematikai transzformációk kerültek felsorolásra, melyek nem módosítják a skálatípust. A műveletek megengedhetőségének feltétele az invariancia, melynek jelentését Anderson szemléletesen mutatja be: „amennyiben egy mutatót számítunk adott változóértékekből, majd transzformáljuk azt, azonos eredményt kell kapnunk, mintha az egyedi értékeket transzformáltuk volna, és így határoztuk volna meg a mutató értékét" (Anderson, 1961, 309).

A skálák számának növelésére, bővítésére több kísérlet is született. Egyrészt az intervallum skála mellett (amelyet egyenlő intervallumok skálájának is neveznek) nem egyenlő intervallumok skáláit is megkülönböztetik esetenként az irodalomban, mint például a logaritmikus skála, ahol tízes alap esetén minden intervallum pontosan a tízszerese az őt megelőzőnek. Az ilyen skálákat azért nem tekintjük külön típusnak, mert megfelelő matematikai művelettel a hagyományos intervallum skálává transzformálhatók. A megengedhető statisztikai műveletek ebben az esetben a hatványtranszformációk. A másik kiterjesztés az ún. abszolút skála, melyet Stevens is megemlít, és kardinális skálán mért számnak nevezi. A megengedett művelet kizárólag a helybenhagyó művelet. Az elképzelhető skálatípusok matematikai meghatározásával a modern méréselmélet foglalkozik, alapvető megállapításuk, hogy Stevens besorolása többékevésbé teljes, más jelentős struktúrák bizonyíthatóan nem léteznek.

A skálák megkülönböztetése mellett Stevens cikkének legnagyobb jelentősége a megengedhető statisztikai műveletek rögzítésében van. Kategorikusan kijelenti, hogy a kutatók által gyakran alkalmazott ordinális változók esetén „a hagyományos, átlagokon és szórásokon alapuló eljárásokat nem szabadna használni, hisz azok többet tételeznek fel, mint csupán az adatok relatív rangsorának ismeretét". Következő mondataiban mindenesetre már megengedőbb a szerző: az eljárások „illegális alkalmazása egyetlen dolog miatt bocsátható meg: sok esetben a vizsgálatok gyümölcsöző eredményekre vezetnek" (Stevens, 1946, 679). Hangsúlyozza, hogy az ordinális skálákon mért változók esetén számított statisztikákat óvatosan kell kezelni, különösen körültekintőnek kell lenni következtetések levonásakor. Ennek ellenére ordinális skálán mért ismérvekből számított átlagokkal azóta is találkozunk, akár a mindennapi életben, akár a tudományos kutatások területén. Ugyan tudjuk, hogy az iskolai osztályzatok nem mondanak többet az ordinalitásnál, az ösztöndíjak például mégis tanulmányi átlagtól, sőt, kreditpontokkal súlyozott átlagoktól függenek.

Stevens nagy hatású cikkei, és az ezekből levonható tanulságok követőkre találtak, melyek a társadalomtudományi módszertanokkal foglalkozó tankönyvekben hamarosan meg is jelentek (Siegel, 1956). A tudományos vita egyik, Stevens tanulmányára (is) visszavezethető fő irányvonala a paraméteres és nemparaméteres eljárások hívei között zajlott, ez az irány vezet el a statisztikai eszköztár két részre bontásához: a paraméteres és nemparaméteres eljárások megkülönböztetéséhez. A paraméteres eljárások legalább egy sokasági érték, azaz paraméter becsléséből indulnak ki, méghozzá legygyakrabban normális eloszlású sokaságból származó minták esetére. Mindez legalább intervallum erősségű skálát követel meg a Stevens tanait követő konzervatívok szerint. A nemparaméteres eljárások nem követelik meg sokasági paraméterek becslését, nem teszik fel az egység állandóságát a skála teljes értelmezési tartományán, nincsenek olyan erős előzetes feltevések a sokasági eloszlást illetően sem. A két megközelítés hipotézisrendszerei azonban nem minden esetben feleltethetők meg egymásnak teljes mértékben.

A Stevens elméletét támogató kutatók mellett mások határozottan támadták véleményét, ellenezték az abból levont következtetéseket, ezzel komoly vitákat generálva (Lord, 1953, Behan-Behan, 1954), olyannyira, hogy a témakör még napjainkban is foglalkoztatja a tudományos közösséget (Scholten-Borsboom, 2009). A Stevens tételeit egyértelműen tagadó ultraliberális szemlélet mellett rengeteg tanulmány született a paraméteres próbák robusztusságával kapcsolatban is. Ezek a tanulmányok elsősorban azt a vitát voltak hivatottak eldönteni, mely a paraméteres (pl. Anderson, 1961, Baker et al., 1966, Labovitz, 1967, Gaito, 1980) és a nemparaméteres próbákat előnyben részesítő kutatók (pl. Siegel, 1956, Thomas, 1982, Townsend-Ashby, 1984) között alakult ki. A központi kérdés emellett az ordinális és intervallum skálák megkülönböztetése volt. A paraméteres és nemparaméteres eljárások alkalmazhatósága (Gardner, 1975) a kutatókat azóta is megosztja.

Megtalálhatóak tehát szélsőséges vélemények, miszerint a nemparaméteres eljárások szinte teljesen feleslegesek, mert majdnem minden esetben alkalmazhatóak a jól ismert paraméteres megfelelőik, azok robusztussága miatt, ráadásul – érvelnek egyes kutatók – a paraméteres próbák statisztikai ereje jóval nagyobb. Ez az okfejtés eredményezte a jellemző paraméteres próbák (t-próba, F-próba stb.) robusztusságával kapcsolatban megjelenő tanulmányokat (pl. Godard et al., 1940, Eisenhart, 1947, Cochran,

1947, Bartlett, 1947, Gayen, 1949, Box, 1953, Boneau, 1960, Glass, 1972, Blair és Higgins 1980, 1985, Zimmerman és szerzőtársai 1989, 1990, 1993, Wiley et al. 2000). A Statisztikai Szemle hasábjain Vargha (2003) foglalkozott az egymintás t-próbával, szimulációk segítségével bizonyítva, hogy az alapeloszlás csúcsosságától és ferdeségétől is függ a robusztusság. Vargha (2004) mutatja be magyar nyelven a kétszempontos sztochasztikus összehasonlítás modelljét, mely ordinális változókra alkalmazható.

A későbbiekben a kutatók egyre inkább úgy vélték, hogy a skálák közötti választás, a skála erősségének meghatározása nem olyan egyértelmű, mint ahogy azt Stevens gondolta. Később Knapp (1990) már „ordinális", „kevesebb mint ordinális" és „több mint ordinális" skálákról is beszél. A Soha, Ritkán, Gyakran, Mindig ismérvváltozatokból álló skálát a legtöbb kutató úgy elemezné, hogy számokat rendel a négy kategóriához (lineárisan, vagy akár nem lineárisan). Amennyiben azonban a lehetséges válaszok például: Soha, Esetenként, Néha, Mindig lennének, úgy a két középső lehetőség sorrendjének megállapítása komoly problémát jelent. Az ordinális és intervallum skálák között létezik bizonyos átmenet, mely a használatban lévő mérési skálák jó részének sajátja. Az ilyen, „átmeneti" skálák esetén alkalmazható módszerek köre természetesen ugyancsak kérdéses.

Joel Michell (1986) cikkében mintegy válaszul az eddigiekben bemutatott tudományos vitára a méréselmélet tanait három nagy iskolára osztotta, és a kutatók közötti ellentéteket erre vezette vissza. Tanulmányában célja csupán a különböző irányzatok követőinek azonosítása, és az általuk képviselt tanok bemutatása volt, ahogy ez esetünkben is igaz. A három különálló – reprezentációs, operacionalista és klasszikus – elméletet Michell szemléletében (néhol az általa is hivatkozott, alapvető munkák mélyebb ismertetésével) mutatjuk be a következőkben. A tanulmányban a pszichológia területén tetten érhető iskolákat tárja fel a szerző, de jelenlétük valamennyi, kvantitatív, statisztikai módszereket alkalmazó tudományágban kimutatható. A reprezentációs elmélettel bővebben foglalkozunk, egyszerűen azért, mert képviselői jóval nagyobb irodalmat tudhatnak magukénak, az elmélet összetettsége, matematikai alapjai miatt.

2.1.1. Reprezentációs elmélet

Michell a reprezentációs elmélet korai megjelenésének tartja Helmholtz (1887), Hölder (1901), Russel (1903) és Campbell (1920) műveit, melyek alapul szolgáltak

Suppes (1951, 1959 és Suppes-Zinnes, 1963) munkásságához, aki az elmélet matematikai alapjait fektette le. Hölder eredeti, német nyelvű munkáját Michell és munkatársa fordították angolra (Michell-Ernst, 1996, 1997). A reprezentációs elmélet fejlődése főként néhány kutató nevéhez köthető, a megjelent tanulmányok, cikkek jó részét ők jegyzik, melyek közül néhányat az irodalomjegyzékben felsoroltam. Ennél részletesebb jegyzék található például Anwer és Hardeo (1993) és Luce (1996) munkáiban. Az alábbiakban a már megismert, Stevens-féle skálákkal kapcsolatban mutatom be a reprezentációs elmélet definícióinak jellegét, eltekintve a pontos matematikai leírástól.

Tegyük fel, hogy a vizsgálandó jelenség a hajszín. A reláció ebben az esetben, hogy két személy hajszíne megegyezik, vagy sem. Úgy kell számokat személyekhez rendelnünk, hogy bármely két személy esetén, ha x hajszíne megegyezik y-éval, akkor és csak akkor $M_x = M_y$, ahol M_x az x-hez, M_y az y-hoz rendelt szám. Nevezzük ezt, a Stevens által nominálisnak elnevezett skálát X hajszínskálának. Megengedhető (admissable) skálatranszformáció bármely egy-egy értelmű hozzárendelés.

Tekintsünk következőként egy gyenge sorrend relációt (weak order relation). Az alábbi megállapítást tehetjük: x dolgozata legalább olyan jó, mint y-é. Ha ez a reláció tranzitív és kapcsolt (connected), akkor azt leírhatjuk a matematikai $\geq$ jellel. Ekkor a hozzárendelést úgy kell elvégezni, hogy x minősége akkor és csak akkor legalább olyan jó, mint y-é, ha $M_x \geq M_y$ állítás igaz. Az eredmény egy ordinális skála. Az ordinális skálák esetén csak a (szigorúan) monoton növekvő transzformációk megengedettek.

Tekintsük valamely attribútumra vonatkozóan különbségek sorrendjét. Ekkor – néhány könnyen tesztelhető feltételezés fennállása esetén – a számok hozzárendelése megtörténhet. Amennyiben w és x közötti különbség legalább akkora, mint y és z között, akkor és csak akkor $M_w - M_x \geq M_y - M_z$. Az eredmény egy intervallum skála, és a megengedhető transzformációk halmaza valamennyi pozitív lineáris transzformáció.

Végezetül tekintsünk egy sorrendi relációt objektumok valamely jellemzőjének összegére (összekapcsolására) vonatkozóan. Példaként a fizikai hosszúságot véve, legyen A szilárd rudak egy halmaza. Bármely, A-ban lévő x és y rúdra vonatkozóan

legyen $x \cdot y$ a két rúd (végeiknél való) összeillesztéséből származó rúd ($\cdot$ az összekapcsolás jele). Amennyiben a feltételek megfelelnek az extenzív struktúra (extensive structure) követelményeinek, úgy a hosszúsági sorrend leképezhető numerikusan. Amennyiben $w \cdot x$ legalább olyan hosszú, mint $y \cdot z$, akkor és csak akkor $M_w + M_x \geq M_y + M_z$. Az eredmény egy arányskála, és a megengedhető transzformációk a hasonlósági transzformációk.

A reprezentációs elmélet hívei úgy gondolják, hogy a szóban forgó relációk jellegétől függ, hogy milyen módszerek alkalmazhatók velük kapcsolatban. A fő probléma annak a meghatározása, hogy mely skálák esetén milyen eljárások ezek. Stevens szerint az egyes skálák esetén az objektumokhoz rendelt számokkal kapcsolatban vannak nem megengedhető műveletek, melyek eredményei nem invariánsak az elvégezhető, megengedett skála-transzformációk tekintetében. Stevens invariancia definíciója azonban távolról sem volt pontos, ráadásul a későbbi vélemények szerint „a tudományban minden tény megengedhető" (Michell, 1986, 399). A megfelelő (appropriate) statisztikákról többek között Suppes, valamint Adams és szerzőtársai (1965) pontosították Stevens elképzeléseit. A logikai, többnyire intuitív, vagy néhány példán alapuló érvelést felváltotta a tételek matematikai bizonyítása a reprezentációs elmélet hívei között.

A megengedhetőség, megfelelőség fogalmához szorosan kapcsolódik a statisztikai értelmesség (meaningfulness) koncepciója, melyet Patrick Suppes fektetett le, miszerint „egy empirikus hipotézis, vagy bármilyen állítás, mely numerikus mennyiségeket tartalmaz, csak abban az esetben értelmes (meaningful), ha igazságtartalma változatlan a numerikus mennyiségek megfelelő transzformációi esetén is" (Suppes, 1959, 131). Az egyik klasszikus, gyakran idézett példamondat a következő: „Kétszer olyan magas vagyok, mint a Sears Tower" (Marcus-Roberts, Roberts, 1987, 384). A kijelentés jól láthatóan hamis, azonban értelmes, hisz igazságtartalma változatlan centiméterben, méterben, lábban, vagy hüvelykben mért magasság esetén is.

Michell az értelmesség két változatát, megközelítését különbözteti meg: a skála-specifikus megállapításokra vonatkozó és a skála-független megállapításokra vonatkozó értelmességet. Míg a skála-specifikus megállapítások tartalmaznak egy bizonyos mérési skálára vonatkozó hivatkozást, addig a skála-független megállapítások esetén ez nem igaz. Suppes előzőekben bemutatott definíciója jól láthatóan skála-specifikus megállapí-

tások esetén alkalmazható. A megközelítés egyik problémája, hogy attól függetlenül, hogy egy állítás értelmetlen, még lehet tudományos értelemben hasznos, segítségével valós következtetéseket vonhatunk le a vizsgált egyedekről. Így az értelmetlen megállapítások nem „száműzhetők" automatikusan. Amennyiben például azt állítjuk, hogy a mai hőmérséklet Celsius fokban a tegnapi kétszerese, állításunk könnyen beláthatóan értelmetlen (hisz pl. Fahrenheitben a viszony nem kétszeres lenne). Ennek ellenére hasznos, hisz tudjuk, hogy melegebb van, mint tegnap volt (0 fok feletti hőmérsékletet feltételezve). Egy másik példa szerint a hajszínt számokkal jelölve az X hajszínskálán értelmetlen megállapítást tehetünk, amennyiben azt mondjuk, hogy a mintánkban a hajszínek összege 10. Ha tudjuk azonban, hogy a vörös hajúakat 3-as számmal jelöltük, akkor az értelmetlen megállapítás haszna az, hogy tudjuk, nem minden egyed vörös hajú.

Adams és szerzőtársai (1965) az értelmesség skálafüggetlen definícióját javasolták, hiszen a mérés során célunk általában nem az, hogy skálafüggő megállapításokat tegyünk, hanem a jelenségekről szeretnénk skála-független információkhoz jutni. A szerzők által javasolt definíció lényege, hogy egy skála-független kijelentés akkor és csak akkor értelmes, ha az igazságtartalma valamennyi skála-specifikus változatának azonos. Egy skála-független kijelentés skála-specifikus változatát úgy nyerjük, hogy minden mérendő változót valamely skálára vonatkozóan írunk le, természetesen minden értékhez azonos skálát rendelve. A skála-független kijelentések értelmessége különös jelentőséggel bír, de ebben az esetben is szembesülhetünk nehézségekkel. A következő két állítás minden kétséget kizárólag értelmetlen skála-független megállapítás: A magassága 6,4; B magassága 3,2 (Michell, 1986). A két állítás alapján azonban a következő állítás tehető: A magassága kétszerese B-ének, ami kétségkívül értelmes skála-független megállapítás, és akár igaz is lehet.

A reprezentációs elmélet követői rengeteg tételt dolgoztak ki és bizonyítottak az utóbbi évtizedekben. Az értelmesség definícióján túl, ehhez kapcsolódóan az invariancia (invariance), homogenitás (homogeneity) és a megfelelő statisztikák (appropriate statistics), valamint ezek kapcsolódási pontjai álltak a kutatások középpontjában. A stevensi tanokhoz kapcsolódóan sikerült nagyrészt tisztázni azt a kérdést, hogy milyen mérési skálák lehetségesek. A kutatások alapján jól látszik, hogy Stevens nem tévedt nagyot akkor, amikor igen kevés skálát vezetett be elméletében (Narens, 1981a, 1981b).

A hasznosságelméletek területén értek el további jelentős eredményeket a reprezentációs elmélet követői. Az extenzív struktúrákon kívül egyéb struktúrák feltárása is sikerrel járt, melyek közül a conjoint struktúra (Luce-Tukey, 1964, Tversky, 1967) a legjelentősebb, melynek gyakorlati alkalmazása mára szélesebb körben elterjedt, igen jelentős a már említett hasznosság mellett az attitűdök, képességek stb. mérésében is. Bizonyítható, hogy amennyiben a matematikai axiómák teljesülnek, úgy az eredmény intervallum skála erősségű lesz (Krantz et al., 1971).

A reprezentációs elmélet nézeteinek szélesebb körben történő elterjedését leginkább az hátráltatja, hogy a gyakorlatban alkalmazott módszerek döntő többsége az elmélet szerint nem minősül mérésnek, így az azokat felhasználó kutatók nem mutatnak kellő érdeklődést. A matematikai-logikai tételek ráadásul nehezen érthetőek, sok szempontból túlságosan elméletiek (Velleman-Leland, 1993). A mérési hiba jelensége nem került beépítésre a reprezentációs elmélet logikai keretrendszerébe, ami kritikákat váltott ki, melyre azonban született reakció (Luce-Narens, 1994). A témakörrel foglalkozó, legfrissebb kézikönyveket Narens (2002, 2007) jegyzi. Az érdeklődőknek ajánljuk továbbá a Krantz, Luce, Suppes és Tversky által jegyzett háromkötetes sorozatot (1971, 1989, 1990).

2.1.2. Operacionalista elmélet

A társadalomtudományok területén rengeteg olyan információforrás van, amely kvantitatív eredményeket szolgáltat, reprezentációs értelemben mégsem tekinthetjük azokat mérésnek. Gondoljunk egy szellemi képességeket mérő tesztre. A teszt több kérdésből áll, és miután az egyén kitöltötte, az eredmény a kérdéseknek megfelelő számú helyes/helytelen válaszokat tartalmazó sor lesz. Az adatokkal kapcsolatos fontos empirikus reláció az, hogy A személy legalább azokat a kérdéseket jól megválaszolta-e, mint B. Ebben az esetben A teljesítménye legalább olyan jó, mint B-é. Amennyiben ez a reláció tranzitív és kapcsolt valamennyi válaszadóra, úgy az eredmény ordinális skálán mért (a reprezentációs elmélet alapján). A fenti eset azonban a legritkábban fordul elő, így ez a teszt még az ordinális erőt sem éri el. A fenti empirikus kapcsolat vizsgálata helyett a tesztet végzők összeadják a helyes válaszok darabszámát, és ezt az értéket tekintik az adott személy tesztértékének, ebben az esetben azonban nem világos az, hogy mi az a reláció, amit vizsgálunk. Az operacionalista definíció szerint a mérés nem

más, mint egy művelet, ami számot eredményez, ami hasonló Stevens értelmezéséhez. A számok tehát műveletek eredményei: „a szigorúan vett operacionalista számára a tudomány egyszerűen a műveletek tanulmányozása, nem pedig a valóságé" (Michell, 1986, 404). Ebben a tekintetben pedig a számok valóban nem tudják, honnan jöttek, azaz szabadon végezhetőek velük műveletek, a skáláknak és a statisztikai módszereknek egymáshoz nincs közük. Mindez azonban a mérés és a tudomány kapcsolatát egészen más megvilágításban mutatja be, mint a reprezentációs elmélet esetén.

2.1.3. Klasszikus elmélet

A klasszikus elméletet Michell egészen Arisztotelészig és Euklideszig vezeti viszsza. Az elmélet szerint a mérés nem más, mint annak a megállapítása, hogy az egység hányszor szerepel egy adott mennyiségben. Mindez a mérés, a mérhető jellemzők körét erősen leszűkíti, bár a reprezentációs elmélettel szemben nem követeli meg a vizsgált objektumok közötti empirikus kapcsolatrendszer létét. A klasszikus elmélet szerint a mérés nem számok hozzárendelését jelenti objektumokhoz, ahogyan azt a reprezentációs és operacionalista tábor véli. A klasszikusok szerint a mérés nem más, mint számszerű kapcsolatok felfedezése, feltárása a kvantitatív jellemzők között.

Az eltérő definícióból adódik, hogy a klasszikusok esetén nem beszélhetünk mérési skálákról, hiszen a számok mindig ugyanabból a folyamatból, a mennyiségi kapcsolat (arányosság) feltárásából származnak. Ez a felfogás leginkább a stevensi arányskála tulajdonságait hordozza, az elvégezhető műveletek köre a lehető legszélesebb, mérési skálák nincsenek, így ezek korlátozást sem jelentenek az alkalmazható statisztikák tekintetében.

2.2. Napjaink gondolatai

A statisztikai szoftverek elterjedésével párhuzamosan ismét felvetődött a kérdés, hogy a mérési skálák befolyásolják-e, és ha igen, mennyiben az alkalmazható módszereket. Szükséges-e, hogy a szoftverek beépítetten korlátozzák a felhasználót az elérhető módszerek tekintetében a mérési szintet figyelembe véve. Velleman és Wilkinson (1993) azt állítják, hogy a korlátozás néhol felesleges, sőt rossz lehet. Legfontosabb megállapításuk, hogy a skála típus nem tisztán az adatok jellemzője, hanem attól is függ, mi a kérdésfeltevésünk. Hasonló gondolattal találkozhatunk Surányi-Vita (1972)

tanulmányában is, akik a keresetek példáján keresztül tárgyalják ugyanezt a jelenséget. Pusztán pénzügyi szempontból a kereset arányskálán (vagy abszolút skálán) mért jellemző, közgazdasági, vagy szociológia szempontú kérdésfeltevés esetén azonban nem egyértelmű, hogy ugyanez igaz-e. Hand (1996, 2004) véleménye szerint minden gyakorlati életben történő mérés egyfajta keveréke a reprezentációs és a pragmatikus nézőpontoknak. A skálák abban különböznek, hogy a két szemlélet különböző „súlyokkal" szerepel bennük.

Michell napjainkban is a mérés elméletével, a társadalomtudományok területén betöltött szerepével foglalkozik. Munkái (1994, 1999, 2005, 2008) szélsőségesen kritikus áttekintést adnak a területről: arról ír, hogy súlyos hiba bizonyíték hiányában elfogadni azt az állítást, hogy egyes jellemzők kvantitatívak. Természetesen Michell saját tudományterületének, a pszichológiának a méréseiről mond véleményt, de gondolatai más területek mérési módszereire is vonatkoznak. Szélsőséges véleménye szerint komoly erőfeszítések soha nem történtek a különböző jellemzők kvantitatív voltának bizonyítására, a méréseket végzők azt vizsgálatok nélkül elfogadják. Michell szerint nem is ez a legnagyobb probléma, hanem az, hogy az empirikus adatokkal dolgozó kutatóknak eszükbe sem jut, hogy a vizsgált jelenség esetleg nem is kvantitatív, azaz a pszichometria (és minden méréseken alapuló társadalomtudomány) ilyen értelemben „kóros tudomány" (pathological science). Michell szerint túl nagy hatással volt a tudományra Stevens tág mérésdefiníciója, ami minden olyan folyamatot, ami számértékeket eredményez, mérésnek tart. Két olyan körülményt azonosít Michell (2008) tanulmányában, melyek megmagyarázhatják ezt az állapotot. Az első ok ideológiai: a tudományosság (scientism). Sokan a mai napig is úgy gondolják, hogy tudományos megismerés csak mérés útján érhető el, így az egyben a tudományosság mérőfoka. A mérés és a statisztikai módszertan bevezetése a (kvantitatív) tudományok körébe emelte az pszichológiát. Ehhez kapcsolódik a gyakorlatiasság (practicalism) szükségességének elterjedése, ami a tudomány sikerét gyakorlati problémák megoldásában méri, szemben azzal a klasszikus tudományfelfogással, ami a vizsgált rendszer megértését helyezi előtérbe. A másik, Michell által említett ok gazdasági jellegű: komoly, méréseken alapuló tudományok könnyebben kaptak a világháború után állami, vagy ipari megbízásokat, így a pszichológiai jellemzőknek egyszerűen kvantitatívnak *kell*ett lenniük. Ez a vélemény természetesen messzemenőkig szélsőséges, a tudományos közvélemény által erősen

vitatott, azonban megemlítését fontosnak tartom. Michell természetesen ezzel nem csak a pszichológiai mérést véleményezi, hanem gyakorlatilag a társadalomtudományok döntő többségét és némely természettudományt is. A túlságosan radikális vélemény azonban felhívhatja a figyelmet arra, hogy nem csupán attól válhat valami tudománnyá, tudományossá, ha mérések társulnak hozzá; ha egyes jelenségek nem (megfelelően) mérhetők, merjük ezt kijelenteni.

Hasonló folyamatoknak más területeken is tanúi lehetünk: a marketingkutatás, a szociológia és sok társadalomtudomány attitűdök, képességek, belső értékek vizsgálatát végzi. Rengeteg mérési módszer került kidolgozásra, melyek a mérni kívánt jellemzők széles skáláját fedik le, például: szükségletek, preferenciák, önképek, értékek, érzelmek, reakciók stb. mérése. Bearden és Netemeyer (1999) például egy több száz mérési módszert felsorakoztató kötetet állított össze, azok eredeti megjelenésével, rövid leírásával, az eddigi kutatási tapasztalatokkal. Ezek a mérési módszerek így mintegy best practiceként terjednek egy-egy nevesebb kutató, vagy kutatócsoport publikációi nyomán. Mindez azt okozza, hogy a hasonló témakörben tevékenykedő tudósok hasonló módon készítik el kérdőíveiket. Kérdés természetesen, hogy adott jelenség mérhetőségét mi módon állapíthatjuk meg.

Meg kell ugyanakkor jegyeznünk, hogy napjainkra egyszerűen nem tartható Anderson „kifogása” az ordinális és nominális ismérvekkel kapcsolatos módszerek szűkösségével kapcsolatban. A nemparaméteres eljárások és kategóriás adatok elemzésére alkalmas módszerek sokasága került kifejlesztésre az utóbbi évtizedekben, nem kis mértékben a fentiekben bemutatott méréselméleti vita folyományaként. A nemzetközi standard kézikönyvként Agresti (2002) művét alkalmazzák, oktatják a klasszikus statisztika területén, a kategóriás adatok bayesi modelljeit bemutató munkát pedig Congdon (2005) jegyzi. Magyar nyelven a legátfogóbb összefoglalót Vargha (2008) műve adja, mely részletesen bemutat ordinális skálákon végezhető műveleteket és teszteket is, különös tekintettel a sztochasztikus egyenlőségek és különbségek vizsgálatára. A szükséges módszerek tehát rendelkezésre állnak, a kutató felelőssége azok megismerése és alkalmazása, amennyiben az adatok jellemzői ezt megkívánják.

Ne feledjük azonban, hogy a számok nem tudják honnan jöttek, de tegyük hozzá, hogy a szoftverek sem, így a kutatónak kell azt észben tartania. Az adatainkból bármilyen mutatót kiszámíthatunk, bármilyen modellt illeszthetünk, de vigyázzunk a belőlük

levonható következtetésekkel! Amennyiben hipotéziseket tesztelünk, csak értelmes (meaningful) kérdéseket tegyünk fel, nehogy levont következtetésünk csak a skála sajátja legyen. Ha szükséges, tegyünk skálafüggő kijelentéseket skálafüggetlenek helyett, ezzel jelezve, hogy más mérési módszerrel akár más eredményeket is kaphattunk volna. A és B csoport vevői elégedettségének mérése esetén megállapításunkat fogalmazzuk meg úgy, hogy A csoport alacsonyabb elégedettségi pontszámmal rendelkezik, mint B csoport az adott mérési módszerrel, ne pedig úgy, hogy A csoport kevésbé elégedett a termékkel/szolgáltatással. Ha nem vagyunk biztosak abban, hogy az elégedettség mérése ordinálisnál erősebb skálát eredményez, alkalmazzunk ennek megfelelő teszteket. Ezzel szemben tegyünk nyugodtan skálafüggetlen kijelentéseket, ha ezt a körülmények megengedik. Az elemzési módszertan kiválasztásakor legyünk egészségesen szkeptikusak azok mérési szintjével kapcsolatban, de vegyük figyelembe a felhasználási területet is. Nem várhatjuk, hogy a már említett iskolai osztályzatok alapján két tanulócsoport teljesítményét ezentúl valamilyen nemparaméteres próbával hasonlítja össze a tanulmányi osztály. Ennek ellenére egy tudományos munkában a megfelelő eljárások és számítógépes háttér birtokában mindez már nem okozhat problémát.

Ne felejtsük el, hogy egy-egy változó több információt tartalmazhat, mint az első ránézésre látszik. A magyar gépkocsik rendszáma például nominális skálán mért. Ez nem jelenti azt, hogy egy autó rendszámából (és egyéb kiegészítő információkból) ne tudnánk igen fontos következtetéseket levonni! Az autó kora és rendszáma között igen szoros a kapcsolat: egy autókereskedő a rendszámból és az autó típusából igen pontosan meg tudja mondani, hogy Magyarországon eladott, vagy külföldről behozott autóról van-e szó? Mivel a rendszámokat az okmányirodák „csomagokban” kapják, az azonos betűkből álló rendszámok egy területen csoportosulnak, a rendszám így a forgalomba helyezés helyére is utalhat. Ugyancsak többletinformációt szolgáltat az, hogy kért rendszámmal rendelkezik a gépkocsi: nagy valószínűséggel céges gépjárműről van szó, vagy tehetős tulajdonosról. A rövid példa annak érzékeltetésére szolgál, hogy a skálatípus nem feltétlenül keverendő össze az információtartalommal. Vegyük észre, hogy a gyakorlati esetekben a változók mérési skálához rendelése közel sem triviális, valamint a mérési skála egyazon adatsornál függhet a felhasználási céltól. A rendszám nem vált ugyan ordinális skálán mért ismérvvé, a változó által hordozott információt azonban kár

lenne elveszíteni. Ilyen értelemben a reprezentációs elmélet képviselője számára a szigorú skálafeltételek fontosak, az operacionalista számára pedig az információtartalom.

A fejezetben a méréselmélettel és skálákkal kapcsolatos tudományos vitát mutattam be az adott időszakokra jellemző publikációkon keresztül, ami jól mutatja, hogy a tudományos világban ma sincs konszenzus a témakörrel kapcsolatosan. A méréselmélet fogalmainak és történetének megismerése az alkalmazott kutatásokat végző tudósok számára alapvető fontossággal bír, és különös jelentősége van további témánk, a kategóriás adatok megkülönböztetése során. Annak eldöntése, hogy a vizsgálatban szereplő változók teljesítik-e az intervallum skála kritériumait, minden esetben a kutató feladata és felelőssége. A következőkben bemutatandó módszertani megfontolások legalább intervallum skálán mért változókat tételeznek fel, azaz azt, hogy a szórás és a variancia megfelelő, értelmes (meaningful) leírását adja a heterogenitásnak. A továbbiakban a kevés kimenettel rendelkező, (közel) intervallum skálán mért változókat az egyszerűség kedvéért Likert-féle skálán mért, vagy röviden Likert-skálás változónak nevezem.

3. Mintaelemszám tervezés Likert-skála esetén

Equation Section (Next)Ebben a fejezetben a Likert-skála esetén alkalmazható mintaelemszám meghatározás módszerével foglalkozom. Elsőként magával a skálával, majd a mintaelemszám meghatározásának hagyományos (kétkimenetelű változókra épülő) módszerét tekintem át, bemutatom a meghatározáshoz szükséges képleteket (a levezetéseket jórészt a Függelékben), valamint a szükséges mintaelemszám várható értékét is meghatározom.

3.1. A Likert-féle skála története, jellemzői

A Likert-skálát első alkalmazójáról, Rensis Likert-ről nevezték el[1]. Létrehozásának célja adott egyén adott tevékenységekkel, illetve fogalommal kapcsolatos attitűdjének vizsgálata volt. Szerkezetét tekintve ezen attitűd-skála 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 egyre gyakrabban 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 gyakran alkalmazzák kérdés-csoportok formájában is, vagyis egy-egy vizsgálandó területre vonatkozóan nem egy, hanem több – estenként 20, sőt 100 – állítást fogalmaznak meg, és ezen állításokra adott összegzett, illetve átlagolt válasz-értékekkel dolgoznak tovább.

A szakirodalomban sokan foglalkoznak azzal a kérdéssel, hogy hány fokozatú skála alkalmazása javasolt (Cox, 1980, illetve Preston, Colman, 2000), páros, vagy páratlan számú válaszlehetőség megadása megfelelőbb-e (Coelho, Esteves, 2007). Hogyan

[1] Rensis Likert (1903-81), a róla elnevezett skála első kifejtését tartalmazza Likert (1932).

kell a kérdéseket megfogalmazni, érdemes-e minden lehetőséget szövegesen megnevezni, vagy csak a két végpontot, a számok hozzárendelését pedig a kitöltő végezze el saját belátása szerint. További vizsgálatok születtek akár azzal kapcsolatosan is, hogy befolyásolja-e a válaszadókat az, hogy milyen sorrendben szerepelnek a válaszlehetőségek: balról jobbra, vagy jobbról balra emelkednek-e az értékek (Nicholls et al., 2006). Azt az eredményt kapták, hogy a vizuálisan bal oldalon elhelyezett válaszlehetőségeket előnyben részesítik a kitöltők, így amennyiben kedvező képet akarunk festeni, bal oldalra a kedvező válaszokat érdemes elhelyezni. Az alkalmazás széleskörűségét mutatja, hogy a Likert skálákkal kapcsolatban vizsgálták a kultúrák közötti összehasonlíthatóság mértékét (Heine et al., 2002), hogy milyen módon alkalmazható gyermekek (Laerhoven et al., 2004) valamint szellemi fogyatékosok esetén (Hartley, MacLean, 2006). A szakirodalmat áttekintve általánosan elmondható, hogy ezeket a skálákat leggyakrabban az értékek átlagával írják le a kutatók, azaz implicite élnek azzal a feltételezéssel, hogy intervallum erősségű változókról van szó, hiszen – ahogyan azt az előző fejezetben tárgyaltam – csak az ilyen skálák esetén van értelme az átlagszámításnak.

3.2. A mintanagyság tervezésének általános módja

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. Amikor a mintavétel tervezője és megrendelője egy gyakorlati probléma megoldása során egymással szembe kerül, gyakorlatilag ellentétes érdekeik vannak: a megbízhatóság és/vagy 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, ám e tekintetben a módszertudomány is elég kevés kézzelfogható választ kínál.

Mint említettük a reprezentatív mintavétel alapján történő kutatások tervezésének egyik legfontosabb problémája (Pintér-Rappai, 2001) a minta nagyságának (mintaelemszám) meghatározása. A közismert statisztikai gyakorlat a minta nagyságá-

nak 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, az aszimptotikus hibahatár:

$$\Delta = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \tag{3.1}$$

ahol a statisztikában szokásos jelölésekkel $z_{1-\alpha/2}$ a standard normális eloszlás megfelelő kvantilise, n a mintaelemszám, p pedig az arány.

A (3.1) egyenlet mintaelemszámra történő rendezése és a leggyakrabban alkalmazott 95,5%-os megbízhatósági szint $(1-\alpha)$ feltételezése mellett

$$n = \frac{4p(1-p)}{\Delta^2} \tag{3.2}$$

A szükséges mintaelemszám tehát függ a keresett sokasági aránytól, így a leggyakoribb választás, hogy a $p(1-p)$ kifejezés maximumát keressük meg, ami a $p = 0{,}5$ esetben adódik. Ebből a szükséges mintaelemszám gyakran alkalmazott becslése könnyen adódik:

$$n = \frac{1}{\Delta^2} \tag{3.3}$$

Néhány kitüntetett, gyakran alkalmazott hibahatár mellett szükséges elemszámokat mutatja a 3-1. táblázat:

3-1. táblázat: Szükséges mintaelemszámok,
95,5%-os megbízhatósági szint és különböző hibahatárok esetén

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

A 3-1. táblázat értelmezése szerint, ha egy eldöntendő kérdésre adott válasz esetén az igen válasz aránya $100p\%$, és a felmérést végzője törekszik arra, hogy 95,5%-os megbízhatósággal azt állíthassa: az alapsokaság $100p \pm 1\%$-a válaszolna igennel, akkor

maximum 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. Az alkalmazott képet némiképpen árnyalhatja az egyszerű véletlen (EV) mintavétel, illetve a maximálisnál kisebb variancia feltételezése, ám mindez a továbbiak megértését nem érinti. Amennyiben a felmérés kezdete előtt prekoncepcióval rendelkezünk a válaszadók arányával kapcsolatban, úgy kisebb mintaelemszám is elégséges lehet, főként abban az esetben, ha p, vagy $1-p$ relatíve kicsi értéket vesz fel.

Az igen–nem típusú feleletválasztós kérdések vonatkozásában a legfontosabb alapstatisztika a korábban már vizsgált arány. Ugyanakkor gyakran fordul elő, hogy egy két-kimenetelű kérdésre adható feleletet 1-gyel, illetve 2-vel jelöljük, és ezt követően nem az arányra, hanem a válaszok várható értékére vagyunk kíváncsiak. Ekkor a 3-2. táblázatban feltüntetett, százalékpontban felírt hibahatárok helyett használhatunk „pontértékben" mért Δ-t is, vagyis a szükséges mintaelemszám az alábbiak szerint alakul[2]:

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

3-2. táblázat: Szükséges mintaelemszámok,
95,5%-os megbízhatósági szint és különböző hibahatárok esetén

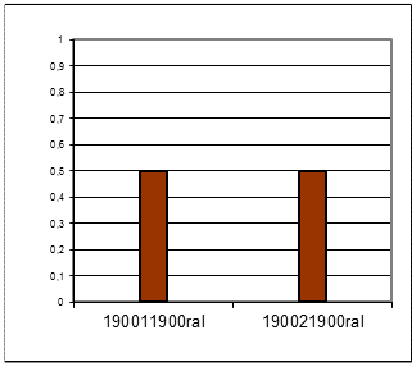
Δ (pontérték)	n
0,005	40 000
0,010	10 000
0,025	1 600
0,050	400

A 3-2. táblázat eredményei – az előzőek analógiájára – tehát úgy interpretálhatók, hogy ha egy konkrét válaszra adott feleletek átlaga, várható értéke esetében a második tizedes helyi értékben „biztos" akarok lenni, akkor 40 000 elemű; ha csak az első tizedes fontos számomra, akkor 400 elemű mintára van szükség. Összességében tehát kijelenthetjük, hogy a fenti hüvelykujj-szabállyal, viszonylag kevés statisztikai előképzett-

[2] 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.

séggel 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 közvéleménykutatást.

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, vagyis (a későbbiekben alkalmazandó jelöléseket használva) a válaszok empirikus eloszlása az alábbi:



3-1. ábra: Teljesen megosztott válaszadók, két kimenetelű kérdés esetén

Ugyanakkor szintén nem elhanyagolható probléma, hogy egy felmérés kérdéseinek jelentős része (zöme) nem eldöntendő, hanem több-kimenetelű feleletválasztós (diszkrét), illetve mért adat (folytonos). Az előbbi kérdéstípus esetén a társadalomtudományokban elterjedtek a pontozási, vagy attitűd skálák, ezen belül is a már említett Likert-skála, amely k fokozatú skálának felel meg. Az 1. fejezetben bemutatottak alapján természetesen felmerül a kérdés, hogy az ilyen módon nyert számszerű értékek milyen mérési skálának felelnek meg. Amennyiben a reprezentációs elmélet híveinek véleményét, nézeteit vizsgáljuk, nem igényel különösebb bizonyítást, hogy az így nyert adatok nem tekinthetők arányskálán mérteknek, sőt, talán a sorrendiség sem bizonyítható különböző megfigyelések esetén[3]. A gyakorlati kutatások során azonban elsősorban az operacionalista elvek kerülnek előtérbe, a Likert-skálákat alkalmazók vagy azt tétele-

[3] A kutatók előtt ismert az a jelenség, hogy bizonyos válaszadók csupán a skála egyik, vagy másik oldalát használják, így vannak pozitívan és negatívan gondolkodók is.

zik fel, hogy a válaszlehetőségek közötti távolságok egyenlők, vagy azt, hogy a távolságkülönbségek kicsik, elhanyagolhatóak, legalábbis az alapvető következtetéseket nem változtatják meg.

A fejezet további részében a mintanagyság tervezésének kérdéseivel foglalkozom Likert-skálán mért válaszokat tartalmazó kérdőívek esetén.

3.3. A szükséges mintaelemszám meghatározása Likert-skálás kérdések esetén

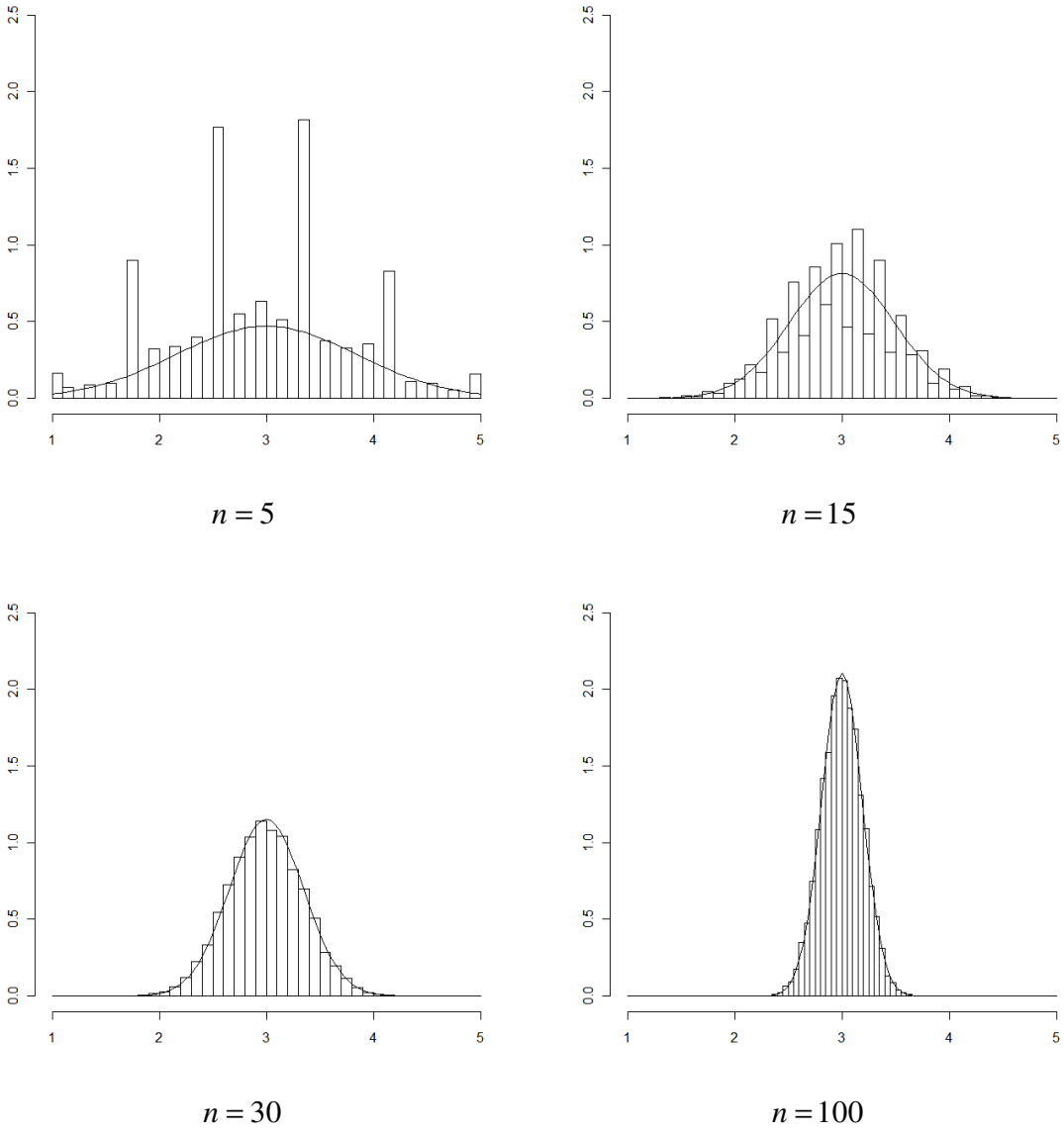
A Likert-skálás lekérdezések, vagyis a kettőnél több válaszlehetőséget tartalmazó kérdések esetén, a mintanagyság meghatározásának problémája azonos a korábban tárgyalttal: meg kívánjuk határozni a szükséges mintaelemszámot, előre adott hibahatár és rögzített megbízhatósági szint mellett. Elégségesen nagy mintaelemszám mellett a mintaátlag normális eloszlást követ, azaz a hibahatár általános képlete az alábbira módosul:

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

ahol σ a kérdésre adott válaszok elméleti (alapsokasági) szórása. 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, példaként álljon itt egy szimulációs eredmény.

A bemutatott szimulációt a „normálistól" erősen eltérő, öt kimenettel rendelkező eloszláson végeztem, az egyes kimenetelekhez tartozó valószínűségek rendre 0,44; 0,04; 0,04; 0,04; 0,44. Az egyes mintaelemszámok mellett 10 000-10 000 iterációt végeztem, majd ábrázoltam az átlagok empirikus és a normális eloszlással közelítő elméleti eloszlását.

A 3-2. ábrán látható négy hisztogram és sűrűségfüggvény alapján jól látható, hogy $n = 100$ esetén a normális eloszlás már alkalmazható, azaz a központi határeloszlás tétele már érvényesül. Amennyiben a válaszlehetőségek száma nagyobb, illetve az eloszlás egymóduszú, a konvergencia még gyorsabb. A tetszőleges k-ra és eloszlásra futtatható R program a függelékben megtalálható, mely közvetlenül a bemutatott típusú ábrákat szolgáltatja.



3-2. ábra: Mintaátlag eloszlása különböző mintaelemszámok mellett

A hibahatár képletéből kifejezhető a szükséges mintaelemszám (a korábban már említett, leggyakrabban alkalmazott feltevések és 95,5%-os megbízhatóság mellett):

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{\Delta} \right)^2 = \left(\frac{2\sigma}{\Delta} \right)^2 \tag{3.4}$$

Láthatjuk, hogy a minta nagysága az előre adott feltételektől, valamint az alapsokasági varianciától függ. Ez utóbbi Likert-skála esetén a – viszonylag kevés számú – válaszlehetőségből tulajdonképpen könnyen kifejezhető. Jelen fejezetben azzal foglalkozom, hogy milyen típusú alapsokasági eloszlások feltételezése lehet reális, illetve

melyik eloszlástípus milyen alapsokasági varianciát eredményez, áttételesen mekkora mintaelemszámot tesz szükségessé. A gondolatmenet tehát a következő: különböző eloszlástípusokat definiálok, majd ezek esetében meghatározom az elméleti (adott típusú eloszlást követő alapsokaság esetén az alapsokasági) szórást, majd ennek felhasználásával felírom a standard hibát, majd kiszámítom a szükséges mintaelemszámot.

Általánosságban a sokasági variancia az alábbiak szerint határozható meg:

$$MSS = p_1(1-\bar{x})^2 + p_2(2-\bar{x})^2 + \ldots + p_k(k-\bar{x})^2 = \sum_{j=1}^{k} p_j(j-\bar{x})^2 \qquad (3.5)$$

ahol:

- p_j a j. kategória relatív gyakorisága $\sum_{j=1}^{k} p_j = 1$;
- $\bar{x} = \sum_{j=1}^{k} p_j j$ az átlag.

Annak érdekében, hogy szórás nagyságát könnyebben meg tudjuk határozni, illetve, hogy a bemutatandó eloszlások érthető struktúrába kerüljenek, a továbbiakban kétféle alapsokasági eloszlástípust különítek el:

- szimmetrikus eloszlások,
- aszimmetrikus megítélésű kérdések.

Néhány levezetésben a későbbiekben szét kell választani a páros és páratlan válaszlehetőségek esetét. Ezekben az esetekben a képletekben $*$ lesz a páratlan, $**$ pedig a páros esetekre vonatkozó jelölés. A csillagok nélküli képletek általános érvényűek, azaz nem függenek k paritásától.

3.3.1. Szimmetrikus eloszlású válaszadások

Könnyen belátható, hogy szimmetrikus eloszlások esetén a kérdésekre adott válaszok átlaga páros és páratlan számú válaszlehetőségek esetén is az alábbi:

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

Ekkor a fenti (3.5) képlet az alábbi, egyszerűbb formában írható:

$$MSS^{(k)} = \sum_{j=1}^{k} p_j \left(j - \frac{k+1}{2} \right)^2 \tag{3.6}$$

Illetve

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

A páratlan tagszámú esetekben az átlag egy tényleges kimenetel értékét veszi fel, így egy tag az eltérés-négyzetösszeg számítása esetén kiesik, ami a kalkulációt gyakran megkönnyíti.

$MSS^{(k)}$ értéke maximális (ami egyben „globális" maximumot is jelent, nem csak a szimmetrikus esetek között), ha

$$p_1 = p_k = \frac{1}{2} \quad \acute{e}s \quad p_2 = p_3 = \ldots = p_{k\text{-}1} = 0$$

vagyis az eloszlás kétmóduszú, mégpedig a két módusz az extrém értékeknél található. A továbbiakban ezt az esetet extrém kétmóduszúnak (EKM) nevezzük. Ekkor a variancia:

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

ami az alábbi szórást eredményezi:

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

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

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

Vagyis képezhető a 3-3. táblázat analógiájára, különböző méretű Likert-skálák esetére[4]:

[4] Vegyük észre, hogy a korábban tárgyalt alternatív (két-kimenetelű) ismérv speciális eset.

3-3. táblázat: Szükséges mintaelemszámok, 95,5%-os megbízhatósági szint és különböző hibahatárok mellett az extrém kétmóduszú sokaságok esetén

Δ	Válaszlehetőségek száma (k)				
	5	6	7	8	9
0,005	640 000	1 000 000	1 440 000	1 960 000	2 560 000
0,010	160 000	250 000	360 000	490 000	640 000
0,050	6 400	10 000	14 400	19 600	25 600
0,100	1 600	2 500	3 600	4 900	6 400

Láthatjuk a 3-3. táblázat alapján, hogy Likert-skála alkalmazása során mindig lényegesen nagyobb mintára van szükségünk, mint a korábban feltételezett. Ne feledjük azonban, hogy a fenti értékek extrém eloszlású válaszadást feltételeznek, vagyis vélelmezhetően túlbecsülik a szükséges mintaelemszámot. Egy 9 fokozatú skálán 0,005 pontosság egészen más értelmet nyer, mint bináris változó, illetve az ennek megfelelő „kétfokozatú Likert-skála" esetén. A későbbiekben erre a problémára visszatérek.

Az előzőekben tárgyalt maximális variancia mellett, nyilvánvalóan felírható a minimális $MSS^{(k)}$ is, ami az alábbi (nem feltétlenül szimmetrikus) esetben áll elő:

$$p_1 = p_2 = \ldots = p_{c-1} = p_{c+1} = \ldots = p_{k\text{-}1} = p_k = 0$$

$$p_c = 1 - \sum_{\substack{j=1 \\ j \neq c}}^{k} p_j = 1$$

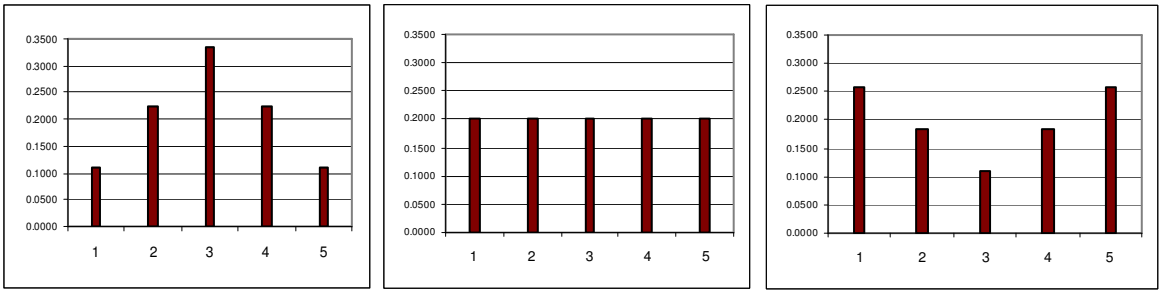
ilyenkor $MSS^{(k)}$ értéke 0. Ez elméletileg azt jelenti, hogy nincs szükség mintavételre, hisz nincs szóródás a sokaságon belül.

Mindez tehát azt jelenti, hogy Likert-skálás lekérdezés esetén a szükséges mintaelemszám 0 és $\left(\frac{k-1}{\Delta}\right)^2$ intervallumban mozog. Célom, hogy ennél a tág intervallumnál szűkebb intervallumot határozzak meg a szükséges mintanagyság tervezésénél, annál is inkább, hiszen egyik eset sem túl valószerű! Az extrém kétmóduszú esetben nehezen érthető, hogy miért van szükség többfokozatú skálára, hiszen a válaszadók csak két kimenetelt használnak; az extrém egymóduszú esetben pedig mintavételre sincs szükség, hiszen mindenki azonos módon viseltetik a megfogalmazott állítással szemben. Ebből következően a továbbiakban olyan válaszadási megoszlásokkal foglalkozom, melyeknek statisztikai szempontból jó tulajdonságaik vannak, és emellett az extrém eseteknél életszerűbbek.

Az alábbiakban a szimmetrikus eloszlás-típusok két csoportját mutatom be, az ún. lépcsős és a normálison alapuló eloszlásokat.

Lépcsős eloszlások

A lépcsős eloszlások jellemzője, hogy alapvetően a piramis (PIR) típusú eloszlásra épülnek, mely úgy épül fel, hogy a különböző lehetőségekre adott válaszok gyakoriságai egymás többszörösei egészen a móduszig, majd a gyakoriságok folyamatosan csökkennek. A különböző lépcsős eloszlásokat jól szemlélteti ötfokozatú Likert-skálák esetén a 3-3. ábra.



3-3. ábra: Lépcsős eloszlások $k = 5$ esetén
(piramis, egyenletes és fordított piramis)

Legyenek a válaszadások relatív gyakoriságai rendre:

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

illetve

$$p^{**}; 2p^{**}; 3p^{**}; \ldots; \left(\frac{k}{2}\right)p^{**}; \left(\frac{k}{2}\right)p^{**}; \ldots; 3p^{**}; 2p^{**}; p^{**}$$

Mivel a súlyok összege 1, ezért adódik:

$$p^* = \frac{4}{(k+1)^2} \qquad p^{**} = \frac{4}{(k+1)^2 - 1}$$

Ismert, hogy az átlag:

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

Ekkor páratlan esetben

$$p^* = \frac{1}{\overline{x}^2} \text{ és } \frac{k-1}{2} = \overline{x} - 1 .$$

A szórásnégyzet a következőképpen adódik (páratlan esetre lásd Függeléket, a páros eset analóg módon vezethető le):

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

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

3-4. táblázat: Szükséges mintaelemszámok, 95,5%-os megbízhatósági szint és különböző hibahatárok mellett a piramis típusú eloszlások esetén

Δ	Válaszlehetőségek száma (k)				
	5	6	7	8	9
0,005	213 333	306 667	400 000	520 000	640 000
0,010	53 333	76 667	100 000	130 000	160 000
0,050	2 133	3 067	4 000	5 200	6 400
0,100	533	767	1 000	1 300	1 600

Amennyiben az eloszlás az egyenletes eloszlás felé közelít, úgy a szórás egyre nagyobb lesz a változatlan átlag mellett. Az egyenletes eloszlás esetén a variancia az alábbi 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 szükséges mintaelemszám:

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

Az egyenletes esetben a páros és páratlan elemszámok esetére a képletek megegyeznek, a szükséges mintaelemszámok a 3-5. táblázatba rendezhetőek:

3-5. táblázat: Szükséges mintaelemszámok, 95,5%-os megbízhatósági szint és különböző hibahatárok mellett az egyenletes eloszlású sokaságok esetén

Δ	Válaszlehetőségek száma (k)				
	5	6	7	8	9
0,005	320 000	466 667	640 000	840 000	1 066 667
0,010	80 000	116 667	160 000	210 000	266 667
0,050	3 200	4 667	6 400	8 400	10 667
0,100	800	1 167	1 600	2 100	2 667

Amennyiben a szélsőséges válaszok felé történő átrendeződés folytatódik, egyre nagyobb lesz a szórás. A következő sarkalatos eloszlás az ún. fordított piramis (FPIR) eloszlás. Az eloszlás a következő képlet alapján határozható meg:

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

ahol p_j a megfelelő tagszámú piramis típusú eloszláshoz tartozó valószínűség.

A fenti képlet biztosítja, hogy

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

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

teljesüljenek, vagyis az eloszlás két azonos valószínűséggel előforduló, különböző maximummal rendelkezzen, mégpedig a két szélső, extrém értéknél, valamint azt is, hogy a súlyok összege 1 legyen.

Ekkor a mintaelemszám a következőképpen adódik (lásd Függelék a páratlan esetre, a páros eset analóg módon levezethető):

$$n^{FPIR*} = \frac{(k-1)\left(k^2-3\right)}{3(k-2)\Delta^2} \qquad n^{FPIR**} = \frac{(k-1)\left(k^2-3\right)-1}{3(k-2)\Delta^2}$$

3-6. táblázat: Szükséges mintaelemszámok, 95,5%-os megbízhatósági szint és különböző hibahatárok mellett a fordított piramis eloszlású sokaságok esetén

Δ	Válaszlehetőségek száma (k)				
	5	6	7	8	9
0,005	391 111	546 667	736 000	946 667	1 188 571
0,010	97 778	136 667	184 000	236 667	297 143
0,050	3 911	5 467	7 360	9 467	11 886
0,100	978	1 367	1 840	2 367	2 971

A fenti három különböző lépcsős eloszlástípus súlyrendszere felírható a következő általános képlet 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 adja vissza. Különböző a értékek esetén eltérő lesz az eloszlások „lapultsága". Ennek megfelelően a különböző értékei segítségével is kifejezhető lenne a szórás. Mivel a lekérdezés tervezésekor még nem állnak rendelkezésünkre ezen információk, a mintaelemszám tervezés esetére megelégszem a fentiekben részletesebben bemutatott esetek tárgyalásával. Úgy gondolom, hogy azok jó támpontot nyújthatnak a mintatervezés folyamán. Ráadásul az a paraméter nem minden értéke esetén értelmezhető a fenti eloszlás, hisz némely értékek esetén negatív relatív valószínűségeket eredményez.

Normalitáson alapuló eloszlások

A társadalmi, gazdasági élet sok jelenségét írja le pontosan, vagy legalább közelítően a normális eloszlás, amely alapgondolata, hogy a jellemző, átlagos tulajdonság gyakran, az extrémumok pedig ritkábban fordulnak elő. Emiatt, valamint a némileg eltérő szórás és mintaelemszámok miatt tárgyalom a következő eloszlásokat[5]:

[5] Nem tárgyaljuk ismét az egyenletes eloszlást, hiszen ennek elemzése a lépcsős eloszlásokkal foglalkozó alpontban megtörtént, ám – könnyen beláthatóan – az egyenletes eloszlás éppen úgy levezethető lenne a normálison alapuló eloszlás-családból is!

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

Mivel ezen csoport összes tárgyalt alesete a „kvázi normális” eloszláson alapul, ez utóbbi eloszlás-típushoz némi magyarázat tartozik. A tömegjelenségek esetén sokszor feltételezhető, és a mintavétel megrendelői körében is viszonylagos ismertségnek örvendő, normális eloszlás – mint tudjuk – folytonos. A továbbiakban „kvázi-normálisnak” nevezem 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 az alábbi elven képzem:

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

ahol $\Phi(x)$ a standard normális eloszlás eloszlásfüggvény-értéke az x helyen; és $[-z;z]$ az az intervallum, ahol a standard normális eloszlást értelmezzük[6]. Látható, hogy az bevezetésben említett alternatív ismérv esetén, ez a

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

értéket jelenti.

[6] 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 egyet adjon – korrigálnunk kell! A továbbiakban mindvégig a $(-3;3)$ intervallummal számolunk, ekkor a nevezőben szereplő korrekciós faktor $1 - 2\times\Phi(-3) \approx 0{,}9973$.

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

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

A lépcsős eloszlásokhoz hasonlóan fenti képlet biztosítja, hogy

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

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

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

Az extrém egymóduszú eloszlás-típus is igényel némi kifejtést. Ez a típus sem definiálható úgy, hogy csak egy eloszlás legyen hozzárendelhető; ám törekszem arra, hogy olyan eloszlásokat határozzak meg, melyek a korábban ismertetettek alapján általánosíthatók. Egymóduszú eloszlásokat a következő elven képezem: származzanak az egyes kimenetelekhez tartozó valószínűségek az alábbi formulából:

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

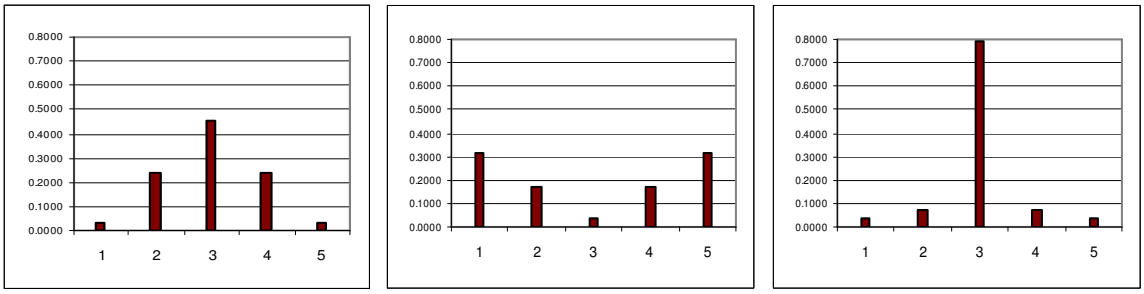
A fenti eljárás alapján keletkező valószínűségekre igaz, hogy:

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

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

és – könnyen beláthatóan – a kvázi normális eloszlásnál csúcsosabb empirikus sűrűségfüggvény keletkezik. A lépcsős eloszlások esetén a hasonló elven képezhető eloszlás megegyezik a piramis eloszlással, így ott ezt az eloszlást nem emeltem ki külön.

A normálison alapuló eloszlások sematikus képe az alábbi ábrán látható:



3-4. ábra: Normálison alapuló eloszlások $k = 5$ esetén
(kvázi-normális, fordított normális, extrém egymóduszú)

A korábban definiált kvázi-normális eloszlás esetén a mintaelemek varianciája az alábbi módon írható fel (kihasználva az átlagos eltérés-négyzetösszegről korábban írottakat):

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

A hibahatárból kifejezhető a szükséges mintanagyság:

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

Ismét képezhető a már ismert táblázat:

3-7. táblázat: Szükséges mintaelemszámok, 95,5%-os megbízhatósági szint és különböző hibahatárok mellett „kvázi normális” eloszlású válaszadás feltételezésével

Δ	Válaszlehetőségek száma (k)				
	5	6	7	8	9
0,005	120 853	168 414	224 636	289 516	363 049
0,010	30 213	42 104	56 159	72 379	90 762
0,050	1 209	1 684	2 246	2 895	3 630
0,100	302	421	562	724	908

A „kvázi normális” eloszlás feltételezése mellett a mintanagyság függ a lehetséges kimenetelek számától, a válaszlehetőségek számával párhuzamosan növekszik.

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

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

illetve

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

A variancia alapján megállapítható a szükséges mintaelemszám (mivel a gondolatmenet azonos a korábbiakkal csak a „végeredménynek” számító mintanagyságokat közlöm):

3-8. táblázat: Szükséges mintaelemszámok, 95,5%-os megbízhatósági szint és különböző hibahatárok mellett, fordított normális sokaságok esetén

Δ	Válaszlehetőségek száma (k)				
	5	6	7	8	9
0,005	452 765	615 793	806 145	1 023 495	1 267 700
0,010	113 191	153 948	201 536	255 874	316 925
0,050	4 528	6 158	8 061	10 235	12 677
0,100	1 132	1 539	2 015	2 559	3 169

A korábban definiált extrém egymóduszú (EEM) esetén a variancia alapján a szükséges mintaelemszám az alábbi képletekkel határozható meg (lásd Függelék a páratlan esetre):

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

$$n^{EEM**} = \frac{\frac{\varphi_1^{(k)}k(k-2)(k+2)(k+4)}{24}+1}{\Delta^2}$$

Ebből felírható a szükséges mintanagyságok táblázata.

3-9. táblázat: Szükséges mintaelemszámok, 95,5%-os megbízhatósági szint és különböző hibahatárok mellett, extrém egymóduszú sokaságok esetén

Δ	Válaszlehetőségek száma (k)				
	5	6	7	8	9
0,005	66 574	108 666	94 413	144 679	135 812
0,010	16 643	27 167	23 603	36 170	33 953
0,050	666	1 087	944	1 447	1 358
0,100	166	272	236	362	340

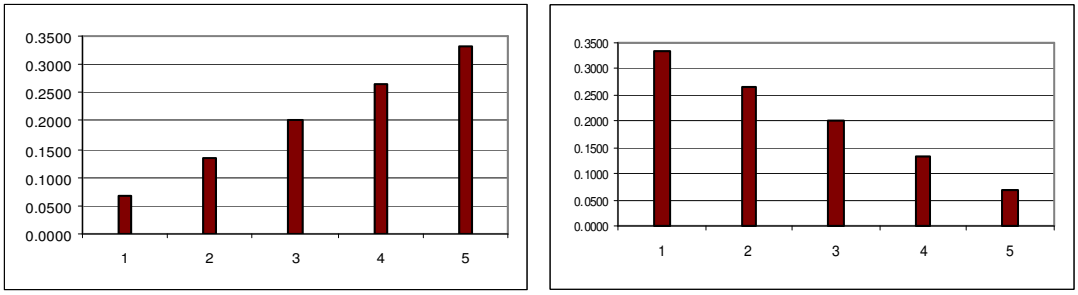
Fel kell hívni a figyelmet arra a tényre, miszerint a normálison alapuló eloszlások esetén nem tudjuk a mintanagyságot csupán a válaszlehetőségek száma, valamint a hibahatár alapján kifejezni, ezen esetekben szükséges a standard normális eloszlás bizonyos kvantiliseinek ismerete is. Ezek azonban ma már könnyen meghatározhatóak, akár valamely kézikönyv táblázatainak, vagy statisztikai programcsomag segítségével.

Az extrém egymóduszú eloszlás érdekes jelenségre hívja fel figyelmünket. Eddig minden esetben növekvő válaszlehetőségekkel egyben növekedett a szükséges mintaelemszám is. Ennél az eloszlásnál (ami a jelentősen eltérő variancia, illetve szükséges mintaelemszám képletből is látható) igazán érződik, hogy a páros számú kimenetellel rendelkező eloszlások rákényszerítik a válaszadót az állásfoglalásra, ami – szimmetrikus esetben – növeli a válaszok heterogenitását. Míg az extrém egymóduszú, páratlan esetben a semleges eset „lenyeli” a varianciát, addig páros esetben mindenképp magasabb varianciára számíthatunk.

3.3.2. Aszimmetrikus eloszlású válaszadások

Természetesen egy, a gyakorlatban végrehajtandó mintavétel esetén nem garantálható, hogy a válaszadók véleménye a semleges megfontolásra szimmetrikusan alakuljon ki (sőt, sok esetben a megrendelő a nem semleges eredményt (pl. pozitív attitűd a termék iránt) várja). Éppen ezért célszerű megvizsgálni az aszimmetrikus vélemények esetén kialakuló eloszlások esetét is. Az alábbiakban – a korábbinál nem kevésbé vitathatóan egyszerűsített – két esetet vizsgálok meg:

- egyenletesen növekvő valószínűséggel adott válaszok esete; valamint
- egyenletesen csökkenő eloszlások esete.



3-5. ábra: Aszimmetrikus eloszlások $k = 5$ esetén
(egyenletesen növekvő, egyenletesen csökkenő)

Elsőként vizsgáljuk meg azt az esetet, melyben a Likert-skála válaszlehetőségeinek előfordulási gyakorisága a teljes elutasítástól a teljes azonosulásig egyenletesen növekszik[7]. Ekkor az egyes osztályzatokra adott válaszok előfordulásának relatív gyakorisága:

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

Mivel a súlyok összege egy, ezért

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

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

$$\overline{x} = 1\times\frac{2}{k(k+1)}\times 1 + 2\times\frac{2}{k(k+1)}\times 2 + \ldots + k\times\frac{2}{k(k+1)}\times k = \frac{2k+1}{3}$$

A variancia ebből a következőképpen adódik:

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

Amiből a szokásos módon

[7] Az esetre a továbbiakban, mint aszimmetrikus egyenletesen növekvő eloszlásra, az AEN kóddal hivatkozok.

$$n^{AEN} = \frac{\frac{2}{9}(k-1)(k+2)}{\Delta^2}$$

A szükséges mintaelemszámok:

3-10. táblázat: Szükséges mintaelemszámok, 95,5%-os megbízhatósági szint és különböző hibahatárok mellett, egyenletesen növekvő valószínűségű válaszok esetén

Δ	Válaszlehetőségek száma (k)				
	5	6	7	8	9
0,005	248 889	355 556	480 000	622 222	782 222
0,010	62 222	88 889	120 000	155 566	195 556
0,050	2 489	3 556	4 800	6 222	7 822
0,100	622	889	1 200	1 556	1 956

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

$$\overline{x} = k \times \frac{2}{k(k+1)} \times 1 + (k-1) \times \frac{2}{k(k+1)} \times 2 + \ldots + 1 \times \frac{2}{k(k+1)} \times k = \frac{k+2}{3}$$

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

Az aszimmetrikus eloszlások vizsgálata természetesen távolról sem teljeskörű, azonban az egyenletes esetekek kívül más eloszlások definiálására nem vállalkozok, aminek két fő oka van:

- Ne felejtsük el, hogy a fenti eloszlásokban előzetes elképzelést fogalmazunk meg, mintatervezés esetén a megrendelőnek ezek közül a „minták” közül kell tudnia azt kiválasztani, amelyik – véleménye szerint – a leginkább tükrözi a valóságot. Nem túl életszerű annak a feltételezése, hogy a lekérdezés előtt

olyan mélységben ismerjük a sokaságot, amilyen mélységekig bonyolultabb aszimmetrikus eloszlások vezetnének.

- Általános aszimmetrikus eloszlás definiálása különböző k tagszámok esetén nem triviális. Ötfokozatú Likert-skálánál a 2-es, illetve 4-es lehetőségekre adott válaszok móduszként való megjelenítése még érthető, de 7, vagy 9 lehetséges kimenetel esetén a lehetséges eloszlások számossága túl nagy, nem is beszélve a páros kimenetelű esetekről.

3.4. A szükséges mintaelemszámok összehasonlítása

A folytatásban a korábban említett eseteket kísérelem meg összevetni, ezáltal néhány gyakorlati tanácsot kívánok adni a Likert-skálás lekérdezések alkalmazóinak. „Etalonként" a bevett gyakorlat szerint az alternatív ismérven alapuló mintanagyság meghatározást alkalmazom, ezzel összevetve 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.

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%-a a „megcélzott" hibahatár, akkor – különböző fokszámú Likert-skálák esetén – felírhatjuk Δ általunk elvárt értékét az alábbi módon:

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

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

$$\Delta^{(k)} = \frac{\Delta(k-1)}{100} \tag{3.7}$$

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

3-11. táblázat: Hibahatárok összehasonítása különböző terjedelmű Likert-skálák esetén

Δ (százalékpont)	$\Delta^{(5)}$	$\Delta^{(6)}$	$\Delta^{(7)}$	$\Delta^{(8)}$	$\Delta^{(9)}$
0,5	0,02	0,025	0,03	0,035	0,04
1,0	0,04	0,05	0,06	0,07	0,08
2,5	0,1	0,125	0,15	0,175	0,2
5,0	0,2	0,25	0,3	0,35	0,4

A fenti (3.7) 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 mutatom be; a nagyobb terjedelmű skálákhoz tarozó eredmények könnyedén számíthatók, és a bemutatotthoz hasonlóan értelmezhetők.

3-12. táblázat: Szükséges mintaelemszámok ötfokozatú Likert-skála, relatív hibahatár és különféle eloszlás-típusok esetén, $1-\alpha = 0,955$

$\Delta^{(5)}$	**Extrém két-móduszú**	**Fordított normális**	**Fordított piramis**	**Egyenletes**	**Piramis**	**Kvázi normális**	**Extrém egy-móduszú**
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 3-12. táblázat alapján jól látható, hogy az extrém kétmóduszú esetben a szükséges mintaelemszám megegyezik a bináris esetben számítottal (lásd 3-1. táblázat). Mindez nem meglepő, 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 megfigyelhetjük a szükséges mintaelemszámokban mutatkozó különbségek is. Bármilyen esetben a szükséges mintaelemszám a bináris esetnél szigorúan kisebb, hisz a szórás maximumát éppen abban az esetben veszi fel. A két extrém

eloszlás esetén a különbség közel tízszeres(!) a szükséges mintaelemszám tekintetében, de normális eloszlást feltételezve is csak mintegy ötöde a maximálisnak a szükséges elemszám. A fejezet legfontosabb eredményeit tulajdonképpen a 3-12. táblázat értelmezésével nyerhetjük. 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 ésszerű 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 minta tervezőjének. Az eddig is részletesebben bemutatott ötfokozatú skálán a különböző hibahatárokhoz tartozó elemszámok a következő, 3-13. táblázatban szerepelnek. Az összehasonlíthatóság érdekében feltüntettem az egyenletes, valamint a kvázi normális eloszláshoz szükséges elemszámokat is.

3-13. táblázat: Az aszimmetrikus, illetve néhány szimmetrikus eloszlás esetén szükséges mintaelemszámok, $1-\alpha = 0,955$

$\Delta^{(5)}$	**Egyenletes**	**Egyenletesen növ.**	**Kvázi normális**
0,02	20 000	15 556	7 553
0,04	5 000	3 889	1 888
0,10	800	622	302
0,20	200	156	76

Természetesen tisztában vagyok 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 és tetszőleges eloszlások szimulációi azt mutatják, hogy a valamilyen szempontból extrém eloszlásokon kívül a legtöbb aszimmetrikus eloszlás ebbe az intervallumba esik, amely megállapítás jó támpontot nyújt a szükséges mintanagyság megtervezéséhez általános esetben.

3.5. A variancia és a mintaelemszám érzékenysége

Ahogy azt már említettem, az előző fejezetekben bemutatott eloszlástípusok csak irányadóak. Érdemes megvizsgálnunk, hogy a valós sokasági eloszlás (akár minimális) eltérése a bemutatott idealizált eloszlásoktól mennyiben befolyásolja eredményeinket. Az alfejezetben prezentált eredményeim úgy is interpretálhatók, hogy ha a tényleges eloszlás helyett attól kismértékben eltérő, idealizált (az előző fejezetben részletesen bemutatott) eloszlás feltételezésével számítjuk ki a szükséges varianciát, majd a szükséges mintaelemszámot, mennyit és milyen irányban tévedhetünk. Gyakori szóhasználattal élve a mintanagyság meghatározás fentiekben bemutatott módszerének érzékenységvizsgálatát végzem el.

Tekintsük a sokasági varianciát egy k-változós függvényként, melynek változói[8] $p_1, p_2, \ldots, p_k$. Tegyük fel, hogy rendelkezünk közelítő értékekkel ezekre a változókra nézve, melyeket jelöljünk $p_1^*, p_2^*, \ldots, p_k^*$ módon. Ekkor $f(p_1, p_2, \ldots, p_k) = f(\mathbf{p})$ kifejezhető, vagy közelíthető az $f(p_1^*, p_2^*, \ldots, p_k^*) = f(\mathbf{p}^*)$ függvényértékekkel.

A variancia átlagfelbontása alapján:

$$\sigma^2(\mathbf{p}) = \sum_{l=1}^{k} l^2 p_l - \left(\sum_{l=1}^{k} l p_l \right)^2 \tag{3.8}$$

Amennyiben nem minden l-re igaz, hogy $p_l^* = p_l$, úgy nagy valószínűséggel a varianciák sem egyeznek meg. Jelöljük és értelmezzük a varianciában elkövetetett hibát, illetve eltérést

$$\Delta_{\sigma^2} = \sigma^2(\mathbf{p}) - \sigma^2(\mathbf{p}^*) \tag{3.9}$$

módon, ahol $\sigma^2(\mathbf{p})$ az ismeretlen variancia, míg $\sigma^2(\mathbf{p}^*)$ a közelítő értékekkel számolt variancia.

---

[8] Zavaró lehet, hogy az eddig kimenetelek arányaként említett értékeket változóként említem a jelen alpontban, de a variancia függvény változóinak tekindendőek.

A másodrendű Taylor közelítés segítségével a variancia megváltozását alkalmasabb formára hozhatjuk. Mivel a kvadratikus függvények esetén a másodrendű Taylor közelítés azonosság, ezért esetünkben tetszőleges **p** vektorra teljesül a következő:

$$\Delta_{\sigma^2} = \nabla\sigma^2\left(\mathbf{p}^*\right)\left(\mathbf{p}-\mathbf{p}^*\right)+\frac{1}{2}\left(\mathbf{p}-\mathbf{p}^*\right)^T\nabla^2\sigma^2\left(\mathbf{p}^*\right)\left(\mathbf{p}-\mathbf{p}^*\right) \tag{3.10}$$

ahol ∇ a gradiens, vagy a parciális deriváltak vektora. Vezessük be a $\varepsilon_l = p_l - p_l^*$ és $\overline{x}^* = \sum_{l=1}^{k} l p_l^*$ jelöléseket. Egyszerű számolással kapjuk, hogy

$$\nabla\sigma^2\left(\mathbf{p}^*\right)\left(\mathbf{p}-\mathbf{p}^*\right) = \sum_{l=1}^{k}(l^2 - 2\overline{x}^* l)\varepsilon_l$$

és

$$\left(\mathbf{p}-\mathbf{p}^*\right)^T\nabla^2\sigma^2\left(\mathbf{p}^*\right)\left(\mathbf{p}-\mathbf{p}^*\right) = \sum_{l=1}^{k}\sum_{j=1}^{k}(-2lj)\varepsilon_l\varepsilon_j = -2\left[\sum_{l=1}^{k} l\varepsilon_l\right]^2$$

Fenti összefüggések figyelembe vételével (3.10) a következő alakban írható:

$$\Delta_{\sigma^2} = \sum_{l=1}^{k}\left(l^2 - 2l\overline{x}^*\right)\varepsilon_l - \left[\sum_{l=1}^{k} l\varepsilon_l\right]^2 \tag{3.11}$$

Ezen összefüggés alapján azt mondhatjuk, hogy a variancia közelítésének hibája kifejezhető a közelítő eloszlás átlagának, és a közelítő és tényleges súlyok (ismeretlen) különbségének segítségével. A következőkben olyan eseteket vizsgálok meg, amikor az ismeretlen $\varepsilon_l = p_l - p_l^*$ értékek valamilyen mintázatba rendezhetők, illetve értékeikről valamilyen feltételezéssel élünk. Célom a varianciában elkövetett hiba nagyságának vizsgálata.

Tegyük fel, hogy az m-edik kategória arányában tévedünk, azaz $p_m \neq p_m^*$. Könynyű belátni, hogy – a változók lineáris függősége miatt (összegük egy) – ekkor legalább egy másik változóban is hibát kell vétenünk. Arra vonatkozóan, hogy az m-edik változóban elkövetett hibánk hol „csapódik le” többféle feltételezéssel is élhetünk. Két esetet vizsgálok meg részletesen. Az első esetben azt tesszük fel, hogy egy kategóriában történő hibánk közvetlenül egy másikban csapódik le (ún. közvetlen eset). Ennek fontos speciális esete az egymás melletti kategóriák közötti tévedés. A másik szituációban azzal a

feltételezéssel fogok élni, hogy egy kategóriában elkövetett hiba az összes többi kategóriát egyenletesen érinti (ún. egyenletesen szétterülő hiba). Természetesen egyéb feltevések is elképzelhetők, amik a bemutatottakhoz hasonlóan vezethetők le a (3.11) általános képletből.

Elsőként megvizsgálom, hogy milyen hatása van annak, ha az m-edik kategória valószínűségében hibázunk az n-edik kategória javára, illetve kárára, ahol a hiba mértéke ε. A hiba előjelére nem teszek kikötést. A tény és közelítő valószínűségek és a hiba vektorait mutatja be a 3-14. táblázat.

3-14. táblázat: Tény és közelítő valószínűségek és a hiba a közvetlen esetben

Kimenetel	1	…	m	…	n	…	k
p (tény)	p_1	…	p_m	…	p_n	…	p_k
$\mathbf{p}^*$ (közelítés)	p_1	…	$p_m+\varepsilon$	…	$p_n-\varepsilon$	…	p_k
ε (hiba)	0	…	$-\varepsilon$	…	ε	…	0

A varianciában történő módosulás felírható (3.11) segítségével. Vegyük észre, hogy m és n nagyságrendi sorrendjétől függetlenül értelmezhető a felírás.

$$\Delta_{\sigma^2}=\sum_{l=1}^{k}\left(l^2-2l\overline{x}^*\right)\varepsilon_l-\left[\sum_{l=1}^{k}l\varepsilon_l\right]^2=\left(m^2-2m\overline{x}^*\right)(-\varepsilon)+\left(n^2-2n\overline{x}^*\right)\varepsilon-\left[m(-\varepsilon)+n\varepsilon\right]^2$$

A zárójeleket felbontva, a tagokat rendezve, kiemeléseket végezve kapjuk az alábbi, ε-ban másodfokú összefüggést:

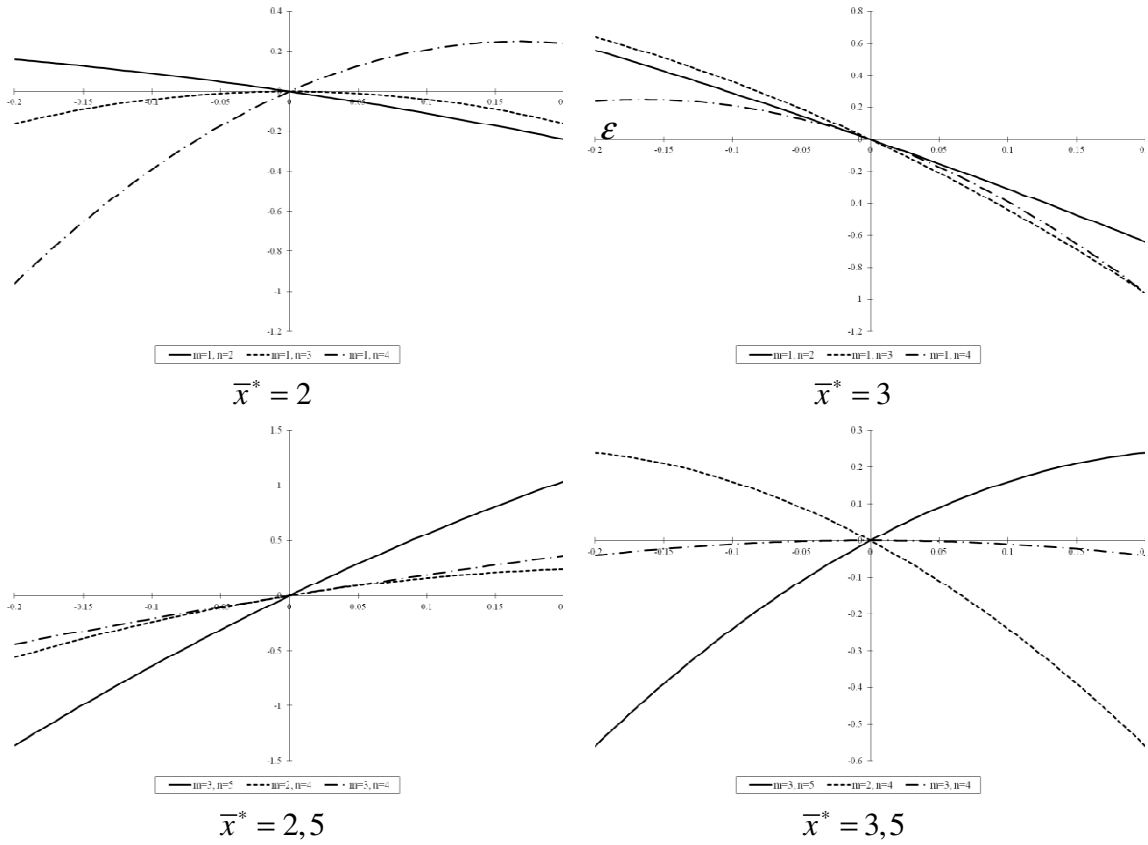
$$\Delta_{\sigma^2}^{\text{közvetlen}}(\varepsilon)=-(n-m)^2\varepsilon^2+\left[\left(n-\overline{x}^*\right)^2-\left(m-\overline{x}^*\right)^2\right]\varepsilon \qquad (3.12)$$

Ennek a függvénynek a képe (az $m\neq n$ esetet feltételezve) konkáv parabola az $\varepsilon_1=0$ triviális és az $\varepsilon_2=\dfrac{\left(n-\overline{x}^*\right)^2-\left(m-\overline{x}^*\right)^2}{(n-m)^2}$ zérushelyekkel.

A (3.12) összefüggés alapján tehát a hibás kategóriák egymástól és az átlagtól való távolsága és a hiba nagysága együttesen határozzák meg a varianciában jelentkező hibát. Mivel Δ_{σ^2}-t a tény variancia és a közelítő variancia különbségeként definiáltuk, pozitív értéke azt jelenti, hogy a tény variancia nagyobb, mint a közelítő, azaz a képlet alapján megállapított mintaelemszám a szükségesnél kisebb lesz.

A paraméterektől függően a variancia hibája alapvetően három formát ölthet, amennyiben azt ε függvényeként tekintjük.

Amennyiben $\left|n-\overline{x}^*\right|>\left|m-\overline{x}^*\right|$, akkor $\Delta_{\sigma^2}(\varepsilon)$ a $0<\varepsilon<\dfrac{\left(n-\overline{x}^*\right)^2-\left(m-\overline{x}^*\right)^2}{\left(n-m\right)^2}$ tartományban pozitív. Amennyiben $\left|n-\overline{x}^*\right|<\left|m-\overline{x}^*\right|$, akkor $\Delta_{\sigma^2}(\varepsilon)$ a $\dfrac{\left(n-\overline{x}^*\right)^2-\left(m-\overline{x}^*\right)^2}{\left(n-m\right)^2}<\varepsilon<0$ tartományban pozitív. Amennyiben $\left|n-\overline{x}^*\right|=\left|m-\overline{x}^*\right|$, akkor $\Delta_{\sigma^2}(\varepsilon)$ minden $\varepsilon\neq 0$ esetén negatív.



$\overline{x}^*=2$ $\overline{x}^*=3$

$\overline{x}^*=2,5$ $\overline{x}^*=3,5$

3-6. ábra: A variancia hibája ε függvényében

Egy fontos speciális eset annak vizsgálata, ha egymás melletti kategóriák esetén hibázunk, annak milyen következményei lehetnek a varianciára, és ezen keresztül a szükséges mintaelemszámra. Legyen $m=n-1$, ekkor (3.12) helyett

$$\Delta_{\sigma^2}^{\text{szomszédos}}=-\varepsilon^2-\left[2\overline{x}^*-2n+1\right]\varepsilon \tag{3.13}$$

írható. Ekkor, ha például $k=5$, a tényleges eloszlás valószínűségei rendre 0,2; 0,19; 0,21; 0,2; 0,2, de egyenletes eloszlást feltételezünk, úgy $m=2$, hiszen a második kategóriában magasabb a feltételezett valószínűség a ténylegesnél, $n=3$, hisz ezen kategória terhére tévedtünk. Az egyenletesnek feltételezett eloszlás esetén $\overline{x}^*=3$, így (3.13) felhasználásával:

$$\varepsilon=0,01,\ \overline{x}^*=3,\ n=3\Rightarrow\Delta_{\sigma^2}=-0,0101$$

ami azt mutatja, hogy a közelítés varianciája nagyobb, mint az eredeti eloszlásé, így a hibás feltételezés magasabb elemszámú minta vételét okozza ebben az esetben.

Másodikként vizsgáljuk meg azt az esetet, hogy valamelyik kategória esetén tévedünk, a tévedés pedig egyenletesen oszlik el az összes többi kategória között. Ezt a feltételezést mutatja be a 3-15. táblázat.

3-15. táblázat: Tény és közelítő valószínűségek és a hiba az egyenletesen szétterülő esetben

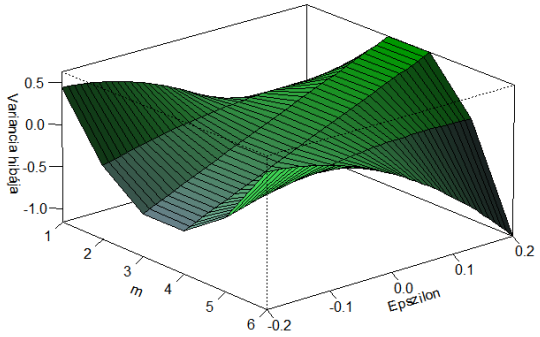
Kimenetel	1	…	m	…	k
p (közelítő)	p_1	…	p_m	…	p_k
$\mathbf{p}^*$ (tény)	$p_1-\frac{\varepsilon}{k-1}$	…	$p_m+\varepsilon$	…	$p_k-\frac{\varepsilon}{k-1}$
ε (hiba)	$\frac{\varepsilon}{k-1}$	…	$-\varepsilon$	…	$\frac{\varepsilon}{k-1}$

A formula levezetése matematikailag valamivel összetettebb, de hasonlóan elvégezhető, mint abban az esetben, amikor azt feltételezzük, hogy csak két kategória esetén van eltérés. Egyszerűsítések és összevonások elvégzése után ε-ban másodfokú alakban írhatjuk a varianciák eltérését:

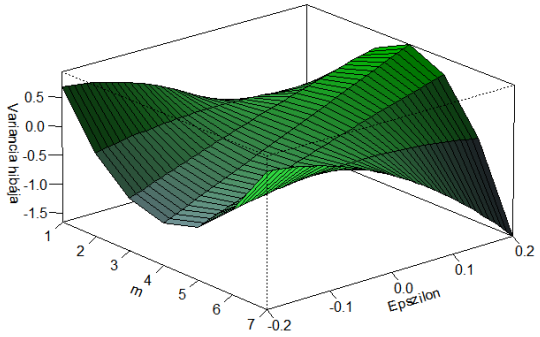
$$\Delta_{\sigma^2}^{\text{egyenletes}}=\frac{k(k+1)\left(2k-6\overline{x}^*+1\right)-6k\left(m^2-2m\overline{x}^*\right)}{6(k-1)}\varepsilon-\frac{(k-2m+1)^2k^2}{4(k-1)^2}\varepsilon^2 \qquad (3.14)$$

Bár a függvény ránézésre összetettnek tűnik, az eredmények bemutatása némileg egyszerűbb. Amennyiben ugyanis tudjuk a skála fokszámát, valamint a feltételezett eloszlás átlagát, úgy a variancia hibája már csak m-től és ε-tól függ, így ábrázolni tudjuk

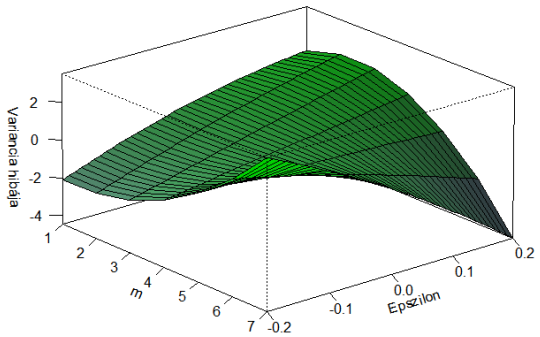
egy felületen[9]. A 3-7. ábra a variancia hibáját mutatja be tehát ε és m függvényében négy különböző esetben. A függvények természetesen ε folytonosak csak, de a jobb összehasonlíthatóság érdekében a teljes felületet ábrázoltam.



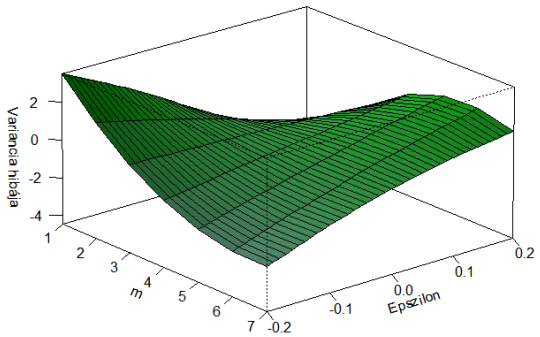
$k = 6, \overline{x} = 3,5$



$k = 7, \overline{x} = 4$



$k = 7, \overline{x} = 2$



$k = 7, \overline{x} = 6$

3-7. ábra: A variancia hibája ε és m függvényében az egyenletesen szétterülő esetben

[9] Az R függvényt lásd a Függelékben: varfelulet(k, x), az ismert paraméterek mellett megrajzolja a variancia függvényét.

A felső két ábrán egy páros és egy páratlan fokszámú skálára mutatom be a grafikonokat, ahol $-0,2 \le \varepsilon \le 0,2$, m pedig végigfut a teljes skálán, valamint az átlag a skála közepén helyezkedik el. A két grafikon szimmetrikus, adott nagyságú hiba hatása pedig annál erősebb, minél extrémebb értéknél fordul elő. Azokban az esetekben, amikor az átlag nem középütt helyezkedik el, az átlagtól távolabb eső kategóriák mutatnak nagyobb érzékenységet.

Ebben az alpontban azt vizsgáltam meg, hogy a mintaelemszám tervezés esetén mennyire érzékeny a kiinduló feltevés tényleges sokasági eloszlástól való különbségére. Az általános képlet meghatározása után három speciális esetet elemeztem részletesebben. Összességében azt tapasztaltam, hogy amennyiben a feltételezett eloszlásban rejlő hiba nem túlságosan nagy, úgy a varianciában megjelenő hiba sem nagymértékű. Az érzékenységvizsgálat egyben feltárta azokat a tényezőket, melyek a hiba nagyságára hatással vannak.

3.6. A szükséges mintaelemszám várható értéke

Az előző alfejezetekben feltételeztem, hogy előzőleges információkkal rendelkezünk a sokasági eloszlásról. Amennyiben nem rendelkezünk ilyen jellegű információval, akkor felvetődhet a szükséges mintaelemszám várható értékének meghatározása. Ehhez először azonosítanunk kell az összes lehetséges kimenetelt[10], majd a bekövetkezésük valószínűségét kell rögzítenünk. Az összes lehetséges kimenetel vizsgálatával az a célom, hogy megadjam a kimenetek varianciájának eloszlását (elsősorban várható értékét), és ezzel együtt a szükséges mintaelemszám várható értékét.

A következőkben elsőként az ötfokozatú Likert-skálán mutatom be a gondolatmenetemet. Ezután levezetem a szükséges mintaelemszám várható értékének általános képletét.

[10] Kimenetel vagy kimenet alatt az alfejezetben minden esetben egy, a lehetséges válaszok közötti gyakoriság megoszlást értek, azaz egy kimenetel általánosságban azt jelenti, hogy n_1 darab 1-es, n_2 darab kettes 2-es választ kaptunk, végül n_k darab válasz érkezett a k lehetőségre.

3.6.1. Az ötfokozatú Likert-skála esete

Tekintsünk egy ötfokozatú Likert-skálát. Az ötfokozatú Likert-skála esetén a mintaelemszám várható értékének meghatározásához a következő lépésekre van szükség:

- a lehetséges kimenetelek azonosítása, a kimenetelek számosságának meghatározása;
- a kimenetelek varianciájának kiszámítása;
- a meghatározott varianciák várható értékének számszerűsítése;
- a szükséges mintaelemszám várható értékének meghatározása.

Az alábbi alfejezetekben a fenti lépéseket veszem sorra.

A lehetséges kimenetelek azonosítása

Elsőként meg kell határoznunk az összes lehetséges kimenetel számosságát, ám ehhez szükségünk van arra az információra, hogy hány válaszadó tölti ki a kérdőívet. Mindez furcsának tűnhet, hisz végső célunk épp annak a meghatározása, hogy hány válaszadóra van szükség. Az alábbi képletben (valamint az egész alfejezetben) n-et a következőképp értelmezhetjük: az egyes lehetőségekre eső gyakoriságok megadásának pontossága. Amennyiben például $n = 100$, úgy egy százalékpontos pontossággal „állíthatjuk be" az egyes válaszadási lehetőségekre eső válaszok arányát. Amennyiben $n = 200$, úgy fél százalékpontos ez a lehetséges pontosság, stb.

Ezek után felírható az összes lehetséges kimenetel száma:

$$\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$$

A fenti kifejezések az összes ismétléses kombináció számát adják meg, bár formájukban némileg eltérnek a matematikában szokásos jelölésektől, hisz jelen esetben k elem n-ed osztályú ismétléses kombinációiról, illetve a kombinációk számosságáról beszélünk. A valószínűség-számításban használatos „n elem k-ad osztályú ismétléses kombinációi" kifejezés helyett alkalmazzuk ezt a némileg megtévesztő betűzést annak érdekében, hogy a statisztikában hagyományos jelöléseinket megtarthassuk.

Tegyük fel, hogy 1 százalékpontos pontossággal adhatjuk meg, hogy melyik lehetőséget milyen arányban választják $(n=100)$ a válaszadók. Ekkor összességében $\binom{n+k-1}{k-1}=\binom{104}{4}=4\,598\,126$ különféle kimenetel képzelhető el a kitöltetés végeredményeképp. Szemléltesse a 3-8. ábra a különböző kimeneteleket:

1	**2**	**3**	**4**	**5**	
0	0	0	0	100	
0	0	0	1	99	
0	0	0	2	98	$\rightarrow \binom{104}{4} = 4\,598\,126$ db
⋮	⋮	⋮	⋮	⋮	
99	1	0	0	0	
100	0	0	0	0	

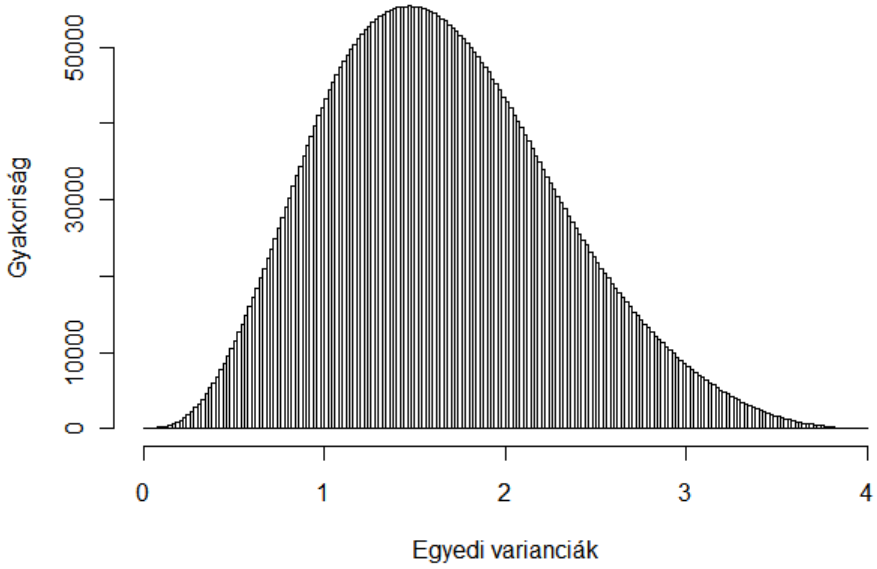
3-8. ábra: Az ötfokozatú Likert-skála kimeneteleinek illusztrációja

A fenti illusztrációban egy sor egy kimenetelt reprezentál. A lekérdezés végzőjének lényegtelen, hogy pl. a második sor esetén ki volt az a válaszadó, aki 4-es választ adott, csak az egyes lehetőségekre adott válaszok száma érdekes.

A kimenetelek varianciájának meghatározása

A különböző kimenetelek esetén a variancia meghatározása a szokásos módon történik. Az egyedi kimenetek varianciája 0 és $\left(\frac{k-1}{2}\right)^2$ között változhat, azaz esetünkben 0 és 4 között. Míg a maximális variancia egyetlen esetben, az ún. extrém kétmóduszú esetben fordul elő, addig a minimális öt (általánosságban k) kimenetelnél, ha minden egyes megkérdezett azonos választ jelöl.

Számítástechnikai eszközökkel a fenti, nagyszámú variancia könnyedén meghatározható. A varianciákat hisztogramon ábrázolva az előforduló varianciák eloszlását jobban megismerhetjük. A varianciák hisztogramja egy százados lépésközökkel ábrázolva a 3-9. ábrán látható.



3-9. ábra: A varianciák hisztogramja, ötfokozatú Likert-skála, $n = 100$

A hisztogram jól láthatóan enyhe jobboldali aszimmetriát mutat. Kérdésként merülhet fel, hogy az így definiált valószínűségi változót diszkrétnek, vagy folytonosnak kell-e tekintenünk. A továbbiakban a várható érték meghatározásánál az eloszlás diszkrét tulajdonságát használom fel, a várható értéket a varianciák és a valószínűségek szorzataként állapítom meg.

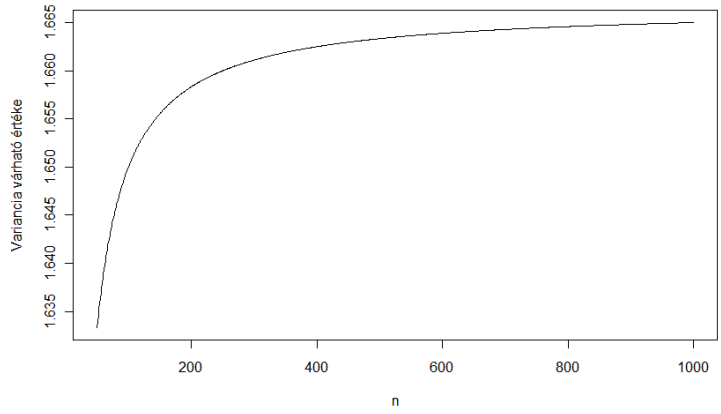
A varianciák várható értékének meghatározása

A varianciák várható értékének meghatározásához elengedhetetlen, hogy súlyokat, valószínűségeket rendeljünk az egyes kimenetelekhez. A kérdés gyakorlati szempontból úgy merül fel, hogy mekkora valószínűséggel kapjuk azt az eredményt n ember megkérdezése után, hogy pl. mindannyian 1-es választ adtak, vagy a válaszadások a lehetőségek között egyenletesen oszlottak meg. Azzal a feltételezéssel éltem, hogy minden egyes kimenetel azonos valószínűséggel rendelkezik. Mindez azt jelenti, hogy ugyanannyira tartom elképzelhetőnek a lekérdezés előtt azt, hogy minden válaszadó egyformán válaszol, mint azt, hogy egyenlő arányban születnek a válaszok, vagy bármilyen, tetszőleges válasz-struktúra (kimenetel) születik.

Számítógép segítségével, valamennyi eset generálása után könnyedén meghatározható a 4 598 126 variancia várható értéke – a feltételezések miatt – egyszerű számta-

ni átlagként. Az eredmény: 1,65, azaz $n=100$ esetén az ötfokozatú Likert-skálán a varianciák várható értéke 1,65.

Amennyiben továbbra is a $k=5$ esetre a számítást különböző n-ekre is elvégezzük, a varianciák várható értékére az $n=100$ esetétől eltérő eredményeket kapunk. Az eredményeket az alábbi ábrán mutatom be:



3-10. ábra: Az ötfokozatú Likert-skála esetén a variancia várható értéke, különböző *n*-ek esetén

A fenti, 3-10. ábra alapján jól látható, hogy az ötfokozatú Likert-skálák esetén n növelése a variancia várható értékét növeli. Az az érzésünk azonban, hogy a várható érték konkáv módon nő, a lehetséges pontosság végtelenbe való növelésével a variancia várható értéke nem a végtelenbe tart, hanem minden határon túl megközelít egy véges határértéket. A variancia várható értékének határértéke ötfokozatú Likert-skála esetén közelítőleg $5/3$, amit a későbbiekben általánosan bizonyítok is.

A 3-10. ábra természetesen nem csak az ötfokozatú skálára készíthető el, hanem bármely más Likert-skálára is, a grafikon által sugallt határérték ezekben az esetekben magától értetődően más lesz.

A szükséges mintaelemszám várható értékének meghatározása

A szükséges mintaelemszám ($\dot{n}$) várható értéke a variancia várható értékének ismeretében könnyedén meghatározható:

$$E(\dot{n}) = E\left(\frac{4\sigma^2}{\Delta^2}\right) = \frac{4E(\sigma^2)}{\Delta^2} = \frac{20}{3\Delta^2}$$

A fenti képlet alapján az ötfokozatú Likert-skála esetén szükséges mintaelemszámok várható értékét az alábbi táblázatban foglalhatjuk össze.

3-16. táblázat: Szükséges mintaelemszámok várható értéke ötfokozatú Likert-skála esetén, előre adott hibahatárok mellett

Δ	$E(\dot{n})$
0,005	266 667
0,010	66 667
0,050	2 667
0,100	667

Az ötfokozatú Likert-skála megismerése után határozzuk meg a szükséges mintaelemszám várható értékét általános esetben.

3.6.2. A szükséges mintaelemszám várható értékének általános meghatározása

A szükséges mintaelemszám várható értékének meghatározását a variancia várható értékének kiszámításával kezdem, majd a várható érték segítségével számítom ki a szükséges mintaelemszámot.

A variancia várható értékének meghatározása

Tegyük fel, hogy k fokozatú Likert-skálát vizsgálunk, azaz a lehetséges válaszlehetőségek: $1, 2, \ldots, k$. Általános esetben nem járhatunk el a konkrét esettel megegyezően, hisz nem tudjuk az összes lehetséges kimenetelt meghatározni, majd a varianciát számszerűsíteni.

A kimenetelek bekövetkezésének valószínűségére általános esetben ugyanazzal a feltételezéssel élek, azaz valamennyi kimenet azonos súlyt kap. Hosszadalmas levezetés (lásd Függelék) után kapjuk a meglepően egyszerű végeredményt:

$$E\left(\sigma_{n,k}^2\right) = \frac{k(k-1)(n-1)}{12n} \tag{3.15}$$

Ellenőrzésképpen helyettesítsünk $n = 100, k = 5$ értékeket a kapott általános képletbe: $E\left(\sigma_{100,5}^2\right) = \frac{5 \cdot 4 \cdot 99}{12 \cdot 100} = 1,65$, azaz a várt eredményt kaptuk.

A fenti összefüggés alapján igazolódott a sejtésünk a konvergenciára vonatkozóan. Amennyiben $n \to \infty$, $E\left(\sigma_{n,k}^2\right) \to \frac{k(k-1)}{12}$, azaz $E\left(\sigma_{n,5}^2\right) \to 5/3$. A 3-10. ábra: tehát valóban jól mutatta a konvergenciát, a határérték általános esetben $\frac{k(k-1)}{12}$, azaz $5/3$ az ötfokozatú Likert-skála esetén.

A várható érték felírható a korrigált varianciára is, eredményként $E\left(s_k^2\right) = \frac{k(k-1)}{12}$ adódik, ami független n értékétől.

Az alábbi táblázatban néhány tipikus k -ra érvényes várható értéket láthatunk:

3-17. táblázat: A korrigált variancia várható értékei néhány Likert-skála esetén

k	3	4	5	6	7	8	9	10
$E\left(s_k^2\right)$	0,5	1	$5/3$	2,5	3,5	$14/3$	6	7,5

A variancia várható értékének meghatározása után már csak szükséges mintaelemszám várható értékének kiszámítása van hátra.

A mintaelemszám várható értékének meghatározása

A variancia várható értékének meghatározása után a mintaelemszám várható értékének meghatározása nem okoz problémát. Az

$$E(\dot{n}) = \frac{4E\left(s^2\right)}{\Delta^2}$$

összefüggésbe a (korrigált) variancia várható értékét helyettesítve könnyedén számíthatóak a keresett értékek.

Az előző alfejezetben bemutatott típusonkénti táblázatokhoz hasonlóan megszerkeszthető az adott hibahatárokhoz tartozó, a szükséges mintaelemszámok várható értékét bemutató összefoglaló táblázat is. Az alábbiakban tehát az adott fokszámú Likert-skálához és adott hibahatárhoz tartozó szükséges várható mintaelemszámokat olvashatjuk.

3-18. táblázat: Szükséges mintaelemszámok várható értéke néhány Likert-skála esetén, előre adott hibahatárok mellett

Δ	$k=5$	$k=7$	$k=9$	$k=10$
0,005	266 667	560 000	960 000	1 200 000
0,010	66 667	140 000	240 000	300 000
0,050	2 667	5 600	9 600	12 000
0,100	667	1 400	2 400	3 000

Természetesen a mintaelemszám várható értékére is alkalmazható az ún. relatív hibahatár fogalma. A meghatározott általános képlet segítségével a 3-12. táblázat kiegészíthető a várható érték oszlopával.

3-19. táblázat: Szükséges mintaelemszám néhány előre definiált eloszlástípus esetén, valamint a szükséges mintaelemszám várható értéke ötfokozatú Likert-skálán

$\Delta^{(5)}$	Extrém két-móduszú	Fordított normális	Fordított piramis	Egyenletes	Várható érték	Piramis	Kvázi normális	Extrém egy-móduszú
0,02	40 000	28 298	24 444	20 000	**16 667**	13 333	7 553	4 161
0,04	10 000	7 074	6 111	5 000	**4167**	3 333	1 888	1 040
0,1	1 600	1 132	978	800	**667**	533	302	166
0,2	400	283	244	200	**167**	133	76	42

A 3-19. táblázat alapján elmondhatjuk, hogy a mintaelemszámok várható értéke a piramis és az egyenletes eloszlás esetén a szükséges elemszámok közé esik. Az empirikus tapasztalatok is azt mutatják, hogy ez az az intervallum, melybe a „reálisabb” kimenetelek többsége beleesik. Mindezeket figyelembe véve úgy gondolom, hogy tanácsolható a mintaelemszám várható érték alapján történő meghatározása gyakorlati szakemberek számára.

A Likert-skálás felmérések egyre elterjedtebbek az attitűd-vizsgálatokban, társadalomtudományi felmérések során. A minta nagyságának tervezése ezekben a vizsgálatokban is kulcsfontosságú, hiszen az eredmények pontossága, illetve megbízhatósága jelentős részben ettől függ.

A felvázolt eloszlás-típusok csak egy részét képezik a viszonylag jól meghatározható, könnyen számszerűsíthető varianciával kecsegtető eloszlásoknak.

A fejezet második felében olyan megoldást találtam, amelyben nincs szükség olyan kiegészítő információkra, melyek a lekérdezés előtt még nem állnak rendelkezésünkre. Előbb egy konkrét példán keresztül mutattam be a gondolatmenetemet, majd sikerült meghatároznom a szükséges mintaelemszám várható értékét adott fokszámú

Likert-skála és adott nagyságú hibahatár esetén. A meghatározáshoz felhasználtam azt a feltételezést, hogy a különböző kimenetelek azonos valószínűséggel fordulhatnak elő.

A fejezetben bemutattam, hogy amennyiben létezik előzetes feltevésünk a vizsgált jelenség (válaszok) eloszlása tekintetében, akár tizedrészére csökkenthető a szükséges mintanagyság. Természetesen felmerülhet a kérdés, hogy mi értelme a mintavételnek, ha ilyen pontosan ismerjük a leendő válaszadók véleményét! Megítélésem szerint a fejezet fejtegetései egy kétfázisú mintavételi eljárást sugallnak: először egy kisebb elemszámú (pl. a kvázi normális eloszlás feltételezésével meghatározott) minta alapján meghatározzuk a válaszok várható eloszlását; majd ennek ismeretében – szükség szerint – kiegészítjük a korábbi mintavételt. Az előzetes feltevések és a mintabeli adatok összesítésében a későbbi fejezetekben a bayesi statisztika eszköztárát kívánjuk segítségül hívni. Mivel a bayesi statisztika alapeleme az előzetes információk megléte, ezért az ilyen jellegű információk modellbe építését is megvizsgáljuk. A dolgozat következő két fejezete tehát az eddigi klasszikus megközelítés mellett a lehetséges bayesi gondolkodást is bemutatja. A 4. fejezetben alapfogalmakat és technikákat mutatok be, majd az 5. fejezetben ezeket a technikákat alkalmazom a dolgozatban tárgyalt probléma, a mintaelemszám tervezés témakörére.

4. Előzetes és mintabeli információk bayesi kombinálása

Equation Section (Next)A bayesi statisztika az utóbbi néhány évtizedben reneszánszát éli. Művelői szerint egy koherens módszertani gondolatvilágot alkot, kritizálói logikai tetszetősségét elismerik ugyan, de egyúttal gyenge pontjaira is felhívják a figyelmet. Mivel a hazai szakirodalomban néhány kivételtől eltekintve alig van jelen a bayesi gondolatkör, dolgozatomban bemutatom az alapvető gondolatokat, a módszerekhez közvetlenül kapcsolódó szimulációs technikákat, valamint érintőlegesen az egyik ezekhez szükséges lehetséges programcsomagot. Mindez szervesen kapcsolódik az előző fejezet záró gondolatához, az eloszlásról rendelkezésre álló előzetes információk és egy kis, tájékozódást segítő minta eredményeinek összesítéséről.

4.1. A bayesi statisztika rövid története és gondolatvilága

Thomas Bayes – akinek a neve fémjelzi jelen témakörünket – valószínűleg soha nem gondolta, hogy egy egész (nem csak statisztikai) gondolatkör fogja a nevét viselni. Életében mindössze két tanulmánya jelent meg, egy teológiai és egy matematikai munka. Az „An Essay Toward Solving a Problem in the Doctrine of Chances” című munkája (Bayes, 1763), melyben a binomiális eloszlás paraméterére vonatkozó következtetésekkel foglalkozott posztumusz jelent meg. További történeti érdekesség, hogy a bayesi statisztika elnevezést R. A. Fisher terjesztette el, aki egyébként messzemenőkig nem értett egyet a bayesi elvekkel, a jelzőt erősen negatív értelemben használta.

A bayesi statisztikusok szerint a bayesi gondolatvilág egy olyan keretrendszer, mely ötvözi a bizonytalansággal kapcsolatos személyes véleményt, leírja, hogy a racionális, haszonmaximalizáló egyénnek milyen döntéseket kell(ene) hoznia a folyamatosan változó külvilághoz alkalmazkodva. A szemlélet szépsége, hogy tulajdonképp minden megfontolás egy egyszerű gondolaton, a Bayes-tételen alapul, a többi „csupán” formális matematika. Ezzel el is érkeztünk a módszer egyik nagy sajátosságához, annak számításigényességéhez. A felhasznált numerikus módszerek ugyan ismertek voltak, legalábbis alapjaikban már korábban is, nagy lendületet azonban a számítógépek, valamint az egyszerűbb programozási nyelvek elterjedése adta a területnek a 80-as évek második felében. Ma a leggyakrabban alkalmazott, bayesi statisztikát is támogató szoftverek az

R, a MATLAB, valamint a BUGS (Bayesian inference Using Gibbs Sampling) különböző verziói (WinBUGS, OPENBugs, R-rel, MATLAB-bal, SAS-sal együttműködő modulok stb.).

Magyar nyelven csupán néhány szerző munkáját tudom megemlíteni. Theiss (1971) teljesen elméleti, bevezető jellegű munkájában elsősorban regressziós, makro modellezési alkalmazásokra koncentrál. Varga (1991) autoregresszív idősori modellek paraméterbecslésére használja az elméleti keretet. Wickman (1999) magyarra fordított műve még teljességgel nélkülözi a modern bayesi szimulációs módszereket. Hunyadi (2001) összefoglaló módszertani munkájában egy fejezetet szentel a bayesi módszerek ismertetésére, elsősorban regressziós példán keresztül, melyet a 80-as években tanulmánya és kandidátusi értekezése előzött meg. Várpalotai disszertációjában (Várpalotai, 2008) simasági priorokkal foglalkozik, két modellt mutat be részletesen, az egyik az üzleti ciklusok szinkronitásáról, a másik infláció előrejelzésről szól. A dolgozatban röviden bemutat szimulációs technikákat is, melyek alapvető jelentőségűek. Ezen kívül több munkájában használt bayesi módszereket is, melyek részint MNB Munkafüzetek, részint tanulmányok formájában jelentek meg. Horváth (2001) a hierarchikus Bayes-módszerekkel és a conjoint analízissel foglalkozott az ezredforduló tájékán egy cikk és jó néhány konferencia előadás keretében Magyarországon és külföldön egyaránt.

Bayesi kötődésű munka egyrészt Hunyadi (2011a) szellemes irodalomismertetése, valamint átfogó, a bayesi statisztika gondolatvilágát bemutató írása (2011b). Kovács és Balogh (2009) sertésár idősorokkal kapcsolatos elemzésben alkalmazza a bayesi módszertant. Legjobb tudomásom szerint bayesi alapokon nyugvó, diszkrét eloszlásokkal és modellekkel foglalkozó dolgozat még nem jelent meg magyar nyelven.

A bayesi statisztika növekvő népszerűségét a bayesi tanulmányok számának növekedése is mutatja a nemzetközi szakirodalomban. A módszerek elterjedése többek között annak köszönhető, hogy már rendelkezésre állnak standard kézikönyvek a szükséges ismeretek elsajátításához. Ezek közül szeretnék kiemelni néhányat, melyeket jelen fejezet elkészítése során is felhasználtam (Albert, 2009; Congdon, 2005; Gelman et al., 2004; Geweke, 2005; Koop, 2003; Koop et al., 2007).

A bayesi statisztika alapvető szemléletében eltér a klasszikus felfogástól, a különbség lényegi pontjait Hunyadi (2011b) foglalja táblázatba az alábbi módon:

4-1. táblázat: A klasszikus és bayesi statisztika jellemzői

Tulajdonság	Klasszikus statisztika	Bayesi statisztika
1. Paraméter	Rögzített	Valószínűségi változó
2. Valószínűség	Objektív	Szubjektív is lehet
3. Külső információ	Nincs vagy csak kevés	Van és lényeges
4. Minta	Valójában csak egy van, de feltételezzük az ismételt mintavétel lehetőségét	Csak egyetlen mintát értékelünk

Forrás: Hunyadi (2011b)

A négy, táblázatban felsorolt szempont mindegyike jelentős különbséget tár fel, melyek egymással is összefüggésben vannak. Első látásra furcsa lehet, hogy a modell paraméterek a bayesi statisztikában valószínűségi változókként definiáltak, mivel azonban az értéküket sohasem ismerjük meg pontosan, ez tulajdonképp nem okoz különösebb problémát. A szubjektív valószínűség elsősorban a prior felállítása esetén merül fel, ami a bayesi logika egyik alapköve. A klasszikus statisztika képviselői szerint ez nem lenne megengedhető, amire a bayesiek a nem informatív priorok felállításával válaszolnak. A külső információk felhasználásával kapcsolatban még egy szemponttal egészíteném ki a táblázatot, mégpedig azzal, hogy a bayesi statisztikában a külső információ felhasználása minden esetben jól dokumentált a prioron keresztül, míg ez a klaszszikus statisztikában nem mindig áll fenn. A klasszikus statisztika ellentmondó feltevésével szemben – miszerint a mintavételt tetszőleges alkalommal képesek vagyunk megismételni – a bayesi statisztika nem támaszkodik ilyen tételre, csak az egyetlen mintában lévő (és előzetes) információkra támaszkodunk.

A statisztikus szakmán belül kezdetben a bayesi megközelítés is erős volt, majd a XX. század elején a frekventista, klasszikus megközelítés vette át az egyeduralmat, a szakmai kisebbségnek számító bayesi statisztikusok azonban az utóbbi évtizedek munkásságának és technikai fejlődésének köszönhetően egyre nagyobb teret és elismerést nyernek.

4.2. Prior és poszterior eloszlás, bayesi frissítés

Az előző alpontokban már említett, mindenki által ismert Bayes-tétel legegyszerűbb formájában az alábbi:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{4.1}$$

ahol A és B események, $P(A|B)$ pedig feltételes valószínűség. Az A és B eseményeket leggyakrabban valószínűségi változók segítségével definiáljuk. „A Bayes-tétel szavakban kifejezve azt a megállapítást tartalmazza, hogy valamilyen eseményre vonatkozó hipotézis valószínűsége a további megfigyelések figyelembevétele folytán mennyire változik meg. Ez az indukció alapvető jellegzetességének szabatos megfogalmazása" (Theiss, 1971). A bayesi statisztika gyakorlatilag ezt az egy összefüggést alkalmazza különböző helyzetekben.

Mivel a bayesi statisztikusok a szóban forgó eloszlások paramétereit nem adott értékként, konstansként, hanem egy eloszlással rendelkező véletlen változóként fogják fel, a fenti képletben helyettesíthetjük A-t a paraméterekkel (θ), B-t pedig a megfigyelésekkel (y). A paramétereket gyakran nem megfigyelhetőnek (unobservable), az adatokat pedig megfigyelhetőnek (observable) nevezik az irodalomban. Ekkor (4.1) az alábbi alakban írható (egyelőre egy paramétert feltételezve):

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4.2}$$

ahol $p(\theta)$ az a priori eloszlás (prior density), azaz az a tudás, amely már az adatok vizsgálata előtt rendelkezésre állt a vizsgált paraméterről, ha úgy tetszik, ez a szubjektív elem. $p(y|\theta)$ az adatok likelihoodja adott modell esetén, $p(y)$ pedig az adatok marginális eloszlása, melyre az alkalmazások nagy részében nincs szükség, ezért sokszor csak konstansként tekintjük, illetve el is hagyjuk. Természetesen θ értékét a vizsgálat után sem tudjuk, azonban ismerjük az a poszteriori eloszlását (posterior distribution), melyet $p(\theta|y)$ jelöl. A dolgozatban a fenti eloszlásokat röviden prior és poszterior eloszlásoknak, vagy röviden csak priornak és poszteriornak fogom nevezni. A poszterior eloszlás összegzi a megelőző ismereteinket, valamint az adatokból nyert információkat. Úgy is fogalmazhatunk, hogy a poszterior azt mutatja meg, hogy menynyit tudunk a nem megfigyelhető paraméterekről a megfigyelhető adatok elemzése után.

A klasszikus statisztika (amit a bayesi szakirodalomban a classical, vagy még gyakrabban a frequentist jelzővel illetnek) csak a (maximum) likelihoodra koncentrál, azt keresi, hogy melyek azok a paraméterértékek, melyek esetén a megfigyelt adatok bekövetkezésének valószínűsége maximális. Ebben az esetben természetesen nincs szükség a prior információk ismeretére, explicit megfogalmazására.

Ahogy említettem, (4.2) jobb oldalának nevezője általában érdektelen, így a legfontosabb bayesi formula az alábbira egyszerűsödik, ahol $\propto$ egyenes arányosságot jelöl:

$$p(\theta|y) \propto p(y|\theta)\, p(\theta) \tag{4.3}$$

A poszterior (eloszlás) arányos a likelihood és a priori (eloszlás) szorzatával.

A bayesi gondolatkör könnyedén lehetővé teszi a beérkező adatok folyamatos feldolgozását, ilyen értelemben a prior és a poszterior fogalmak relatívak. Ami ma még poszterior információ θ-val kapcsolatban, holnap már a következő elemzés priorját alkothatja. Ezt a tulajdonságot bayesi frissítésnek (bayesian update) nevezi a szakirodalom.

Amennyiben szükséges, $p(y)$ is kiszámítható, mégpedig az alábbi módon:

$$p(y) = \int_{\theta \in \Theta} p(y|\theta)\, p(\theta)\, d\theta \tag{4.4}$$

ahol az integrálás határait θ értelmezési tartománya adja meg. Diszkrét eloszlások esetén az integrálás helyébe szumma lép, de ez nagyon ritkán fordul elő a gyakorlatban, hisz a paraméterek eloszlása jellemzően folytonos.

Mindeddig nem szóltam arról, hogy a fenti függvények (elsősorban a poszterior és a prior) milyen formát öltenek. Néhány esetben lehetőségünk van ezeknek „kényelmes” formáját választani, ezekről az esetekről lesz szó a következő alpontban.

4.3. Prior eloszlások, a konjugált prior

A likelihood formáját a modell, a feltételezett adatgeneráló folyamat (Data Generating Process, DGP) határozza meg. Amennyiben például a modellezendő folyamat eredményeképp bináris változót figyelünk meg, a feltételezett folyamat is ennek

megfelelően binomiális eloszlású lesz. Amennyiben darabszámokat modellezünk, a folyamat lehet pl. Poisson, folytonos változó esetén többek között normális stb.

A prior eloszlás alakját a kutató határozza meg, úgy, hogy az megfelelően tükrözze a nem megfigyelhető paraméterekkel kapcsolatos előzetes ismereteket. A bayesi módszerek leggyakrabban támadott eleme épp a prior eloszlás, hisz az objektivitásra törekvő statisztikatudományban nem természetes egy szubjektív prior választásának lehetősége. Ennek kivédésére a kutatók gyakran alkalmaznak úgynevezett nem informatív priorokat[11]. A nem informatív priorok a lehető legkevesebbet mondanak a szóban forgó paraméterről. Ha például egy binomiális eloszlás paraméteréhez kívánunk ilyen priort előállítani, választhatjuk az egyenletes eloszlást a $[0,1]$ intervallumon. Azaz előzetes feltevésünk, hogy a paraméter tetszőleges értéket felvehet az értelmezési tartományán, egyenlő valószínűséggel. Természetesen léteznek informatív priorok, melyek ténylegesen tartalmaznak előzetes információkat a vizsgált paraméter(ek)ről. Meg kell még említenünk az ún. improper prior eloszlásokat, melyek nem tényleges sűrűségfüggvénnyel definiáltak, azaz $\int_{-\infty}^{\infty} p(\theta)\,d\theta \neq 1$. Ilyenre lehet példa egy regressziós paraméter esetén az a feltételezés, hogy értéke a valós számok halmazán bármely értéket azonos valószínűséggel vehet fel, azaz $p(\theta) \propto 1$. Egy másik szélsőséges eset, ha az összes valószínűséget egyetlen pontra tesszük, ezzel a paraméterben rejlő bizonytalanságot megszűntetve, ám ez az improper prior esetéhez képest sokkal ritkább megoldás.

A választható priorok egy másik szempontból lehetnek kifejezetten kényelmesek. Mivel a poszterior eloszlás a likelihood és a prior eloszlás szorzatával arányos, érdemes olyan priort választani, aminek függvényformája megegyezik a likelihoodéval. Mindez egyben azt is jelenti, hogy a priorban lévő információ felfogható az előzetes információk megfigyelésekké alakításaként is. Az azonos függvényforma biztosítja, hogy a poszterior formája is megegyezzen a prioréval, azaz a szorzás elvégzése után is ugyanabban az eloszláscsaládban maradhatunk. Az ilyen, gyakran alkalmazott priorokat hívjuk konjugált priornak. Belátható, hogy amennyiben a DGP az exponenciális eloszlás-

[11] A nem informatív priorok másik elnevezése diffúz prior.

családhoz tartozik, úgy minden esetben létezik konjugált prior (Geweke, 2005, p. 42-43). A konjugált prior eloszlások speciális esete a természetes konjugált prior, ami azt jelenti, hogy a likelihood, a prior és a poszterior ugyanazon eloszláscsaládhoz tartoznak. Így például a normális eloszlás konjugált prior eloszlása a normális, azaz természetesen konjugált priorról beszélhetünk.

Amennyiben nem a konjugált prior eloszlásokból választunk, úgy a poszterior eloszlás jellemzően nem ismert eloszlás formáját fogja felvenni. Nem informatív prior eloszlás alkalmazása esetén az elemzést gyakran kiegészíti egy érzékenységvizsgálat a prior megválasztásának tekintetében. Ezzel elkerülhető az, hogy a végeredmény pusztán a prior megválasztásából adódóan alakul ki.

A bayesi statisztikával foglalkozók megosztottak abban a kérdésben, hogy alkalmazhatók-e az elemzést tárgyát képező adatok valamelyest informatív prior előállítására. A konzervatív réteg szerint nem, hisz a prior az adatok előtti tudást összegzi, míg mások, az empirikus irányzat korlátozottan ugyan, de alkalmazza az adatokat a prior előállítására. Amennyiben teljesen az adatok alapján határoznánk meg a priort eloszlást, kétszer vennénk figyelembe azokat, ami a standard hibák torzításához vezetne.

Az alábbi, 4-2. táblázatban a legfontosabb eloszlások konjugált prior eloszlásait mutatom be. A diszkrét esetek mellett természetesen a gyakran alkalmazott folytonos eloszlások kapcsán is szerkeszthető 4-2. táblázattal analóg felsorolás.

4-2. táblázat: Diszkrét eloszlások konjugált prior eloszlásai

Modell	Modell paraméter	Konjugált prior	Prior paraméterek
Bernoulli	p	Béta	α, β
Binomiális	p	Béta	α, β
Poisson	λ	Gamma	k, θ
Negatív binom	r, p	Béta	α, β
Multinomiális	$\mathbf{p} = (p_1 \quad p_2 \quad \dots \quad p_k)$	Dirichlet	$\boldsymbol{\alpha} = (\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_k)$

Forrás: Bayesi témájú kötetek levezetései alapján saját gyűjtés

Végezetül a prior eloszlások egy gyakran alkalmazott családjára, az ún. Jeffreys-féle (Jeffreys, 1946) priorokra hívom fel a figyelmet, melyek a modell különböző parametrizációira invariáns prior eloszlások. Az ajánlott priorok a Fisher-féle informá-

ciós „mátrix” (lásd pl. Hunyadi (2001, 114)) determinánsának gyökével arányosak, azaz $p(\theta) \propto \left|\Im(\theta)\right|^{\frac{1}{2}}$, ahol $\Im(\theta) = E_{Y|\theta}\left(-\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right)$.

4.4. Modell eredmények bemutatása

Amennyiben azonosítottuk az adatgeneráló folyamatot, kiválasztottuk a prior eloszlás, (4.3) alapján (egy konstans híján) megkapjuk a paraméter poszterior eloszlását, azaz az eddigi eredményünk egy kvázi sűrűségfüggvény. A paraméterrel kapcsolatban azonban nem egy eloszlást, hanem a klasszikus statisztikában megszokott mutatókat szokás bemutatni. Így – a klasszikus pontbecsléshez hasonlóan – eredmény lehet a poszterior átlag, a medián, kiválasztott percentilisek, vagy akár a módusz. A bayesi irodalom ezeket a mutatókat nem „ízlés” szerinti választásként tárgyalja. A döntéshozónak lehetősége van egy ún. veszteségfüggvény definiálására (loss function), mely a következő általános formát ölti: $L(\tilde{\theta}, \theta)$. Amennyiben $L(\tilde{\theta}, \theta) = (\theta - \tilde{\theta})^2$, azaz a veszteség függvény négyzetes, úgy $\tilde{\theta} = E(\theta | y)$, azaz a poszterior átlagot használjuk. Amennyiben a függvény abszolút érték függvény, úgy a mediánt (illetve általános esetben súlyozott abszolút érték esetén tetszőleges kvantilist), az ún. mindent-vagy-semmit veszteség függvény esetén pedig a poszterior móduszt. Vegyük észre, hogy a klasszikus, maximum likelihood alapú statisztikai eljárások tulajdonképp ezt az értéket, ezt a veszteségfüggvényt alkalmazzák. Ennek megfelelően a nem informatív prior és mindent-vagy-semmit veszteségfüggvény mellett végzett következtetések a gyakorlatban a klasszikus ML becslésekre vezetnek.

A pontbecslés mellett szükségünk lehet a klasszikus statisztikából ismert konfidencia intervallum bayesi megfelelőjére is. Mivel ismerjük a poszterior sűrűségfüggvényt, integrál segítségével meghatározható minden esetben olyan „hihető, valószínűsíthető halmaz” (credible set - CS), mely a poszterior eloszlás $100 \times (1-\alpha)\%$-át lefedi. C tehát teljesíti az alábbi összefüggést:

$$p(\theta \in C | y) = \int_{\theta \in C} p(\theta | y) d\theta = 1 - \alpha \tag{4.5}$$

A gyakorlatban jellemzően végtelen sok ilyen halmaz képzelhető el. Amennyiben a poszterior eloszlás egymóduszú, úgy a halmaz helyett intervallumokról beszélhetünk, egyéb esetben semmi sem biztosítja, hogy a halmaz összefüggő. A legegyszerűbb megoldás, hogy a poszterior eloszlás két szélén $\frac{\alpha}{2}$ nagyságú területeket határozunk meg. Gyakran azonban a sok lehetséges halmaz közül azt szeretnénk kiválasztani, amelyik valamilyen értelemben a legkisebb. Az ilyen halmazok neve legnagyobb poszterior sűrűségű halmaz/intervallum (highest posterior density - HPD), ami nem feltétlenül szimmetrikus a „levágott" területek tekintetében. Belátható (lásd pl. Casella-Berger, 2002, p. 448, vagy Lénárt-Rappai, 2001), hogy a HPD akkor áll elő, ha (egymóduszú poszterior esetén) az intervallum két pontjában a poszterior sűrűségfüggvény megegyezik. A hazai szakirodalomban Lénárt és Rappai feszegetik a legrövidebb intervallum kérdését a klasszikus statisztika varianciabecslésének hibahatárával kapcsolatban, ahol a probléma az χ^2 eloszlás esetén merül fel. Tanulmányukban különböző szabadságfokokra és megbízhatósági szintekhez tartozóan megadják a legrövidebb intervallumokat, valamint vizsgálják a „szokásos" és optimális intervallumok egymáshoz viszonyított nagyságát is.

4.5. Modellszelekció és hipotézisvizsgálat

A modellezési gyakorlatban jellemzően nem csupán egy vizsgált modell szerepel, hanem több, és a feladat az ezek közötti választást is magában foglalja. A modellek közötti választás esetén ugyanazt az egyszerű szabályt, a Bayes-tételt alkalmazzuk, mint a poszterior eloszlás levezetésében. Ahogy azt láttuk, a modellt a prior és a DGP definiálja formálisan. Amennyiben feltesszük, hogy m különböző modellünk van, azokat M_i-vel jelölve, ahol $i = 1, 2, \ldots, m$, úgy (4.2) az alábbi, bővített alakban írható:

$$p\left(\theta^{(i)} \middle| y, M_i\right) = \frac{p\left(y \middle| \theta^{(i)}, M_i\right) p\left(\theta^{(i)} \middle| M_i\right)}{p\left(y \middle| M_i\right)} \tag{4.6}$$

A paraméterek függenek a modelltől, ezért amennyiben több modellről beszélünk, az i-edik modell paramétereit $\theta^{(i)}$-vel jelöljük az egyértelműség kedvéért.

Amennyiben modellek közötti választás a feladatunk, a bayesi logika szerint a szokásos módon járunk el. Valószínűségi kijelentést teszünk arról, amit nem tudunk (mi a valós modell) az alapján, amit tudunk (megfigyelt adatok, poszterior). Ki kell számítanunk tehát a modell poszterior (megfigyeléseket követő) valószínűségét. Ahogy a paraméterek esetén, a modellek esetén is szükség van priorok meghatározására, $p(M_i)$-re, amit prior modell valószínűségnek fogunk nevezni, és azt jelenti, hogy milyen valószínűséggel gondoljuk, az adatok ismerete előtt, hogy az i-edik modell helyes. Újra Bayes (4.1) tételét használva, $B = M_i$ és $A = y$ mellett az alábbi összefüggést kapjuk:

$$p(M_i|y) = \frac{p(y|M_i)\,p(M_i)}{p(y)} \tag{4.7}$$

A likelihood helyét elfoglaló $p(y|M_i)$ kifejezés neve marginális likelihood. Meghatározása integrálással történik:

$$p(y|M_i) = \int p(y|\theta^{(i)}, M_i)\, p(\theta^{(i)}|M_i)\, d\theta^{(i)} \tag{4.8}$$

azaz a modell marginális likelihoodja csak a likelihoodtól és a paraméter priorjától függ. Mivel (4.7) nevezőjének meghatározása bonyolult, valamint jellemzően nem egy modell valószínűségére vagyunk kíváncsiak, hanem modelleket szeretnénk összehasonlítani, ezért az úgynevezett poszterior esélyhányadost (posterior odds ratio) használjuk, ami egyszerűen a modellek poszterior valószínűségének hányadosa, azaz

$$O_{ij} = \frac{p(M_i|y)}{p(M_j|y)} = \frac{\dfrac{p(y|M_i)\,p(M_i)}{p(y)}}{\dfrac{p(y|M_j)\,p(M_j)}{p(y)}} = \frac{p(y|M_i)}{p(y|M_j)}\frac{p(M_i)}{p(M_j)} \tag{4.9}$$

A fenti kifejezés segítségével tehát két modellt hasonlíthatunk össze. A jobb oldali hányadosban gyakran $p(M_i) = p(M_j)$ feltételezéssel élünk, azaz a modellek a priori valószínűsége megegyezik, vagyis nem informatív modell priorokat alkalmazunk. Ekkor a poszterior hányados a két marginális likelihood hányadosára egyszerűsödik. A modell-összehasonlításban betöltött kitüntetett szerepe miatt ennek a kifejezésnek külön neve van, az ún. Bayes faktor. Amennyiben a poszterior hányados, illetve a Bayes faktor 1-nél nagyobb értéket vesz fel, úgy az i-edik modell valószínűbb. Az elgondolás

előnye, hogy a bayesi statisztikus csupán a Bayes faktort közli, a modellek közötti döntést a döntéshozó saját prior modellvalószínűségeivel korrigálva hozhatja meg. A (4.8) összefüggés nem adható meg hagyományos módon, amennyiben a prior improper.

Szorosan kapcsolódik a modellválasztás témaköréhez a bayesi hipotézisvizsgálat. Ahogy azt már megszokhattuk, a döntés ebben az esetben is a megszokott logikát követi. A prior esélyhányados a két (null és alternatív) hipotézis egymáshoz viszonyított valószínűségét mutatja meg, míg a poszterior esélyhányados azt, hogy ugyanez az arány hogy néz ki az adatok megismerésének hatására. A poszterior esélyhányados felírható az alábbi módon (Hunyadi, 2011b)

$$POO = \frac{\Pr(H_1)}{\Pr(H_0)} \frac{\int \Pr(y|\theta, H_1) \Pr(\theta|H_1) \partial\theta}{\int \Pr(y|\theta, H_0) \Pr(\theta|H_0) \partial\theta} \tag{4.10}$$

amennyiben értéke egynél kisebb, akkor a nullhipotézist tartjuk valószínűbbnek.

4.6. Előrejelzés

A legmegfelelőbb modell kiválasztása után gyakori feladat – sokszor az egész modellalkotás célja – előrejelzés készítése. Ahogy azt már megszokhattuk, az előrejelzést ismeretlen, nem megfigyelhető értékként fogjuk fel, a későbbiekben y^*-gal jelöljük. Célunk egy feltételes valószínűség, $p(y^*|y)$ meghatározása a poszterior ismereteink alapján, melyet az alábbi formulával tehetünk meg:

$$p(y^*|y) = \int p(y^*|y,\theta)\, p(\theta|y)\, d\theta \tag{4.11}$$

ahol a már megismert poszterior eloszlás és az előrejelzés adatoktól és paraméterektől függő feltételes eloszlása (predictive density) szerepel. Természetesen elképzelhető az a priori ismeretek alapján történő előrejelzés is, ekkor nem szerepel feltételként y. Amennyiben több modell segítségével is szeretnénk előrejelzést készíteni, (4.11)-et ki kell egészítenünk a modellre vonatkozó feltételek jelölésével is.

A fentiekben röviden összefoglaltam a bayesi statisztika alapvető kellékeit, eljárásait. A konkrét alkalmazások „csupán” a fenti elméleti eredményeket alkalmazzák. Mindez sok esetben összetett egy vagy többváltozós függvények integrálásából adódó értékek kiszámítását jelenti, többek közt (4.4), (4.5), (4.8) vagy (4.11) esetén. Ezeket a

műveleteket nem minden esetben tudjuk analitikusan elvégezni a hatékonyság megtartása mellett, numerikus, vagy szimulációs módszereket kell igénybe vennünk. Ahogy már említettem, a bayesi statisztika elterjedésének egyik korlátja sokáig az volt, hogy nem álltak rendelkezésre a szükséges szimulációs feladatokhoz elegendően gyors számítógépek. A következő alfejezetben alapvető, a bayesi módszerek használata esetén elengedhetetlenül szükséges technikákat mutatok be, melyek statisztikai szoftverek segítségével egyszerűen implementálhatóak és gyorsan futtathatóak.

4.7. Szimulációs, Monte-Carlo és Markov-lánc Monte-Carlo technikák

A fentiekben bemutattam a bayesi gondolkodás alapjait. A gondolatmenet tulajdonképp egyszerű, csupán valószínűségelméleti alapok szükségesek hozzá. Könnyen ütközhetünk azonban számítási nehézségekbe, sok esetben kell integrálást végeznünk, vagy egy adott sűrűségfüggvényű eloszlásból véletlen számokat generálnunk. A poszterior eloszlás ugyan nem igényli integrál meghatározását (amennyiben a normalizáló konstansra nincs szükségünk), a benne rejlő információ kinyerése azonban gyakran nem oldható meg analitikusan. Egydimenziós esetben a grafikus ábrázolás kézenfekvő, összetettebb modelleknél azonban ez nem megoldható. A poszterior átlag, variancia, vagy akár a percentilisek integrál, illetve véletlen szám generátor segítségével közelíthetők. Ha nem konjugált analízist végzünk, ezek a poszterior eloszlások nem standard, ismert eloszlások, az integrálás sok esetben csak numerikusan végezhető el.

Monte-Carlo (MC) módszerek összefoglaló névvel illetünk rengeteg eljárást, technikát, melyek közös jellemzője, hogy véletlen szám sorozatok generálásán alapulnak. A módszerek népszerűségének oka rendkívül egyszerű: analitikusan követhetetlen feladatok eredményeit vagyunk képesek tetszőleges közelítéssel meghatározni velük. A robbanásszerű elterjedéshez a matematikai alapok lefektetésén kívül szükség volt egy másik összetevőre, a véletlen értékeket generáló, a számításokat gyorsan elvégző számítógépekre.

A véletlen események felhasználásának ötlete nem új a statisztikában, már a számítógépek megjelenése előtt is voltak alkalmazásai, elég csak a Buffon-féle tűproblémára (π közelítése a padlóra dobott tűk segítségével a XVIII. században), vagy a Gossett nevéhez fűződő, t-eloszlásról szóló cikkre (Student, 1908) utalnom. A véletlen

értékek felhasználásának történetéről, a módszerek fejlődéséről az érdeklődő Olvasónak például Robert-Casella (2011) nyújt kimerítő irodalomjegyzéket[12].

A valós statisztikai alkalmazások felsorolása azok sokszínűsége miatt szinte lehetetlen: hagyományos alkalmazási terület a különböző tesztek erőfüggvényeinek kiszámítása, kritikus értékek, vagy becslőfüggvények jellemzőinek (paraméterek, MSE, percentilisek stb.), konfidencia intervallumok takarási valószínűségének meghatározása. Ebbe a körbe tartoznak a Bootstrap és Jackknife módszerek is. Ezen kívül fontos szerepet töltenek be az MC és Markov-lánc Monte-Carlo (MCMC) módszerek a bayesi statisztikában, ahol a feladat összetett, sokdimenziós sűrűségfüggvények (poszteriorok) leírása, amely szinte minden esetben integrálok meghatározását jelenti a gyakorlatban. A harmadik nagy felhasználási terület a sztochasztikus optimalizáció, amely összetett függvények szélsőértékeit, illetve szélsőérték helyeit keresi. Tipikusan ilyen probléma összetett likelihood függvények maximumának keresése ML becslés meghatározásakor.

A fejezet felépítése a témával foglalkozó szakkönyvek (Albert, 2009; Casella-Berger, 2002; Rizzo, 2008; Robert-Casella, 2004, 2010) struktúráját követi. Az általános bevezető után a különböző eloszlásokból való véletlen érték generálás egyszerű technikáit szemléltetem, majd az egyik gyakran alkalmazott területet, a Monte-Carlo integrálást és az ehhez kapcsolódó varianciacsökkentő módszereket tárgyalom. A sokdimenziós, ismeretlen eloszlásokból történő mintavétel leggyakrabban a Markov-láncokhoz kapcsolódó, ún. Markov-lánc Monte-Carlo technikákhoz kötődik, ennek a módszercsaládnak a rövid bemutatása zárja a fejezetet.

4.7.1. Véletlen szám generálási technikák

A számítógépes véletlen szám generálás alapja az egyenletes eloszlás. Nem kívánok részletesen foglalkozni azzal, hogy a számítógépek csupán ún. pszeudo véletlen szám létrehozására képesek. Az ilyen módszerek egy hosszú sorozatot állítanak elő, melyek matematikai tulajdonságaik alapján megfelelő minőségűnek tekinthetők[13]. Valamennyi, számítási célokat is szolgáló programcsomag tartalmaz egyenletes eloszlásból

[12] A Statisztikai Szemle hasábjain megjelent rövid magyar nyelvű összefoglalást lásd Kehl (2012a).

[13] Létezik olyan R csomag, mely a random.org honlapon keresztül igazi véletlen számokat használ, azonban tudományos célokra a megfelelően jó minőségű pszeudo véletlen számok teljesen elfogadottak.

származó véletlen szám generátort, R-ben ez a függvény a runif(), Excelben a vél(), Matlabban pedig a rand(). A szoftverek alapértelmezésben az ún. Mersenne Twister eljárást használják, amely egy gyors, jó minőségű, pszeudo véletlen számokat generáló algoritmus (Matsumoto-Nishimura, 1998). Véletlen számon tehát ezentúl pszeudo, számítógép által generált véletlen számokat értek.

A különböző eloszlásokból származó véletlen értékek hatékony generálásának komoly irodalma van, aminek ismertetése meghaladja a dolgozat kereteit. Szerencsére R-ben a rendelkezésre álló leghatékonyabb (leggyorsabb) módszerek implementálása megtörtént, a szükséges függvényeket a base csomag, így minden telepített verzió tartalmazza. A kevésbé gyakran alkalmazott eloszlások sok esetben csak előre nem telepített, speciális csomagokban találhatók meg, vagy azokban sem. Az ismert eloszlásokhoz tartozó parancsok elnevezései azonosan épülnek fel, p#() az eloszlás-, d#() a sűrűségfüggvény értékét számítja ki, q#() segítségével a kvantilisek határozhatók meg, míg r#() az adott eloszlásból származó véletlen mintát generál, ahol # a teljesség igénye nélkül az alábbi értékeket veheti fel.

4-3. táblázat: Néhány beépített eloszlás elnevezése és R függvénye

Eloszlás	R (#)	Eloszlás	R (#)
Beta	beta	Khi-négyzet	chisq
Binomiális	binom	Lognormális	lnorm
Cauchy	cauchy	Neg. binomiális	nbinom
Egyenletes	unif	Normális	norm
Exponenciális	exp	Poisson	pois
Gamma	gamma	Student-t	t

Természetesen a függvényeknek meg kell adnunk a szükséges információkat, a kívánt paramétereket, egyes esetekben lehetőség van speciális argumentumok megadására is. A 4-3. táblázatban nem szerepelnek olyan ismert eloszlások, melyekre később szükségünk lesz, ezek vagy egyszerűen implementálhatók, vagy valamelyik kiegészítő csomagban megtalálhatók. Sok esetben hasznos a sample() parancs, segítségével könnyen generálhatunk Bernoulli, vagy multinomiális véletlen változókat[14].

Inverz eloszlásfüggvény módszer

[14] sample(0:1, size = 100, replace = TRUE); sample(1:5, size = 100, replace = TRUE, prob = c(5,4,3,2,1))

A véletlen számok generálásának talán legegyszerűbb módja az ún. inverz eloszlásfüggvény módszer (inverse transform method), hátránya azonban, hogy nem minden esetben alkalmazható. Ha X folytonos véletlen változó $F_X(x)$ eloszlásfüggvénnyel, akkor $U = F_X(X) \sim Unif(0,1)$, azaz egyenletes eloszlású a $[0,1]$ intervallumon (ún. probability integral transform).

Hasonlóan belátható, hogy (amennyiben az inverz létezik) $F_X^{-1}(U)$ eloszlása megegyezik X eloszlásával, azaz a feladatunk $F_X^{-1}(U)$ meghatározása, majd véletlen $u \sim Unif(0,1)$ generálása és $F_X^{-1}(u)$ kiszámítása. Az elmélet kiterjeszthető (Angus, 1994), többek között a folytonos eloszlásokról diszkrétekre az inverz fogalom általánosításával.

Példaként tekintsük a Cauchy eloszlás eloszlásfüggvényét: $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right)$, amiből az $F^{-1}(u) = \mu + \sigma\tan\left(\pi\left(u - \frac{1}{2}\right)\right)$ inverz egyszerű átrendezéssel adódik.

Nincs más dolgunk tehát, mint a kívánt számú 0–1 közötti egyenletes eloszlású érték generálása, majd azokon az inverz transzformáció elvégzése. Hasonló módon állítható elő például exponenciális, logisztikus vagy Rayleigh-eloszlású véletlen változósorozat. A generáláshoz szükséges néhány soros programot a Függelékben mutatom be. A csonkolt normális eloszlás példájával illusztrálja a módszert Várpalotai (2008) dolgozata függelékében.

Direkt transzformációs módszer

A kívánt véletlen értékek előállítása sok esetben megoldható ismert eloszlások közötti matematikai összefüggések segítségével. Standard normális véletlen változók négyzetre emelésével és összegzésével állíthatunk elő χ_k^2 eloszlást. Casella és Berger (2002, p. 627) átfogó képet adnak a gyakran alkalmazott eloszlások kapcsolati hálójáról, ami alapján a módszer könnyedén implementálható. Példaként említhetném még a lognormális véletlen változó generálását standard normális eloszlásból, vagy F-, illetve Student t-eloszlású értékek létrehozását. Transzformáción alapul a normális eloszlású változókat generáló Box-Müller (1958) algoritmus is, amely egy egyenletes változópár-

ból normális eloszlású változópárt állít elő. A direkt transzformációs módszer nyilvánvaló hátránya, hogy nem szokványos eloszlások esetén ilyen lehetőség ritkán áll fent.

Az elfogadás-elutasítás módszere

Az elfogadás-elutasítás módszer (acceptance-rejection method) alkalmazásához szükségünk van egy olyan eloszlásra (forrás eloszlás – g), melyből könnyedén tudunk véletlen számokat generálni, ráadásul megfelelően „közel" van ahhoz az eloszláshoz, melyből generálni szeretnénk (cél eloszlás – f). Legyen X és Y két véletlen változó és jelölje sűrűségfüggvényüket rendre f és g. Tegyük fel továbbá, hogy létezik olyan c konstans, melyre

$$\frac{f(t)}{g(t)} \le c \tag{4.12}$$

fennáll minden olyan t-re, ahol $f(t) > 0$. A cél elegendően alacsony, lehetőleg a legalacsonyabb c megtalálása (4.12)-ben egy olyan g-hez, amely elég hatékony és könnyen generálható.

Amennyiben megtaláltuk a megfelelő eloszlást és a hozzá tartozó konstanst, az alábbi lépéseket kell elvégeznünk:

1. Generáljunk egy véletlen y számot Y eloszlásból.
2. Generáljunk $u \sim Unif\left(0, c \times g(y)\right)$ egyenletes eloszlású véletlen értéket.
3. Amennyiben teljesül $u < f(y)$, fogadjuk el y-t X-ből származó véletlen számként, azaz $x := y$, ellenkező esetben utasítsuk el, majd térjünk vissza az *1.* pontra.

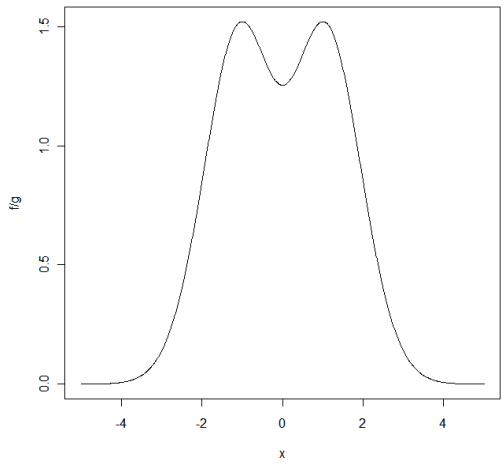
Adott $Y = y$ feltételhez tartozó elfogadási valószínűség tehát $\frac{f(y)}{cg(y)}$ a 3. lépés és az egyenletes eloszlás eloszlásfüggvénye alapján. Bármely iteráció összesített (feltétel nélküli) elfogadási valószínűsége $\int_{-\infty}^{\infty} \frac{f(y)}{cg(y)} g(y)\,dy = \frac{1}{c}$, így egy X-ből származó véletlen szám átlagosan c iterációt, azaz $2c$ (c a forrás, c az egyenletes eloszlásból) vélet-

len szám generálását igényli. Amennyiben nem találjuk meg a megfelelő (minimális) c-t, a módszer alkalmazható marad, de nem hatékony. A c konstans tartalmilag a javasolt forráseloszlás maximális távolságát méri a céleloszlástól.

Az elfogadás-elutasítás módszer bemutatására standard normális változókat állítunk elő. Első lépésként egy, a standard normálishoz hasonló, könnyen generálható eloszlást kell keresnünk. Legyen ez a már megismert Cauchy-eloszlás standard változata, hisz abból könnyedén tudunk generálni a megírt inverz eloszlásfüggvény eljárás vagy beépített függvény segítségével. Megfelelő választás lenne természetesen bármely egyéb, a valós tengelyen értelmezett függvény is. Praktikus, ha olyan függvényt választunk, mely vastagabb eloszlásszéllel rendelkezik, mint a céleloszlás. Második lépésként meg kell határoznunk a lehető legkisebb konstanst (4.12)-ben a kiválasztott g-hez, ehhez írjuk fel a sűrűségfüggvények hányadosát:

$$\frac{f(x)}{g(x)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}{\frac{1}{\pi\left(x^2+1\right)}} = \frac{\sqrt{\pi}\left(x^2+1\right)}{\sqrt{2}} e^{-\frac{1}{2}x^2} \leq c$$

Amennyiben ábrázoljuk a hányadost, láthatjuk, hogy az felülről korlátos, azaz a Cauchy eloszlás megfelelő forrás eloszlás, amennyiben a normális a cél eloszlás.



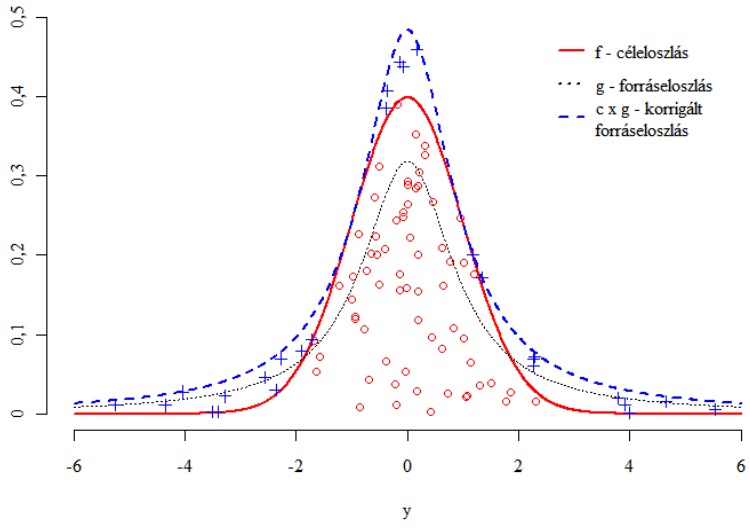
4-1. ábra: A standard normális és standard Cauchy eloszlások hányadosa

Keressük meg azokat az x_0 értékeket, melyeknél a függvény a maximumát veszi fel. A hányados deriváltja alapján könnyen megállapítható, hogy a függvénynek két

maximumhelye van az $x_0 = \pm 1$ pontokban (valamint lokális minimuma az $x = 0$ pontban). A maximumhelyeken a függvény értéke, azaz a lehetséges minimális konstans:

$$\frac{f(x_0)}{g(x_0)} = \sqrt{2\pi} e^{-\frac{1}{2}} = c \approx 1,52$$

A módszer megértését segíti az 4-2. ábra, melyen a standard normális (f), a Cauchy (g) és a konstanssal szorzott Cauchy $(c \times g)$ eloszlásokat, valamint 100 iterációval kapott véletlen értékeket ábrázoltam. A folytonos vonallal feltűntetett normális eloszlásból kívánunk generálni, méghozzá a pontozott vonallal ábrázolt Cauchy-eloszlás segítségével. Ehhez megkerestem azt a legkisebb c-t, amellyel a Cauchy-eloszlást szorozva a megnyújtott görbe lefedi a teljes céleloszlást (szaggatott vonal). Ezután a Cauchy-eloszlásból generálunk egy véletlen számot (y), amiről a 0 és $c \times g(y)$ közötti egyenletes eloszlásból származó véletlen érték és az adott pontban érvényes $f(y)$ dönt: a két görbe között helyezkedik-e el (elutasítás – kereszt) vagy a normális eloszlás sűrűségfüggvénye alatt (elfogadás – kör). A körrel jelölt pontok első koordinátái standard normális valószínűségi változóból generált véletlen értékeket képeznek.



4-2. ábra: Elfogadás-elutasítás módszer

Ahogy említettem, c egyben az egy céleloszlásból származó véletlen számhoz szükséges iterációk átlagos számát is jelenti. Amennyiben például 10 000 standard nor-

mális valószínűségi változót szeretnénk generálni[15], úgy átlagosan 15 200 iterációra, azaz 30 400 véletlen szám előállítására van szükség.

Végül megjegyzendő, hogy a módszer abban az esetben is alkalmazható, ha a céleloszlásnak csupán az alakját ismerjük, a normalizáló konstanst nem, ahogy ez a bayesi statisztikában gyakran előfordul. Ebben az esetben azonban $\hat{c}^{-1}$ nem az elfogadás valószínűsége, mert az ismeretlen normalizáló konstans „beszivárog" $\hat{c}$-be.

A következő alfejezetekben MC integrálási módszereket mutatok be. Integrálás eredményeképp kaphatjuk meg folytonos valószínűségi változók várható értékét, egyéb momentumait, kvantiliseit. A bayesi statisztikában mind a prior, mind a poszterior sűrűségfüggvénnyel írható le, a normalizáló konstans (ami az egységnyi integrálértéket biztosítja) azonban gyakran nem ismert és analitikusan nem is meghatározható. Az ilyen és ehhez hasonló esetek megoldására mutatok be olyan módszereket, melyek analitikusan nem kezelhető határozott integrálok meghatározására szolgálnak. Az ismert determinisztikus módszerek (Kehl, 2012b) a függvényt egyszerű alakzatokkal közelítik, hátrányuk, hogy magasabb dimenziószám esetén konvergenciájuk lassul. A véletlen értékek generálásán alapuló MC eljárások implementálásának egyszerűsége magasabb dimenziószám esetén is megmarad, ezért összetettebb, sokváltozós problémák esetén előszeretettel használják őket. Az MC módszerek leírása után a klasszikus MC becslés varianciáját csökkentő eljárások bemutatása következik.

4.7.2. Monte-Carlo integrálás

A Monte-Carlo-integrálás egy véletlen szám generáláson alapuló statisztikai módszer, mely a 40-es évek vége óta ismert, elsősorban Neumann János és Stanislaw Ulam munkásságának köszönhetően. A véletlen kísérletekből való következtetés gondolata – ahogy azt említettem – azonban már sokkal korábban, a XVIII. században felmerült (Buffon-féle tűprobléma), de az első számítógépek óriási lendületet adtak az alkalmazások elterjedésének.

Tudjuk, hogy ha X véletlen változó $g(x)$ sűrűségfüggvénnyel, akkor $h(X)$ transzformált véletlen változó várható értéke:

[15] A programot lásd a Függelékben.

$$\mu = E\left[h(X)\right] = \int_{-\infty}^{\infty} h(x)\, g(x)\, dx \tag{4.13}$$

Amennyiben rendelkezünk véletlen mintával X eloszlásából, úgy a függvényértékek átlaga (4.13) torzítatlan becslését adja, $n \to \infty$ esetén

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \mu \tag{4.14}$$

ahol X_i az i-edik mintaelemet reprezentáló véletlen változót jelöli.

Legyen $X \sim Unif(a,b)$, így $E\left[h(X)\right] = \frac{1}{b-a}\int_a^b h(x)\, dx$, azaz az integrálást visszavezetjük egy várható érték meghatározásának problémájára. A következő lépéseket kell elvégeznünk az $\int_a^b h(x)\, dx$ integrál Monte-Carlo közelítéséhez:

1. Legyen n független $X_i \sim Unif(a,b)$ véletlen változónk.
2. Számítsuk ki az átlagos függvényértéket: $\overline{h(X)} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$.
3. A közelítő integrál érték: $\hat{\mu} = (b-a)\overline{h(X)}$.

A MC integrál nem determinisztikus, hisz véletlen számokon alapszik. A becslés varianciája a véletlen számok darabszámának növelésével csökkenthető, de ez nyilván idő- és számításigényes. A konvergencia lassabb, mint a determinisztikus esetben (főként a trapezoid és Simpson-féle módszerhez képest), de magasabb dimenziókban is megmarad a konvergencia sebessége, míg a determinisztikus módszerek egyre lassabbá válnak.

Összetettebb problémák esetén jellemzően nem az MC integrálás implementálása okoz tehát nehézséget, hanem a lassú konvergencia. Fontos az olyan módszerek alkalmazása, melyekkel a variancia csökkenthető, viszont a számítási időt egyáltalán nem, vagy nem jelentősen növelik meg. Az alábbiakban tehát két olyan módszert mutatok be, melyek nem a mintaelemszám növelésével csökkentik a MC becslés varianciáját.

4.7.3. A variancia csökkentése MC módszerek esetén

A fejezetben a hagyományos MC-integrálás varianciáját, valamint négy olyan módszert mutatok be, melyek nem a mintaelemszám növelésével csökkentik a MC-becslés varianciáját. Belátható, hogy a mintaátlagon alapuló Monte-Carlo becslés torzítatlan, varianciája:

$$Var\left(\hat{\mu}\right)=\frac{\left(b-a\right)^2}{n^2}Var\left(\sum_i h\left(X_i\right)\right)=\frac{\left(b-a\right)^2}{n}Var\left(h\left(X\right)\right), \tag{4.15}$$

ahol a véletlen értékek függetlenségét használjuk ki. Jelen fejezetben épp ezt a függetlenséget sértjük meg oly módon, hogy a torzítatlanság továbbra is fennálljon, a variancia azonban csökkenjen. Az MC-becslés varianciája tehát az integrálási határoktól, a generált véletlen számok számától és a sűrűségfüggvény alakjától függ. A központi határeloszlás-tétel szerint pedig elégségesen nagy mintaelemszám mellett – ami gyakorlatilag mindig igaz – az integrálra (tehát a függvényértékek átlagára) vonatkozó MC-becslések normális eloszlást követnek, azaz $\hat{\mu} \sim N\left(\mu,\frac{\left(b-a\right)^2}{n}Var\left(h\left(X\right)\right)\right)$.

Példaként az $\int_2^4 e^{-x}dx$ integrál érték MC közelítésének eloszlásához meg kell tehát határoznunk $Var\left(h\left(X\right)\right)$ értékét. A várható értéket, az integrál tényleges értékét, valamint a többi paramétert ismerjük.

Tekintsük általánosan az $X \sim Unif\left(a,b\right)$ valószínűségi változót és határozzuk meg $h\left(X\right)=e^{-X}$ valószínűségi változó első és másodrendű momentumait! Alkalmazhatjuk (4.13)-t, ahol tudjuk, hogy $g\left(x\right)=\frac{1}{b-a}$.

$$E\left(e^{-X}\right)=\int_a^b e^{-x}\frac{1}{b-a}dx=\frac{1}{b-a}\left[-e^{-x}\right]_a^b=\frac{e^{-a}-e^{-b}}{b-a},$$

valamint

$$E\left(\left(e^{-X}\right)^2\right)=E\left(e^{-2X}\right)=\int_a^b e^{-2x}\frac{1}{b-a}dx=\frac{1}{b-a}\left[\frac{e^{-2x}}{2}\right]_a^b=\frac{e^{-2b}-e^{-2a}}{2\left(b-a\right)},$$

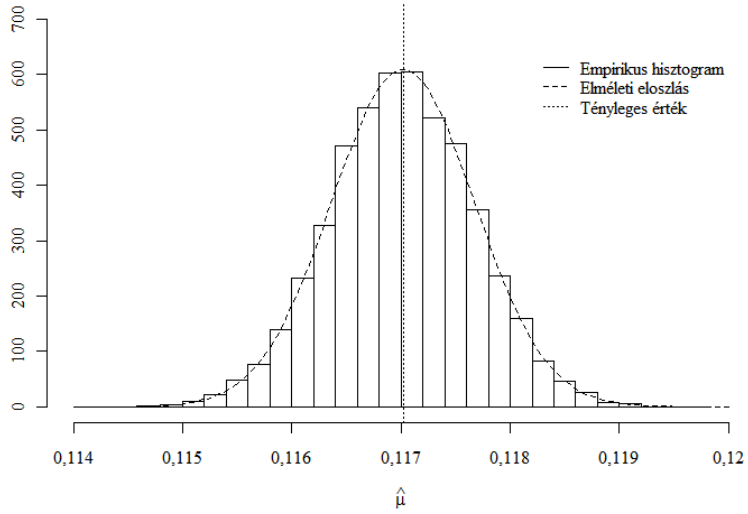
azaz a variancia:

$$Var\left(e^{-X}\right)=E\left(e^{-2X}\right)-\left[E\left(e^{-X}\right)\right]^2=\frac{e^{-2b}-e^{-2a}}{2(b-a)}-\frac{\left(e^{-a}-e^{-b}\right)^2}{(b-a)^2} \tag{4.16}$$

A (4.16) képletet alkalmazva $a=2, b=4$ esetére azt kapjuk, hogy $Var\left(e^{-X}\right)=\frac{e^2-1}{2e^8}\approx 0,001071645$, azaz (4.15) alapján a 10 000 elemű mintából álló becslés elméleti eloszlása:

$$\hat{\mu} \sim N\left(\mu, \frac{(b-a)^2}{n}\times Var\left(h(X)\right)\right)=N\left(e^{-2}-e^{-4}, \frac{2e^2-2}{10\,000e^8}\right).$$

A 4-3. ábrán az elméleti és a 20 000-szer megismételt, egyenként 10 000 véletlen számot felhasználó MC-becslés eredményei láthatók. Jól láthatóan a két eloszlás nagyon közel van egymáshoz. A 4-3. ábra ábrára és a becslés varianciájára később, a variancia-csökkentő eljárások bemutatása során visszautalok, ugyanez az elméleti sűrűségfüggvény szerepel a 4-4. ábrán is.



4-3. ábra: Az integrál érték átlagoláson alapuló MC-becslésének elméleti és empirikus eloszlása, valamint a tényleges érték

Az eloszlás ismerete lehetőséget teremt arra, hogy becslésünk köré konfidenciaintervallumot építsünk a szokásos módon, ám most csupán benchmarkként fogjuk azt felhasználni.

A Monte-Carlo integráláshoz kötődő fontos témakör a variancia csökkentése (variance reduction). Amint az ismert, egy θ_1 becslőfüggvényt hatásosabbnak nevezünk θ_2-nél, ha $\frac{Var(\theta_1)}{Var(\theta_2)} < 1$ és mindkét becslőfüggvény torzítatlan módon becsli θ-t. Ebben az esetben θ_1 használata θ_2 helyett a varianciában

$$\frac{Var(\theta_1) - Var(\theta_2)}{Var(\theta_1)} \times 100 \tag{4.17}$$

százalékos csökkenést eredményez.

A variancia csökkentésére a következőkben négy fontos eljárást mutatok be, az ellentétes (antitetikus) változók (antithetic variables), az ellenőrző változók (control variables), a fontossági mintavételezés (importance sampling) és a rétegző mintavételezés (stratified sampling) módszereit. A módszerek alapötlete nem ismeretlen a statisztikában járatos Olvasó számára: az ellenőrző változók módszere regressziós, a fontossági mintavételezés a hányadosbecsléses, a rétegző mintavételezés a rétegzett mintavétel (Galambosné, 2011) logikáját alkalmazza a variancia csökkentésére.

Antitetikus változók módszere

Tekintsük két azonos eloszlású, X_1 és X_2 valószínűségi változó átlagát. Az átlag varianciája:

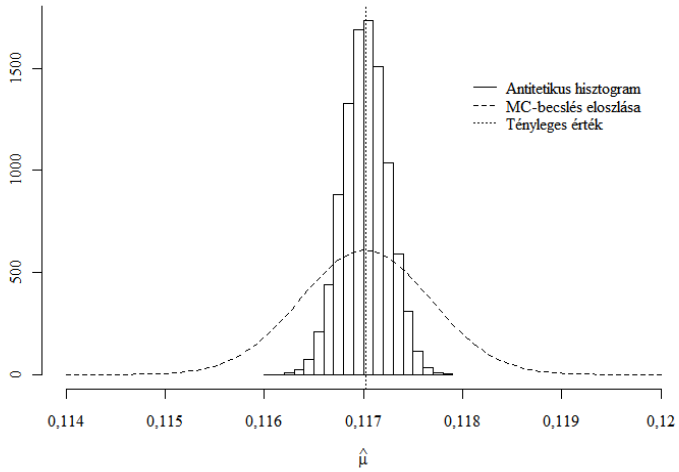
$$Var\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}\left(Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)\right) \tag{4.18}$$

Amennyiben X_1 és X_2 függetlenek, úgy a kovarianciatag (4.18)-ben 0. Ha tehát olyan változókat használunk, ahol a kovariancia negatív, az átlag varianciája csökkenthető a független esethez viszonyítva. Ez az alapvető ötlet húzódik az antitetikus változók módszere mögött. A Monte-Carlo szimulációk esetén legyen a keresett integrál becslése a $[0,1]$ egyenletes véletlen változók valamilyen függvénye: $X_1 = h(U_1, U_2, \ldots, U_n)$. Tekintsük $X_2 = h(1-U_1, 1-U_2, \ldots, 1-U_n)$ antitetikus becslést, ekkor a két valószínűségi változó eloszlása megegyezik. Páronként a véletlen változók

közötti kovariancia negatív, értéke $-\frac{1}{12}$. Ekkor bizonyítható (pl. Rizzo, 2008, p. 129), hogy bármely monoton h függvényre

$$Cov\left(h\left(U_1,U_2,\ldots,U_n\right),h\left(1-U_1,1-U_2,\ldots,1-U_n\right)\right)\le 0$$

A módszer gyakorlati alkalmazása egyszerű. Generáljunk $n/2$ mintaelemet a szükséges egyenletes eloszlásból, majd ezekből további $n/2$ ellentétes változót. A negatív kovariancia miatt az ily módon előállított becslés varianciája alacsonyabb lesz, mint a hagyományos, n darab véletlen számból álló MC becslésé. Az előző alfejezetben a $\int_2^4 e^{-x}dx$ integrál becsléséhez 10 000 véletlen számot használtunk fel amelyek, $U_1 \sim Unif\left(2,4\right)$ eloszlásúak voltak. Amennyiben 5 000 darab véletlen számot generálunk U_1-ből, majd az $U_2 = 6-U_1 \sim Unif\left(2,4\right)$ véletlen értékeket párosítjuk hozzájuk, a becslés varianciája alacsonyabb lesz. A 20 000 alkalommal elvégzett becslés eredményeinek empirikus eloszlását a 4-4. ábra mutatja be hisztogramon, feltüntetve a 4-3. ábrán bemutatott (eredeti MC-) becslés elméleti eloszlását is. Az összehasonlíthatóság érdekében a korábbi ábra vízszintes tengelyét (0,114–0,120) megtartottam. Az antitetikus módszerrel nyert becslések tehát jóval kevésbé szóródnak a tényleges érték körül. A további eljárások esetén a hasonló ábrát nem, csak a variancia csökkenésének mértékét mutatom be.



4-4. ábra: Az antitetikus változók használatának hatása az MC-becslés eloszlására

A (4.17) arány alapján könnyedén meghatározhatjuk a módszer hatékonyságjavulását, ami 88% körüli értéket mutat, azaz antitetikus változók használatával a becslésünk varianciáját nagymértékben sikerült csökkenteni, úgy, hogy a számítási költség nem lett nagyobb.

Kontrollváltozók módszere

Tegyük fel, hogy a célunk $\int_a^b h(x)\,dx$ becslése, és van egy olyan l függvénnyel leírható (kontroll-) változó, mely $\eta = E\left[l(X)\right]$ várható értékét ismerjük, és a két változó korrelál. A két függvényből konstruálható olyan becslőfüggvény, mely torzítatlan[16] bármely c konstans esetén:

$$\hat{\mu}_{cont} = h(X) + c\left(l(X) - \eta\right). \tag{4.19}$$

Ekkor (4.19) varianciája felírható, és célunk c függvényében ennek legkisebb értékét megtalálni:

$$Var\left(\hat{\mu}_{cont}\right) = c^2 Var\left[l(X)\right] + 2cCov\left(h(X), l(X)\right) + Var\left[h(X)\right]. \tag{4.20}$$

A (4.20) összefüggés c-ben másodfokú és konkáv, így minimális értékét a

$$c^* = -\frac{Cov\left(h(X), l(X)\right)}{Var\left[l(X)\right]} \tag{4.21}$$

helyen veszi fel, ahol a variancia értéke

$$Var\left(\hat{\mu}_{cont}\left(c^*\right)\right) = Var\left[h(X)\right] - \frac{\left[Cov\left(h(X), l(X)\right)\right]^2}{Var\left[l(X)\right]}. \tag{4.22}$$

Láthatjuk, hogy a kivonandó taggal csökken a variancia értéke. Ezt tudva kiszámíthatjuk (4.17) alapján a variancia százalékos csökkenését:

$$\frac{Var(\theta_1) - Var(\theta_2)}{Var(\theta_1)} = \frac{\left[Cov\left(h(X), l(X)\right)\right]^2}{Var\left[h(X)\right] Var\left[l(X)\right]} = \left[Corr\left(h(X), l(X)\right)\right]^2. \tag{4.23}$$

---

[16] $E\left(\hat{\mu}_{cont}\right) = E\left[h(X) + c\left(l(X) - \eta\right)\right] = \mu + c \times 0 = \mu.$

A fentiekből egyértelműen látszik, hogy olyan $l(.)$ függvényre van szükségünk, hogy $l(X)$ erősen korrelál $h(X)$-szel. Amennyiben a valószínűségi változók között nincs korreláció, a módszer nem használható, más kontrollváltozót kell keresnünk. A feladat tehát a helyes változó megtalálása, majd az optimális c^* kiszámítása, melyhez (4.21) szerint a varianciára és a kovarianciára van szükség. Amennyiben ezek az értékek analitikusan nem meghatározhatók, szimuláció segítségével találhatjuk meg a megfelelő értékeket.

Tekintsük újra az $\int_2^4 e^{-x}dx$ integrált. Keressük azt a $l(X)$ függvényt, amely momentumait könnyen meg tudjuk határozni és erős korrelációt mutat $h(X)$-szel. A legegyszerűbb választás egy lineáris függvény, $l(X)=\frac{X-2}{2}$. Ekkor $Cov\left(h(X),l(X)\right)=-\frac{e^{-4}}{2}$, $Var\left[l(X)\right]=\frac{1}{12}$, azaz (4.21) alapján $c^*=6e^{-4}$. A variancia értéke itt (4.22) és a már korábban levezetett $h(X)$ momentumai segítségével:

$$Var\left(h(X)+c\left(l(X)-\eta\right)\right)=\frac{e^2-1}{2e^8}-3e^{-8}\approx 0,0000653,$$

ami azt jelenti, hogy a variancia csökkenése (4.23) alapján közel 94%-os:

$$\frac{Var(\theta_1)-Var(\theta_2)}{Var(\theta_1)}=\frac{6}{e^2-1}\approx 0,939$$

Célunk közvetlenül nem csupán $\mu=E\left[h(X)\right]$, hanem az integrál közelítése volt, de a variancia elért csökkenése természetesen az integrál becslésében is hasonlóan jelentkezik. A kontrollváltozó használatával, azaz $h(X)=e^{-X}$ helyett a szintén torzítatlan $h(X)+c\left(l(X)-\eta\right)=e^{-X}+6e^{-4}\left(\frac{X-2}{2}-\frac{1}{2}\right)$ becslőfüggvényt alkalmazva az integrál becslésének varianciáját jelentősen sikerült csökkenteni.

Az antitetikus változók módszere a kontrollváltozó módszer speciális esete, ahol mindkét becslőfüggvény független azonos eloszlású, és a változópárok közötti korrelá-

ció -1, ekkor $c^* = \frac{1}{2}$ optimális érték adódik. Annak ellenére, hogy az antitetikus változók módszere speciális esetként is felfogható, az irodalomban külön tárgyalják őket.

Gyakran alkalmazott technika több kontrollváltozó felhasználása, hiszen $\hat{\mu} = h(X) + \sum_j c_j^* \left(l_j(X) - \mu_j \right)$ szintén torzítatlan becslést ad. Az optimális $\mathbf{c}^* = \left(c_j^* \right)$ vektort a l_j és az h függvények közötti maximális korrelációval érhetjük el. Gyakran alkalmazott módszer, hogy az optimális $\mathbf{c}^*$ meghatározásához egyszerű lineáris regressziót illesztünk, amiből a kontrollváltozós módszer legfontosabb jellemzőit azonnal megkapjuk. A varianciában bekövetkező csökkenés pontosan a lineáris regresszió R^2 értékével egyezik meg, a regressziós paraméterekből pedig $\mathbf{c}^*$ adódik.

A fentiekben egyenletes eloszlású valószínűségi változó felhasználásával közelítettünk függvények várható értékén keresztül integrálokat. A következő alfejezetben az egyenletes véletlen számoknál hatékonyabb módszert ismerünk meg. Ez a módszer egyben a leginkább elterjedt MC integrál alkalmazás.

Fontossági mintavétel

A bemutatott klasszikus MC-módszer és a variancia csökkentésére irányuló eljárások hátránya, hogy nem használhatók közvetlenül olyan esetekben, amikor valamely integrálási határ nem véges, ráadásul egyenletes véletlen számok alkalmazása nem hatékony, ha $h(.)$ nagyon távol esik az egyenletestől. Mivel az integrálást visszavezettük egy átlagolási problémára, általánosíthatjuk megközelítésünket a súlyozatlan átlag helyett súlyozott átlag (azaz az egyenletestől eltérő sűrűségek) alkalmazásával. Az általános módszer neve fontossági mintavétel (importance sampling).

Tekintsük az X véletlen változót g sűrűségfüggvénnyel, ahol bármely x esetén, melyre $h(x) > 0$, szükségképpen $g(x) > 0$. Legyen továbbá $\tilde{h}(x) = h(x)$, ha $a \le x \le b$, ezen kívül $\tilde{h}(x) = 0$, valamint Y véletlen változó $\frac{h(X)}{g(X)}$. Ekkor

$$\int_a^b h(x)\,dx = \int_a^b \frac{h(x)}{g(x)} g(x)\,dx = \int_{-\infty}^{\infty} \frac{\tilde{h}(x)}{g(x)} g(x)\,dx = E\left(\frac{\tilde{h}(X)}{g(X)} \right) = E(Y) \qquad (4.24)$$

Közelítsük $E(Y)$ értékét hagyományos Monte-Carlo-integrálással, azaz számítsuk ki az

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{h}(X_i)}{g(X_i)} \tag{4.25}$$

átlagot, ahol az X_i-k a $g(x)$ sűrűségfüggvényből származó véletlen értékek. A $g(x)$ függvény neve fontossági függvény (importance function). A közelítés varianciáját n és $Var(Y)$ határozza meg, ezért a gyakorlatban célunk, hogy a fontossági függvény $h(x)$-hez hasonló, a hányados közelítőleg konstans legyen. Hasonlóan fontos szempont, hogy $g(x)$ alapján X könnyen szimulálható legyen.

Bizonyítható (*Rizzo* [2008] 143. old.), hogy a variancia minimalizálása a

$$f^*(x) = \frac{|h(x)|}{\int_A |h(x)|\,dx}$$

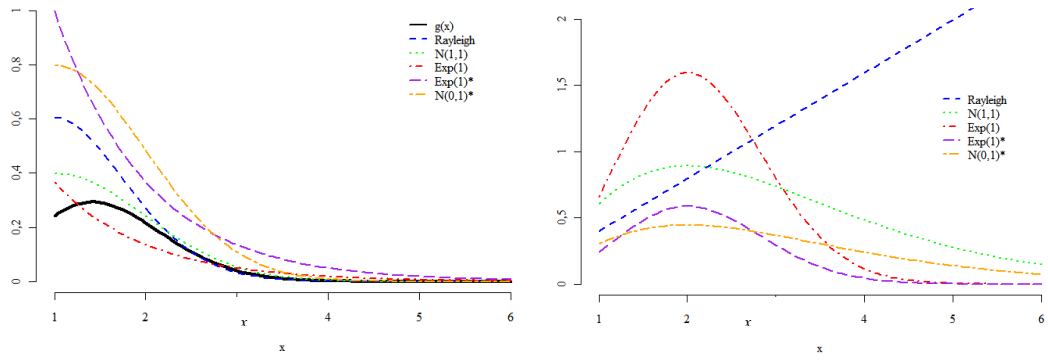
fontossági függvény alkalmazásával érhető el, ahol $A \in \square$ az a halmaz, ahol integrálni kívánunk. Mivel valószínűtlen, hogy ez a kifejezés rendelkezésre áll, gyakorlati probléma esetén leggyakrabban olyan függvényt választunk, amely elégségesen közel van $|h(x)|$-hez A-n.

A fontossági mintavételt a korábbiaktól eltérő $h(x) = \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}}$ függvény $(1,\infty)$ intervallumon vett integrálja segítségével mutatom be, méghozzá öt fontossági függvény felhasználásával, azok hatékonyságát összehasonlítva.

1. Standard Rayleigh-eloszlás: $g_1(x) = xe^{-\frac{x^2}{2}} \quad x \geq 0.$
2. Normális eloszlás $N(1,1)$: $g_2(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}}$.
3. Exponenciális eloszlás $Exp(1)$: $g_3(x) = e^{-x},\ x \geq 0.$
4. Módosított exponenciális eloszlás $Exp(1)^*$: $g_4(x) = e^{-x+1},\ x \geq 1.$

5. Módosított standard normális eloszlás: $N(0,1)^*$: $g_5(x) = = 2\times\phi(x-1),\ x \geq 1.$

A választott eloszlások egy része nem kizárólag az $x \geq 1$ tartományon értelmezett. A módosított sűrűségfüggvényeket olyan módon alakítottam ki, hogy belőlük könnyű legyen véletlen értékeket nyerni és hasonlítsanak a céleloszlásra (mind alakra, mind értelmezési tartományra). A Rayleigh-eloszlásból az inverz eloszlásfüggvény módszerrel, a többi eloszlásból pedig az R beépített függvényei segítségével vettem mintát. A 4-5. ábra $h(x)$ integrálandó szakaszán mutatja be a függvényeket (bal oldal), valamint az $\frac{h(x)}{g_j(x)}$ hányadosokat (jobb oldal). Ahogy említettem, egy fontossági függvény akkor megfelelő, ha pontosan, vagy legalább közel azon az intervallumon értelmezett, ahol az integrálandó függvényt integrálnunk kell, valamint elég közel van a két függvény egymáshoz, azaz a hányados hozzávetőleg konstans.



4-5. ábra: A fontossági mintavétel sűrűségfüggvényei és a függvények hányadosai

A 4-5. ábra a módosított standard normális eloszlás tűnik a legjobb választásnak. Ez a fontossági függvény a módosításnak köszönhetően csak az $x \geq 1$ helyeken értelmezett, míg például g_2 a teljes x tengelyen, g_1 és g_3 pedig a pozitív félegyenesen. Mindez azt jelenti, hogy az g_1, g_2, g_3 függvények esetén a $\tilde{h}/g_j$ hányados sok esetben zérus, az eljárás nem hatékony.

A mintavételek[17] 2000 alkalommal történt elvégzése után az eredményeket a 4-4. táblázatban foglaltam össze $(n = 10\,000)$.

4-4. táblázat: Az integrál közelítésének eredményei öt fontossági függvénnyel

Becslés jellemzői	1.	2.	3.	4.	5.
Az integrál becsült értékeinek átlaga	0,40076	0,40059	0,40026	0,40065	0,40063
Az integrál becsült értékeinek szórása	0,00357	0,00412	0,00584	0,00155	0,00044
Nullák átlagos aránya (százalék)	39,3	50,0	63,25	0,00	0,00

Az öt fontossági függvény közül a módosított normális eloszlással készült közelítés rendelkezik a legkisebb varianciával. Az első három jelölt esetén az integrálás határaitól eltérő értelmezési tartomány miatt a generált véletlen értékek döntő többsége nem hasznosul (a táblázatban nullák aránya), mivel az $\frac{\tilde{h}(x)}{g(x)}$ hányados zérus értékű. Az $Exp(1)$ eloszlás sűrűségének 63,2 százaléka esik a $[0,1]$ intervallumba, az $N(1,1)$ eloszlás esetén 50 százalék, a függvényforma tekintetében hasonló Rayleigh-eloszlásnál pedig a $[0,1]$ közötti integrál alapján $1-e^{-\frac{1}{2}} \approx 0,3935$ a felesleges húzások aránya. Ez utóbbi választás további hátránya, hogy az $\frac{h(x)}{g(x)}$ egy pozitív meredekségű lineáris egyenes, azaz nem stabil. A transzformált eloszlások jobban teljesítettek ebben az esetben, általánosságban hátrányuk, hogy a véletlenszám-generálás nem minden esetben triviális.

Rétegző mintavétel

[17] Az ábrák elkészítéséhez és az integrálok, valamint jellemzőik számításához használható programot a Függelék tartalmazza. Egyéb fontossági függvények a kód alapján könnyedén implementálhatók.

A rétegző mintavétel[18] a variancia csökkenését úgy éri el, hogy az integrálandó területet rétegekre bonja, és ezeken a rétegeken belül kis varianciával próbál becsülni. A k darab rétegben rögzített számú véletlen értéket húzunk, úgy, hogy $n = n_1 + n_2 + \ldots + n_k$, azzal a céllal, hogy

$$Var\left(\hat{\mu}_1(n_1) + \hat{\mu}_2(n_2) + \ldots + \hat{\mu}_k(n_k)\right) << Var\left(\hat{\mu}(n)\right)$$

ahol a bal oldalon a rétegző mintavételt alkalmazó, a jobb oldalon pedig a standard MC-becslőfüggvény látható.

A variancia csökkentése akkor hatékony, ha a rétegekben az integrálandó függvény átlaga jelentősen eltérő, azaz sikerül heterogén rétegeket kialakítani. Amennyiben az integrálandó függvény monoton, ezt könnyű teljesíteni. Jól érzékelhető a hagyományos rétegzett mintavétellel való analógia abból a tényből adódóan is, hogy a rétegző mintavétel mindig kisebb varianciát szolgáltat, kivéve abban az esetben, ha a rétegek átlagai megegyeznek.

A már ismerős $\int_2^4 e^{-x}dx$ példán mindössze két egyenlő hosszúságú réteg alkalmazásával a becslés varianciája kevesebb mint harmadára, négy réteg esetén kevesebb mint 10 százalékára esik azonos mintaelemszám mellett. Ennek eléréséhez csupán annyi a feladatunk, hogy $Unif(2,4)$ véletlen számok helyett például $Unif(2,3)$ és $Unif(3,4)$ véletlen értékek segítségével becsüljük a megfelelő területeket, majd becsléseinket összegezzük.

A rétegző mintavétel előnye, hogy tetszőlegesen kombinálható a többi varianciacsökkentő eljárással, így a szakirodalom ismeri és használja a rétegző fontossági mintavétel (stratified importance sampling) fogalmát is, ahol a különböző rétegekhez különböző fontossági függvények definiálhatók tovább javítva ezzel a hatékonyságot.

Az előző alfejezetekben olyan integrálási eljárásokat mutattam be, melyek egy része determinisztikus, pontosságuk csupán a választott módszertől és az intervallum fel-

[18] A rétegző mintavétel angol terminológiával stratified sampling, azaz az elnevezés megegyezik a rétegzett mintavétel elnevezéssel. A könnyebb megkülönböztethetőség érdekében nevezem rétegző mintavételnek az eljárást.

osztásának finomságától függ. A módszerek egy másik típusa egyenletes eloszlású véletlen értékek generálását igényelte, majd bemutattam a közelítés varianciáját csökkenteni képes algoritmusok közül a legfontosabbakat. A véletlen értékek egyenletes eloszlását feloldva megismerkedtünk a fontossági mintavétellel, ahol a fontossági eloszlás kiválasztása néha nem egyszerű feladat, több szempont egyidejű megfontolását igényli.

4.7.4. Markov-lánc Monte-Carlo módszerek

A Markov-lánc Monte-Carlo (MCMC) módszerek célja, hogy mintát tudjunk venni egy (jellemzően összetett, többdimenziós) valószínűség eloszlásból. Az MCMC technikák közös jellemzője, hogy olyan Markov-láncot állítanak fel, melyek egyensúlyi eloszlása megegyezik a kívánt eloszlással. Ezután minden lépés utáni állapotot a cél eloszlásból származó mintának tekintünk. Jellemzően a Markov-lánc megkonstruálása nem okoz különösebb nehézséget, a gyakorlati alkalmazások esetén a probléma annak meghatározása, hogy a lánc konvergál-e, illetve mikor konvergál az egyensúlyi állapothoz egy adott hibahatáron belül.

Elsőként röviden tekintsük át a Markov-láncok azon jellemzőit, melyek az MCMC módszerek szempontjából jelentőséggel bírnak. A Markov-lánc egy $\{X_t\}$ dinamikus sztochasztikus folyamat $t \geq 0$ indexszel. Az MCMC technikákhoz diszkrét Markov-láncok generálására van szükség, ahol t nemnegatív egész értékeket vesz fel. A megalkotott lánc tehát X_0 indulási érték után $X_1, X_2, \ldots, X_t, \ldots$ állapotokba kerül. Ez a sorozat Markov-lánc, ha

$$P\left(X_{t+1} = a \middle| \mathbf{X} = \mathbf{b}\right) = P\left(X_{t+1} = a \middle| X_t = b_t\right)$$

minden $(a, \mathbf{b})$ párra és $t \geq 0$-ra, ahol $\mathbf{X} = \begin{pmatrix} X_0 & X_1 & \ldots & X_t \end{pmatrix}$ a megelőző és a jelen állapotok vektora és $\mathbf{b} = \begin{pmatrix} b_0 & b_1 & \ldots & b_t \end{pmatrix}$. Mindez azt jelenti, hogy a következő állapot alakulása csak a jelenlegi állapoton múlik, a múltbeli állapotok nem befolyásolják azt. Az összes lehetséges kimenetel halmazát állapottérnek hívjuk. Amennyiben az állapottér véges halmaz, az állapotok közötti átmenetek valószínűségeit leírhatjuk egy $\mathbf{P}$ átmenetmátrix segítségével, ahol a p_{ij} elem annak a valószínűsége, hogy i állapotból j állapotba kerül a lánc egy lépésen belül. Az ún. Chapman-Kolmogorov azonosságok alapján annak a valószínűsége, hogy mindez k lépésben történik meg, az átmenetmátrix

megfelelő $\mathbf{P}^k$ hatványából olvasható le. Egy Markov-láncot irreducibilisnek nevezünk, ha tetszőleges állapotból elérhető tetszőleges másik állapot, vagyis az állapotok kommunikálnak egymással. Egy i állapot rekurrens, ha a lánc visszatér az adott állapotba egy valószínűséggel, egyébként az állapotot tranziensnek, átmenetinek nevezzük. Amennyiben a visszatérés várható ideje véges, úgy az állapotot pozitív rekurrens. Amennyiben egy gráfon ábrázoljuk a láncot, az i állapotból induló és végződő utak hosszainak legnagyobb közös osztóját az állapot periódusának hívjuk. Egy lánc aperiodikus, ha minden hozzá tartozó állapot 1 periódusú. Az aperiodikus, irreducibilis Markov-láncokat ergodikusnak hívjuk. Ergodikus láncok esetén egyetlen egyensúlyi eloszlás létezik, mely minden j-re

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$$

és független a kezdeti i állapottól. Amennyiben a Markov-lánc egyensúlyi eloszlása f, úgy a lánc generálásával tulajdonképp f-ből származó mintavételt végzünk. Így a Monte-Carlo módszernél alkalmazott nagy számok törvényéhez hasonlóan igaz az

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \to E_f\left(h(X)\right) \tag{4.26}$$

határérték reláció. A Markov-láncok konvergenciájára nem térek ki részletesen, a következőkben felvázolt algoritmusok elméletileg csaknem minden esetben konvergens láncot hoznak létre. A gyakorlatban azonban a konvergencia néha nagyon lassú, egyes esetekben pedig úgy tűnik, hogy a Markov-lánc konvergál, azonban ez mégsem teljesül. A következőkben olyan MCMC algoritmusokat írok le röviden, melyek a legáltalánosabban alkalmazottak. Közös tulajdonságuk, hogy olyan Markov-láncot határoznak meg, melynek egyensúlyi eloszlása a kívánt f. Szerencsére ez nem túl nehéz feladat, léteznek olyan általános formulák, melyek – elméletileg – tetszőleges f esetén is alkalmazhatóak. Az elméletek bizonyításai, a konvergencia tulajdonságok megtalálhatók Robert és Casella (2004) átfogó művében.

4.7.5. Metropolis-Hastings algoritmusok

A Metropolis-Hastings algoritmusok névadója egyrészt Nicholas Metropolis (Metropolis et al., 1953) amerikai fizikus, másrészt Keith Hastings (1970) kanadai sta-

tisztikus[19]. Az algoritmus több fajtája létezik, ahogy azt be is mutatom rövidesen. A M-H algoritmusok történetéről ad összefoglalót Hitchcock (2003), aki beszámol a módszer kezdeti ismeretlenségének, majd a napjainkban megfigyelhető népszerűségének okairól. Az eredeti, Metropolis-féle megközelítés csupán egyetlen konkrét, fizikai probléma megoldására vonatkozott, így nem is vált ismertté a statisztikával foglalkozók körében. A Statisztikai Enciklopédia 1982-es több ezer oldalas kiadásában sem a Metropolis-Hastings algoritmus, sem a szerzők, sem az MCMC szócikk nem szerepeltek, a 2006-os második kiadás (Kotz et al., vol 7.) azonban már hét oldalban tárgyalja Luke Tierney tollából. A 60-as években kapott némi figyelmet Metropolis 1953-as cikke, de az alkalmazások továbbra sem ismerték fel, hogy a módszer nem csak a Tellerékéhez hasonló sokdimenziós integrálok kiszámítására használható, hanem tulajdonképp tetszőleges eloszlásból származó véletlen szám generálásra is. Hastings tanulmánya általánosította és továbbfejlesztette a módszert nem szimmetrikus instrumentális eloszlások esetére is. Hastings egyetlen doktorandusza, Peter Peskun (1973) bizonyította, hogy amit ma általános M-H algoritmusnak nevezünk, hatékonyabb, mint az egyéb hasonló eljárások. Ennek ellenére a 90-es évek elejéig továbbra sem alkalmazták a módszert széleskörűen. A 90-es években az ún. Gibbs mintavétel terjedt el, Geman és Geman (1984) munkája nyomán. A Gibbs mintavétel olyan sokdimenziós eloszlások esetén alkalmazható hatékonyan, melyek egyváltozós feltételes eloszlásai ismertek, vagy könnyen generálhatók. A Gibbs mintavétel tehát megelőzte az M-H algoritmus elterjedését, annak ellenére, hogy annak speciális esete (Gelman, 1992). Az igazi áttörést, a mainstreambe kerülést Chib és Greenberg (1995), valamint Tierney (1994) hozták meg, akik közérthetően, példákon keresztül mutatták be az algoritmusok alkalmazási lehetőségeit. Az algoritmus újabb lendületet adott a bayesi statisztika művelőinek, hisz a tetszőleges eloszlásokból

[19] A Metropolis-féle tanulmány magyar kötődése, hogy az öt szerző között szerepel Teller Ede, valamint felesége is. Rosenbluth visszaemlékezései szerint azonban sem Metropolis, sem Teller felesége, Mici nem dolgoztak a problémán, az alapprobléma felvetése Teller nevéhez fűződött, a megoldást Rosenbluth adta, a programozás pedig az ő feleségének munkája volt.

Hastings ezen cikkére több ezer hivatkozás született az elmúlt négy évtizedben, érdekes megemlíteni, hogy ezen a híres cikkén kívül életrajza csupán két további referált cikkét említ.

generálható véletlen értékek lehetőséget adtak a konjugált analíziseken túl egyéb a priori eloszlások, illetve összetett modellek alkalmazására is.

A következő alpontban a M-H algoritmus általános alakját, valamint néhány speciális esetét mutatom be. Mivel a Gibbs mintavétel külön néven vonult be a statisztika szakirodalmába, némileg az alkalmazási területe is más, ezért azt külön alfejezetben ismertetem.

A fő feladat tehát egy Markov-lánc létrehozása, melynek egyensúlyi eloszlása az adott cél eloszlás. Az algoritmusnak definiálnia kell, hogy egy adott X_t állapotból hogyan érjük el X_{t+1}-et. Valamennyi M-H algoritmusban szerepel egy $q(\cdot|X_t)$ instrumentális eloszlás (instrumental vagy proposal distribution), melyből egy jelölt Y véletlen számot generálunk. A jelöltről bizonyos szabály alapján eldöntjük, hogy elfogadjuk, vagy elutasítjuk. Amennyiben elfogadjuk, úgy $X_{t+1} = Y$, egyébként a lánc marad az eredeti állapotában, azaz $X_{t+1} = X_t$. Az instrumentális eloszlás könnyen generálható kell hogy legyen és függhet a kiinduló X_t értéktől. Így például a $q \sim N(X_t, \sigma^2)$ normális eloszlás egy klasszikus választás, valamilyen fix σ szórással. A szórás nagyságának szerepére a későbbiekben még visszatérünk. Amennyiben a lánc irreducibilis, pozitív rekurrens és aperiodikus, a lánc konvergálni fog a cél eloszláshoz. A gyakorlatban ez gyakran olyan instrumentális eloszlást jelent, mely értelmezési tartománya lefedi a cél eloszlását, illetve elegendően nagy a szórása ahhoz, hogy a cél eloszlás értelmezési tartományát képes legyen bejárni.

A klasszikus M-H algoritmus a következő módon képezi a Markov-láncot:

1. Válasszunk egy megfelelő $q(\cdot|X_t)$ eloszlást.
2. Generáljunk, vagy adjunk meg egy kezdő X_0 értéket.
3. Ismételjük az alábbi lépéseket, amíg a lánc bizonyos kritériumok alapján az egyensúlyi eloszlásához nem konvergál.
    a. Generáljunk jelöltet (Y) a $q(\cdot|X_t)$ eloszlásból.
    b. Generáljunk egy $U \sim Unif(0,1)$ véletlen változót.
    c. Ha

$$U \le \frac{f(Y)q(X_t|Y)}{f(X_t)q(Y|X_t)} \tag{4.27}$$

akkor $X_{t+1} = Y$ egyébként $X_{t+1} = X_t$.

d. Növeljük t értékét.

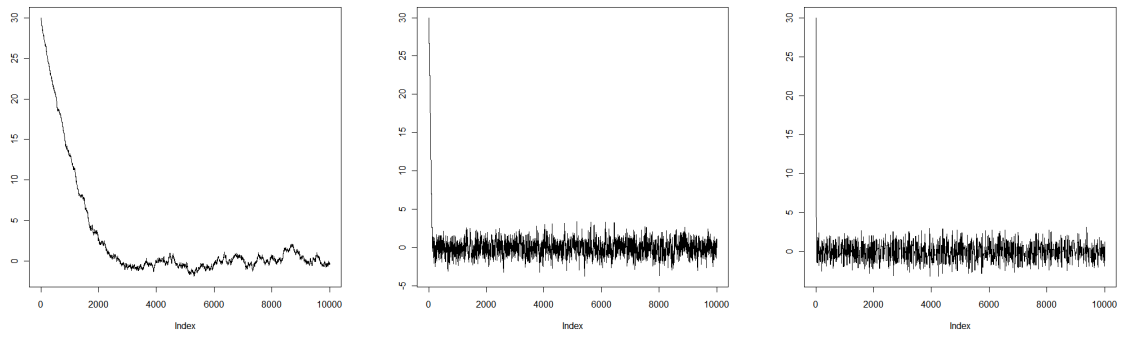
4. A konvergálás előtti értékeket (burn in period) levágva a lánc elejéről megkapjuk a kívánt eloszlásból származó véletlen mintát.

A fenti lépések szoftver segítségével egyszerűen és gyorsan megoldhatók. Minden egyes új Y jelölt elfogadásának esélye $\rho(X_t, Y) = \min\left\{\frac{f(Y)q(X_t|Y)}{f(X_t)q(Y|X_t)}, 1\right\}$. A (4.27), ún. Hastings hányados és az eljárás leírása alapján belátható két fontos jellemző. Mivel f a hányados számlálójában és nevezőjében is szerepel, a normalizáló konstans ismerete nem szükséges a módszer alkalmazásához. A másik fontos tulajdonság az, hogy a MCMC módszerrel nyert véletlen számok egymástól nem függetlenek, a szokásos angol jelöléssel az így nyert minta nem iid. A burn in periódus hosszának megválasztása általában önkényes, néhány ezer iteráció a legtöbb esetben elegendő hosszúnak bizonyul.

Az általános M-H módszer mellett meg kell említenünk néhány speciális esetet is, melyek jórészt a q függvény jellemzőiben térnek el. Amennyiben az instrumentális eloszlás szimmetrikus olyan értelemben, hogy $q(X_t|Y) = q(Y|X_t)$, úgy a (4.27)-ben szereplő hányadosból az instrumentális eloszlás kiesik, így az elfogadás valószínűségét csak f befolyásolja. Amennyiben a jelölt pontban a sűrűségfüggvény magasabb értéket vesz fel mint a lánc aktuális értéke, akkor biztos az elmozdulás. Ez a speciális eset az eredeti, Metropolis-féle mintavételezés. A Metropolis mintavétel egyik tipikus példája a véletlen bolyongás módszere (random walk Metropolis), ami esetén minden iterációban egy véletlen (pl. zérus várható értékű, fix szórású normális) növekedést/csökkenést generálunk, majd az így kapott jelölt pontról a fentiek szerint döntünk. A véletlen bolyongást használó módszer esetén nem könnyű feladat az instrumentális eloszlás szórását jól beállítani. Amennyiben a szórás túl alacsony, azaz a lépések túl kicsik, a lánc csak nagyon lassan konvergál, azaz a gyakorlatban használhatatlan. Amennyiben a szórás túl nagy, úgy az iterációk nagy részében elvetjük az új jelölt értéket, azaz a lánc nem hatékony, a tagok közötti korreláció pedig mindkét esetben magas lesz. A szimuláció elfo-

gadási rátájának azt a relatív értéket nevezzük, ahány alkalommal a jelölt értéket elfogadtuk az összes iteráció számához viszonyítva. Általános szabályként ez az arány a véletlen bolyongás algoritmusnál optimális esetben a $[0.15, 0.5]$ intervallumban található (Roberts et al., 1997).

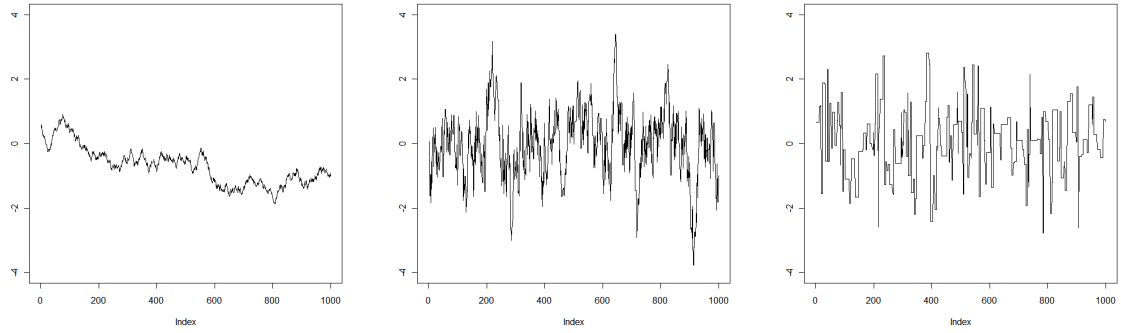
Amennyiben standard normális eloszlású véletlen értékeket szeretnénk generálni a random walk metropolis eljárás, és $q(y|x_t) \sim Unif(x_t - a, x_t + a)$ instrumentális eloszlás segítségével, úgy az eljárás minőségét a fentiek szerint nagyban befolyásolja az a paraméter. Az alábbi három ábrán három $(a = 0,1; a = 1; a = 10)$ eseteket hasonlítok össze. Gyakran alkalmazott technika, hogy a láncot (több) extrém értéktől indítjuk, ez kétmóduszú eloszlásoknál lehet kiemelten fontos. A láncot most $x_0 = 30$ értékről indítva, az x értékeket az indexszel szemben ábrázolva az alábbi, 4-6. ábrán látható mintázatot kapjuk 10 000 hosszúságú láncok generálása után:



$a = 0,1$ $a = 1$ $a = 10$

4-6. ábra: Véletlen bolyongás Metropolis láncok különböző paraméterekkel

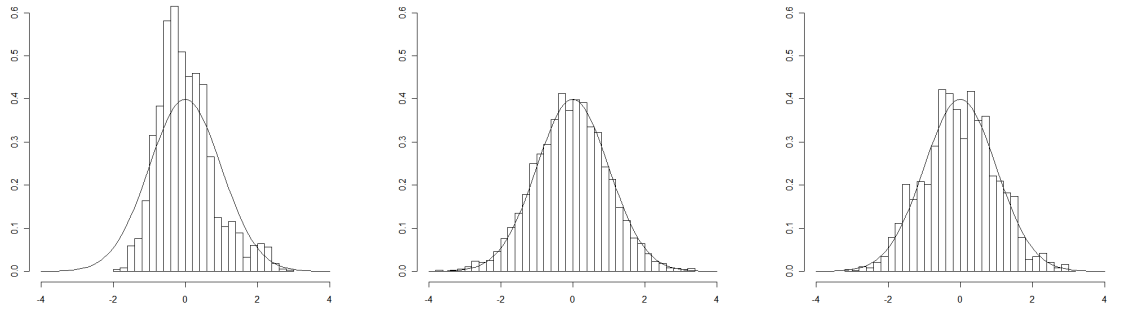
Az első lánc esetén a lépésköz túl kicsi, az extrém értéktől indulva nagyon lassan éri el a lánc az eloszlás „lényeges" részét. Ezután az eloszlás felfedezése szintén nagyon lassú. A második és harmadik esetben a lánc szinte azonnal megtalálja a felfedezendő területet, azonban a harmadik ábrán a nagy lépésköz miatt az ajánlott érték sok esetben elutasításra kerül. A 4-7. ábrán valamennyi láncból egy tipikus részt, a középső ezer lépést ábrázoltuk, ahol mindez jobban látszik.



$a = 0{,}1$ $a = 1$ $a = 10$

4-7. ábra: Véletlen bolyongás Metropolis láncok középső 1000 értéke

Az első ábrán azt látjuk, hogy a jelölt értékeket gyakorlatilag minden esetben elfogadjuk, de az 1 000 iteráció alatt alig sikerült az eloszlást bejárni, a függőleges tengelyen megfigyelhető, hogy az összes érték szűk, alig több mint két egység széles sávban szóródott. A harmadik esetben a sok függőleges vonal azt mutatja, hogy a jelölt érték sok esetben került elutasításra, azaz a $t+1$-edik érték megegyezik a t-edikkel. Lehetőség van az elfogadások, vagy az elutasítások számának rögzítésére, ezzel összehasonlítva a láncokat. Az egyes esetekben rendre 561, 1 987, 8 368 elutasítás történt a 10 000 iteráció során.



$a = 0{,}1$ $a = 1$ $a = 10$

4-8. ábra: Véletlen bolyongás Metropolis hisztogramok

Ebben az egyszerű példában lehetőségünk van a tényleges sűrűségfüggvény mellett ábrázolni a beégetési szakasz (egységesen az első 2 000 Markov-tagot távolítottuk el) levágása után megmaradó véletlen értékeket. Az első esetben jól láthatóan a konvergencia még nem történt meg, ehhez a lánc sokkal hosszabb futtatására lenne szükség, ami gyakorlati problémák esetén gyakorta túlságosan hosszú időt venne igénybe. A második, $a = 1$ esetben az eloszlást jól közelítő hisztogramot kapunk, míg a harmadik eset-

ben a túl gyakori elutasítás miatt vannak olyan osztályközök, ahol túlságosan nagy az esetek száma.

Egy másik jellemző, bár önmagában ritkábban alkalmazott speciális eset az ún. független mintavétel (idependence sampler). Az instrumentális eloszlás ebben az esetben nem függ a lánc előző értékétől, azaz $q\left(\cdot\middle| X_t\right) = q\left(\cdot\right)$. A (4.27) kifejezés jobb oldala tehát a $\frac{f\left(Y\right)q\left(X_t\right)}{f\left(X_t\right)q\left(Y\right)}$ alakot ölti. A független mintavétel akkor alkalmazható igazán hatékonyan, ha az instrumentális eloszlás elég jó közelítése a cél eloszlásnak. Szükséges azonban a hatékonyság érdekében, hogy a használt instrumentális eloszlás vastagabb farokkal rendelkezzen, mint a céleloszlás, ezért a gyakorlatban gyakran alkalmazott az alacsony szabadságfokú többváltozós t-eloszlás.

4.7.6. Gibbs mintavétel

A Gibbs mintavétel előnye az olyan sokdimenziós eloszlások esetén mutatkozik meg, amikor véletlen bolyongás M-H algoritmus építése közel lehetetlen. A Gibbs algoritmus lehetővé teszi a sokdimenziós problémák lebontását kisebb, akár egydimenziós feladatokra. Az egyszerűsített feladatok mennyisége azt okozhatja, hogy a konvergencia lassú, ám a megközelítés ettől függetlenül fontos szerepet tölt be az MCMC módszercsaládon belül.

Tegyük fel, hogy $\mathbf{X} = \left(X_1, X_2, \ldots, X_p\right)$ véletlen vektorváltozó, ahol az X_j-k egy, vagy többdimenziós komponensek, valamint azt is, hogy képesek vagyunk a megfelelő $f_1, f_2, \ldots, f_p$ sűrűségfüggvényekből szimulálni, azaz ismerjük a

$$f_j\left(x_j\middle| x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p\right)$$

ún. teljes feltételes (full conditional) eloszlásokat minden $j = 1, 2, \ldots, p$-re.

Ekkor a Gibbs algoritmus:

1. Válasszunk $\mathbf{X}_0 = \mathbf{x}_0$ kezdőértékeket.
2. Ismételjük az alábbi lépéseket, amíg a lánc bizonyos kritériumok alapján az egyensúlyi eloszlásához nem konvergál:

a. $X_1^{(t+1)} \sim f_1\left(x_1 \middle| x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)}\right)$

b. $X_2^{(t+1)} \sim f_2\left(x_2 \middle| x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}\right)$

c. $X_3^{(t+1)} \sim f_3\left(x_3 \middle| x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_p^{(t)}\right)$

d. …

e. $X_p^{(t+1)} \sim f_p\left(x_p \middle| x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)}\right)$

f. Növeljük t értékét.

3. A konvergencia előtti értékeket (burn in period) levágva a lánc elejéről megkapjuk a kívánt eloszlásból származó véletlen mintát.

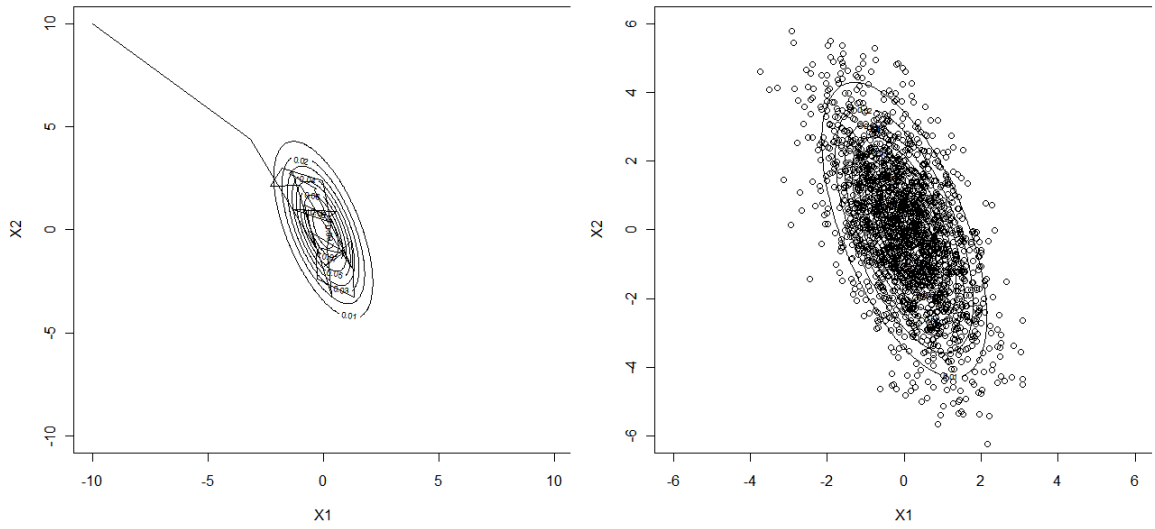
Az átláthatóság érdekében t felsőindexként szerepel, azaz $x_j^{(T)}$ a j. komponens T. lépésben felvett értékét jelöli.

A Gibbs mintavételezés legegyszerűbb esete és klasszikus példája a kétdimenziós normális eloszlásból való mintavétel. Legyen a várható értékek vektora $\boldsymbol{\mu} = (\mu_1, \mu_2)$, a varianciák σ_1^2 és σ_2^2, a korreláció pedig ρ. Ekkor a feltételes eloszlások egyváltozós normális eloszlások, melyekből könnyedén tudunk véletlen értékeket generálni:

$$f\left(x_1 \middle| x_2\right) \sim N\left(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}\left(x_2 - \mu_2\right), \left(1-\rho^2\right)\sigma_1^2\right)$$

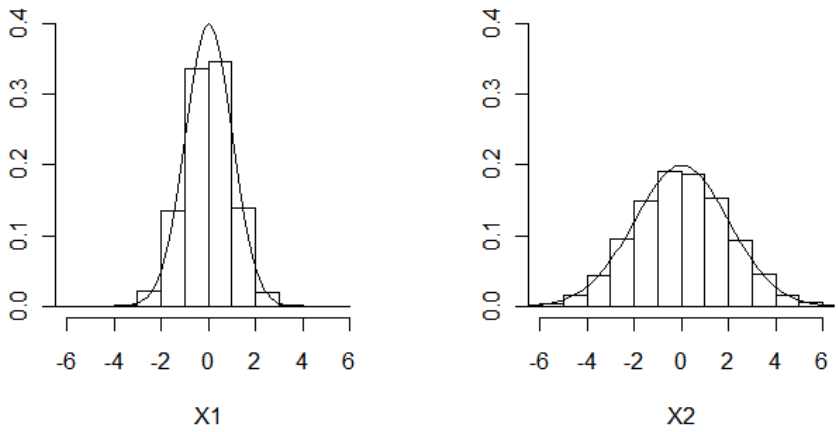
$$f\left(x_2 \middle| x_1\right) \sim N\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}\left(x_1 - \mu_1\right), \left(1-\rho^2\right)\sigma_2^2\right)$$

Példaként bemutatom a véletlen értékek generálását $\boldsymbol{\mu} = (0,0); \sigma_1 = 1; \sigma_2 = 2; \rho = -0,6$ paraméterekkel. Kezdő értékként $\mathbf{X}_0 = (-10,10)$ vektort adtam meg, ami távol esik a függvény tényleges helyétől. Az első 30 lépést és a tényleges sűrűségfüggvény kontúrját mutatja be a 4-9. ábra bal oldala. A jobb oldalon 2000 generált véletlen értékpár szerepel 1000 periódus levágása után (az átláthatóság érdekében).



4-9. ábra: Kétváltozós normális eloszlás Gibbs mintavétellel

A lánc tetszőleges lépésszámra futtatható, az alábbi összefoglaló számításokat 10 000 elem alapján készítettem. Az empirikus átlagok $\bar{\mathbf{x}} = (0{,}0089 \quad 0{,}0042)$, a szórások pedig rendre 1,0058 és 2,0080 értékeket vesznek fel, a korreláció pedig $r = -0.604$. A marginálisok hisztogramjai, illetve az elméleti sűrűségfüggvények láthatók a 4-10. ábrán.



4-10. ábra: Marginális eloszlások hisztogramjai

A fenti, kétdimenziós feladatot tehát két darab egydimenziós problémára vezethetjük vissza. Vegyük észre, hogy magasabb dimenziószám esetén az eljárás a fentiekkel analóg módon használható, ez programozás-technikailag sem okoz nehézséget. A Gibbs

mintavételezés esetén is érvényes, ami a M-H algoritmusok esetén: jellemzően nem a Markov-lánc felállítása okoz problémát, sokkal inkább a konvergencia ellenőrzése.

A fejezetben a bayesi statisztika és a szimulációs algoritmusok alapjait tekintettem át. Az előzetes ismereteinket egy prior sűrűségfüggvényben kell összefoglalnunk, majd azt az adatokat leíró likelihooddal kombinálva kapjuk a poszterior sűrűségfüggvényt. Ez a jellemzően sokdimenziós eloszlás nehezen értelmezhető, ezért általánosan bevett eljárás a marginális eloszlások szokásos statisztikai mutatókkal (átlagok, szórások, kvantilisek stb.) történő leírása. Sok esetben nem csak a poszteriorra, hanem annak valamely függvényére vagyunk kíváncsiak, ebben az esetben szinte bizonyosan szimulációt kell segítségül hívnunk. A különböző statisztikai mutatók, vagy akár a normalizáló konstans az esetek többségében integrálok meghatározását követelik meg.

A fejezet második felében a leggyakrabban alkalmazott szimulációs és integrálási módszereket mutattam be, azok előnyeivel és hátrányaival, alkalmazási lehetőségeivel együtt. Az algoritmusok leírása mellett egy-egy gyakorlati példa is szerepel, a szükséges programot pedig a Függelék minden esetben tartalmazza.

5. Mintaelemszám tervezése bayesi szemléletben

Equation Section (Next)A harmadik fejezet konklúziója az volt, hogy javasolt az előzetes ismeretek alapján egy feltétlenül szükséges nagyságú minta megkérdezésére, majd a minta alapján az elképzelésünket tovább pontosíthatjuk. A negyedik fejezetben az előzetes és a mintabeli információk összesítésének módszertanát, majd az így nyert poszterior leírását mutattam be. Jelen fejezetben a bemutatott eljárásokat alkalmazom a mintaelemszám tervezésére. Az egyszerűség és jobb áttekinthetőség kedvéért elsőként a kétkimenetelű kérdés példáját tekintem, amely azután könnyedén terjeszthető ki a kettőnél több kimenetelű válaszadási lehetőségre, korábbi szóhasználatunkkal a Likert-skálára. Bemutatom a konjugált analízis lépéseit, a poszterior összefoglalását, valamint a poszteriorból nyerhető, varianciára vonatkozó információ felhasználását a szükséges mintaelemszám meghatározására.

5.1. Mintaelemszám tervezése kétváltozós esetben

Amint az ismert, az aránybecslés standard hibájából kiinduló mintaelemszám meghatározás sarokköve a tényleges sokasági arány (lásd (3.2) egyenlet), ami gyakorlatilag egy Bernoulli valószínűségi változó θ paramétere.

Legyen $Y \sim Bernoulli(\theta)$. Amennyiben n elemű mintával rendelkezünk, a likelihood függvény az alábbi $(0<\theta<1)$:

$$L(\theta) = p(y|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \qquad 0 \le \theta \le 1 \tag{5.1}$$

ahol k a kedvező kimenetelek (vagy rövidebben a sikerek) száma a mintában. Ahogy azt a 4-2. táblázat alapján tudjuk, a konjugált prior eloszlás ebben az esetben a béta eloszlás, de elég csupán a hasonlóságot észrevennünk (5.1) likelihood és a béta

eloszlás sűrűségfüggvénye között. Számítsuk ki a poszterior eloszlást a konjugált prior $Beta\left(\underline{\alpha},\underline{\beta}\right)$ mellett (4.2) segítségével[20].

A számláló a likelihood és a prior szorzatával arányos, azaz:

$$p\left(y\middle|\theta\right)p\left(\theta\right)\propto\theta^{k}\left(1-\theta\right)^{n-k}\theta^{\underline{\alpha}-1}\left(1-\theta\right)^{\underline{\beta}-1}=\theta^{\underline{\alpha}+k-1}\left(1-\theta\right)^{\underline{\beta}+n-k-1}$$

ami már önmagában megadja a poszterior eloszlás sűrűségfüggvényének alakját. Ahhoz, hogy a poszterior eloszlás tényleges sűrűségfüggvényét megkapjuk, szükségünk van a nevezőre, ami (4.4) alapján:

$$p\left(y\right)=\int_{0}^{1}\frac{1}{\mathrm{B}\left(\underline{\alpha},\underline{\beta}\right)}\theta^{\underline{\alpha}+k-1}\left(1-\theta\right)^{\underline{\beta}+n-k-1}d\theta=$$

$$=\frac{\mathrm{B}\left(\underline{\alpha}+k,\underline{\beta}+n-k\right)}{\mathrm{B}\left(\underline{\alpha},\underline{\beta}\right)}\int_{0}^{1}\frac{1}{\mathrm{B}\left(\underline{\alpha}+k,\underline{\beta}+n-k\right)}\theta^{\underline{\alpha}+k-1}\left(1-\theta\right)^{\underline{\beta}+n-k-1}d\theta=\frac{\mathrm{B}\left(\underline{\alpha}+k,\underline{\beta}+n-k\right)}{\mathrm{B}\left(\underline{\alpha},\underline{\beta}\right)}$$

Az utolsó lépésben azt használtuk ki, hogy az integrál $Beta\left(\underline{\alpha}+k,\underline{\beta}+n-k\right)$ eloszlássá formálható a normalizáló konstans átalakításával, így az integrálást elvégezve egyet kapunk eredményül. Ez nem minden esetben ilyen egyszerű, hisz ez csupán a konjugált prior miatt tehető meg. Egyéb esetben az integrálást el kell végezni, ami az esetek döntő többségében numerikus integrálást jelent.

Az utóbbi két eredmény alapján a poszterior eloszlás tehát már könnyedén felírható:

$$p\left(\theta\middle|y\right)=\frac{\frac{1}{\mathrm{B}\left(\underline{\alpha},\underline{\beta}\right)}\theta^{\underline{\alpha}+k-1}\left(1-\theta\right)^{\underline{\beta}+n-k-1}}{\frac{\mathrm{B}\left(\underline{\alpha}+k,\underline{\beta}+n-k\right)}{\mathrm{B}\left(\underline{\alpha},\underline{\beta}\right)}}=\frac{1}{\mathrm{B}\left(\underline{\alpha}+k,\underline{\beta}+n-k\right)}\theta^{\underline{\alpha}+k-1}\left(1-\theta\right)^{\underline{\beta}+n-k-1}$$

Újra észre kell vennünk, hogy egy béta eloszlással van dolgunk, azaz egyúttal beláttuk, hogy a béta eloszlás a Bernoulli és a binomiális eloszlás konjugált priorja. Ekkor

[20] A bayesi irodalomban bevett szokás a prior eloszlás paramétereit alulvonással, a poszteriorét pedig felülvonással megkülönböztetni.

$\bar{\alpha} = \underline{\alpha} + k$ és $\bar{\beta} = \underline{\beta} + n - k$ helyettesítésekkel élve kapjuk a Bernoulli eloszlás poszterior eloszlását konjugált priort alkalmazva:

$$p(\theta|y) = \frac{1}{\mathrm{B}(\bar{\alpha}, \bar{\beta})} \theta^{\bar{\alpha}-1} (1-\theta)^{\bar{\beta}-1} \propto Beta(\bar{\alpha}, \bar{\beta}) \tag{5.2}$$

A konjugált prior eloszlás választása kényelmes és gyakori megoldás. Az eloszláscsalád kiválasztása azonban még nem jelenti a konkrét paraméterek megválasztását. Gyakori választás a nem informatív $\underline{\alpha} = \underline{\beta} = 1$ béta eloszlás, amely megfelel az egyenletes eloszlásnak, azaz az a priori információnk csupán annyi, hogy a paraméter valahol 0 és 1 között helyezkedik el. Egy másik gyakori választás a Jeffreys-féle prior. Az $(1\times 1 \text{ méretű})$ információs mátrix $\frac{n}{\theta(1-\theta)}$ formát ölt[21], amiből a $\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$ mag (kernel) adódik, ami pedig a $Beta(0{,}5, 0{,}5)$ eloszlásnak felel meg. A Jeffreys-féle prior tehát proper és némileg informatív, mégpedig az extrém valószínűségekre helyez relatíve nagyobb súlyt.

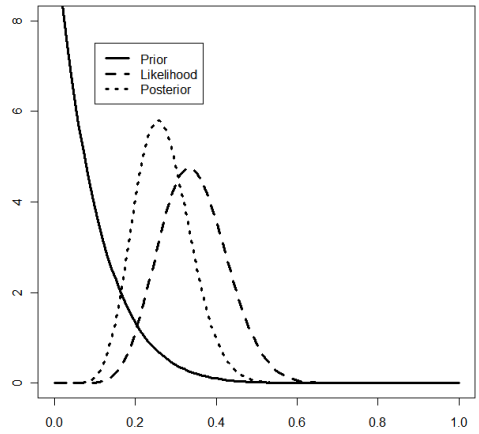
Amennyiben informatív priort alkalmazunk, úgy a prior béta eloszlás alkalmazandó paramétereinek meghatározása nem egyértelmű. A gyakorlatban általában az átlagra és/vagy kvantilisekre, vagy a szórásra adott becslésből számíthatjuk vissza az implicit paramétereket.

Az alábbi, 5-1. ábrán néhány – erősen eltérő – prior eloszlást feltételezve mutatjuk be az előálló poszterior eloszlást. A függvények azonos feltételezett minta mellett készültek, ahol $n = 30$ és $k = 10$. Az ábra a) pontja olyan priort tételez fel, ahol az alacsony értékek kapnak magas súlyt, míg b) esetben a magas értékek. A c) rész a nem informatív prior, feltételezése szerint minden lehetséges arány egyenlő valószínűséggel fordulhat elő. A Jeffreys-féle prior esetét mutatja be a d) ábra, amely az extrémumokra enyhén magasabb súlyt helyez, mint az egyéb értékekre.

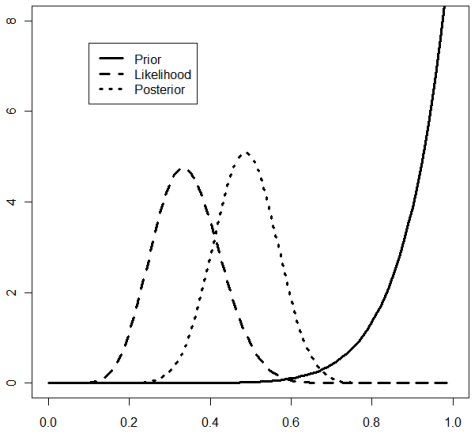
A likelihood természetesen mind a négy esetben megegyezik, hisz az az adatok hozzájárulását testesíti meg. A poszterior a nem informatív esetekben eltér a prior elosz-

[21] Bizonyítást lásd a Függelékben.

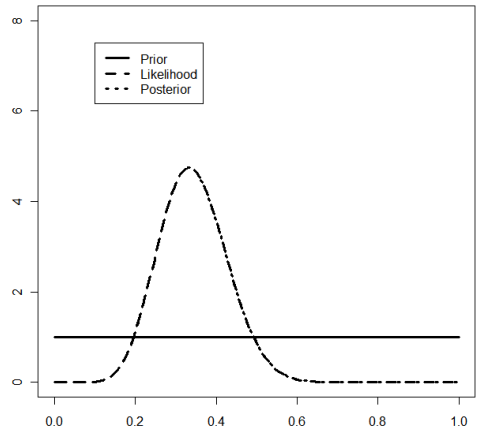
lástól, az elmozdulás mértéke a b) ábrán a legnagyobb, hiszen a likelihood és a prior igen messze esnek egymástól. A mintaelemszám már elég nagy ahhoz, hogy az elmozdulás nagy lehessen.



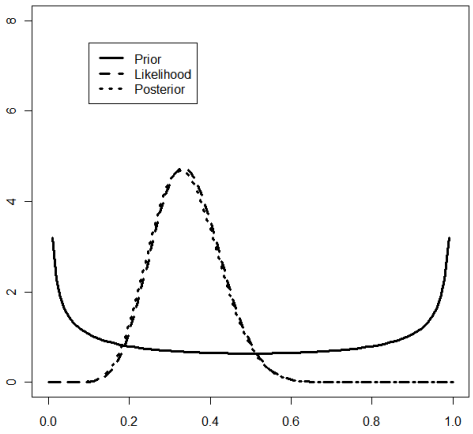
a) $p(\theta) \sim Beta(1,10)$



b) $p(\theta) \sim Beta(10,1)$



c) $p(\theta) \sim Beta(1,1)$



d) $p(\theta) \sim Beta(0{,}5,0{,}5)$

5-1. ábra: Beta poszterior néhány különböző prior esetén

A nem informatív, egyenletes eloszlás esetén a poszterior megegyezik a likelihooddal, míg a Jeffreys-féle prior esetén a különbség minimális. A tetszőleges binomiális mintára (n,k) és béta priorra $(\underline{\alpha},\underline{\beta})$ alkalmazható R program a Függelékben megtalálható. A kód segítségével elvégzett szimulációk alapján grafikusan is meggyőződhetünk róla, hogy elegendően nagy minta esetén a prior jelentősége egyre kisebb, még akkor is, ha az messze van az adatoktól és a priorban erős véleményt fogalmaztunk

meg az ismeretlen sokasági arányról. Nagy minta esetén a likelihood szerepe értékelődik fel, ne felejtsük el azonban, hogy kis minta esetén az a priori feltételezések hatása nagy lehet.

Jelen esetben egyszerű a poszterior eloszlásban rejlő összes információt grafikusan bemutatni, hisz egydimenziós eloszlásról van szó. Sok esetben azonban egyszerűbb, ha csupán a poszterior eloszlás néhány jellemzőjére hagyatkozunk.

A 5-1. táblázat a prior és a poszterior eloszlások jellemzőit is bemutatja. Ez alapján jól látható, hogy a θ-ra vonatkozó poszterior eloszlások jelentősen közeledtek egymáshoz, még egymástól gyökeresen eltérő prior valószínűségek mellett is. Szeretném kiemelni a nem informatív, $Beta(1,1)$ eloszlást, ahol – mint azt a fentiekben láttuk – a likelihood megegyezik a poszterior eloszlással. Ez azonban nem jelenti azt, hogy a klasszikus maximum likelihood becslés megegyezik a bayesi pontbecslés eredményével.

5-1. táblázat: A prior és poszterior eloszlások néhány jellemzője

Modell	Prior			Poszterior		
	Átlag	Szórás	Módusz	Átlag	Szórás	Módusz
$p(\theta) \sim Beta(1,10)$	0,091	0,083	0	0,268	0,068	0,256
$p(\theta) \sim Beta(10,1)$	0,909	0,083	1	0,488	0,077	0,487
$p(\theta) \sim Beta(1,1)$	0,500	0,289	-	0,344	0,083	0,333
$p(\theta) \sim Beta(0{,}5,0{,}5)$	0,500	0,354	-	0,339	0,084	0,328

Míg az ML becslés a log likelihood maximalizálásával a $\hat{\theta}_{ML} = \frac{k}{n}$ becslőfüggvényt eredményezi, addig a bayesi poszterior a

$$p(\theta | y) = \frac{1}{\mathrm{B}(\underline{\alpha}+k, \underline{\beta}+n-k)} \theta^{\underline{\alpha}+k-1} (1-\theta)^{\underline{\beta}+n-k-1} \sim Beta(\underline{\alpha}+k, \underline{\beta}+n-k)$$

formában írható, így ha a négyzetes veszteség függvényt használjuk (más szóhasználattal pontbecslésként a poszterior átlagot alkalmazzuk), úgy $\hat{\theta}_{Bayes} = \frac{\underline{\alpha}+k}{\underline{\alpha}+k+\underline{\beta}+n-k} = \frac{k+\underline{\alpha}}{n+\underline{\alpha}+\underline{\beta}}$ adódik, ami eltér a maximum likelihood becslőtől. Amennyiben a nem informatív priort használjuk, a pontbecslésünk $\hat{\theta}_{Bayes} = \frac{k+1}{n+2}$

lesz, azaz a nem informatív prior „adattartalma” egy siker és egy kudarc. Hasonlóan értelmezhető az általános eset is. A $Beta(\underline{\alpha},\underline{\beta})$ alakú prior adattartalma $\underline{\alpha}$ siker és $\underline{\beta}$ nem siker megfigyelés. Alacsony k,n és/vagy sokasági θ esetén a nem informatív prior így – az ML becsléshez képest – kifejezetten informatív is lehet. A nem informatív priorral nyert poszterior eloszlás módusza – jelen esetben – értelemszerűen megegyezik a maximum likelihood pontbecsléssel.

Természetesen felmerülhet igényként a pontbecslések mellett egy a klasszikus statisztikában intervallumbecslésként ismert halmaz is. Mivel jelen esetben ismert a poszterior eloszlás, így elegendő a poszterior sűrűségfüggvény megfelelő kvantiliseit számszerűsíteni. Így például, amennyiben 90%-os credible set-et (lásd (4.5) egyenlet) szeretnénk kapni, úgy az 5. és 95. percentilis kiszámítására van szükség. Egy másik – jelen esetben felesleges – megoldás a poszteriorból való szimuláció, majd a keresett kvantilisek szimulált adatokból való meghatározása.

A nem informatív prior esetén kapott poszterior eloszlás a $Beta(11,21)$, aminek percentiliseit könnyedén megkaphatjuk bármely statisztikai programcsomag segítségével. A 95%-os valószínűségi tartomány ez alapján 0,192-0,514, ami nem a legrövidebb az ilyen intervallumok közül, azaz nem HPD (lásd 4.4 alfejezet). A legrövidebb ilyen tartomány meghatározásához szintén a számítógépet hívhatjuk segítségül. Az R rendelkezik azokkal az egyszerű optimalizációs eszközökkel, melyek a legrövidebb, a feltételeknek megfelelő intervallum előállításához szükségesek. A legegyszerűbben használható, egymóduszú poszterior esetén alkalmazható függvények a hpd és az emp.hpd, melyek konjugált analízis, illetve szimulált poszterior esetén adnak megoldást. A 95%-os HPD intervallum tehát a nem informatív priorral számolva 0,186-0,507. A két intervallum közötti eltérés minimális, hiszen a poszterior eloszlás közel szimmetrikus. A HPD intervallumtól balra két-, jobbra háromszázaléknyi terület található, azaz ilyen értelemben nem szimmetrikus.

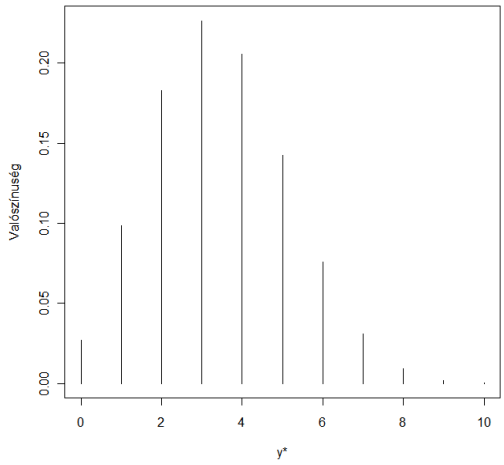
Tegyük fel, hogy meg akarjuk becsülni egy következő, m elemű mintában a sikerek számát a poszterior tudásunk alapján. Az általános, (4.11) képlet alapján:

$$p\left(y^*\middle|y\right)=\int_0^1\binom{m}{y^*}\theta^{y^*}\left(1-\theta\right)^{m-y^*}\frac{1}{\mathrm{B}\left(\bar{\alpha},\bar{\beta}\right)}\theta^{\bar{\alpha}-1}\left(1-\theta\right)^{\bar{\beta}-1}d\theta=$$

$$=\binom{m}{y^*}\frac{1}{\mathrm{B}\left(\bar{\alpha},\bar{\beta}\right)}\int_0^1\theta^{y^*+\bar{\alpha}-1}\left(1-\theta\right)^{m-y^*+\bar{\beta}-1}d\theta=\binom{m}{y^*}\frac{\mathrm{B}\left(\bar{\alpha}+y^*,\bar{\beta}+m-y^*\right)}{\mathrm{B}\left(\bar{\alpha},\bar{\beta}\right)}$$

A fenti eloszlás neve a szakirodalomban béta-binomiális (Gelman et al., 2004).

Amennyiben arra vagyunk kíváncsiak, hogy egy $m=10$ elemű új mintában milyen valószínűséggel érünk el sikert $y^*=0,1,2,\ldots,10$ esetben, úgy a képletbe helyettesítve, majd az eredményt ábrázolva az alábbi, 5-2. ábrán bemutatott valószínűségeket kapjuk. A számításhoz feltételezem a nem informatív $Beta\left(1,1\right)$ priort.



5-2. ábra: Előrejelzés az $m=10$ esetre

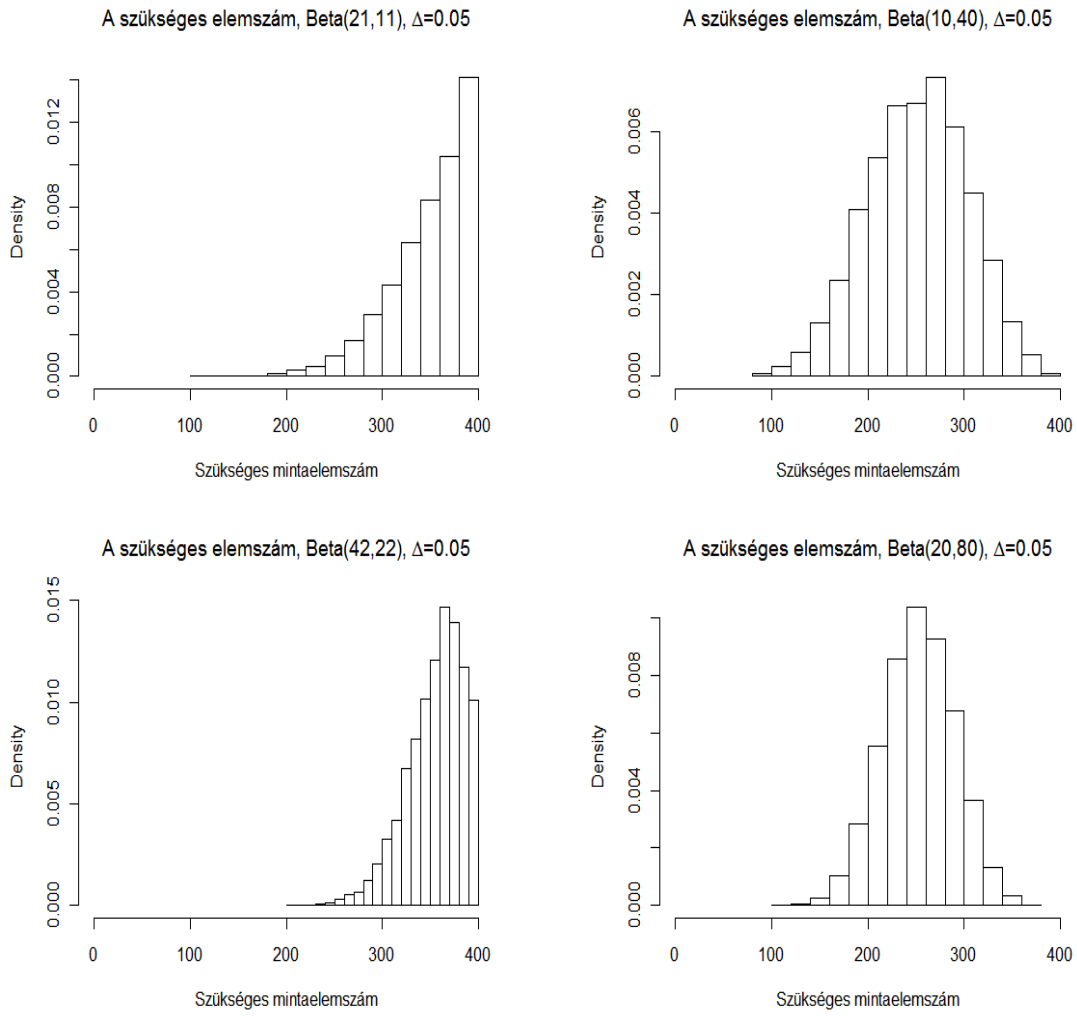
Annak a valószínűsége, hogy a 10 elemű mintában a sikerek száma 1 és 5 között van 85,6%, a 10 siker esélye csupán 0,016% az ismert poszterior alapján.

Az előrejelzés a bayesi felfogásban figyelembe veszi a paraméterekben rejlő és a mintavételből fakadó bizonytalanságot is a teljes értelmezési tartományon történő integrálás segítségével.

A fenti fejtegetést az tette indokolttá, hogy dichotóm kérdések esetén a szükséges mintaelemszámot a sokasági arány határozza meg. A poszteriorban pedig erről a sokasági arányról rendelkezésre álló összes (előzetes és mintabeli) információnk tömörül, így felhasználhatjuk azt a továbbiakban. Bayesi értelemben természetesen a sokasági

arányról nem csupán egy (pl. legvalószínűbb) értékkel rendelkezünk, hanem egy eloszlással. Ennek megfelelően a (3.2) képletbe nem tudunk közvetlenül behelyettesíteni. A legegyszerűbb megoldás a poszteriorból származó véletlen értékek generálása, majd minden generált értékhez a szükséges mintaelemszám kiszámítása. Ezzel megkapjuk a szükséges mintaelemszám közelítő poszterior eloszlását (lásd Függelék). A számolásokhoz továbbra is a 95,5%-os megbízhatóságot használtam fel.

Tegyük fel, hogy a poszterior tudásunk a $Beta(\alpha,\beta)$ eloszlással írható le (függetlenül attól, hogy az tisztán előzetes információból, vagy mintából is származik). A kapott függvény természetesen 0 és $\frac{1}{\Delta_p^2}$ között értelmezett, ahol Δ_p az aránybecslés kívánt hibahatára. A szükséges mintaelemszám ezen az intervallumon felvett eloszlása a poszterior béta eloszlás paramétereitől függ.



5-3. ábra: Szükséges mintaelemszám különböző béta paraméterek esetén

Az 5-3. ábra jól mutaja, hogy a lehetséges mintaelemszámok már alacsony béta paraméterek mellett is elszakadnak a lehetséges maximumtól. Gyakorlatilag nincs a maximum környékén sűrűsége a mintaelemszám eloszlásának, ha a poszterior béta eloszlás a $p = 0,5$ értéket csak kis valószínűséggel tartalmazza.

A fentiekben a binomiális eloszlás példáján keresztül mutattuk be a bayesi alapelveket. A következő alpontban azt vizsgáljuk, hogy a klasszikus statisztikai eszközökkel már hosszasan vizsgált Likert-skálás mintaelemszám tervezés a bayesi keretek között hogyan hajtható végre.

5.2. Mintaelemszám tervezése Likert skálák esetén

A bayesi statisztika eszköztára lehetőséget ad számunkra, hogy a mintaelemszám tervezésekor ne csupán előzetes ismereteinket használjuk fel a priorba ágyazva, hanem – amennyiben erre lehetőség van – kombináljuk azt egy kezdeti mintavétel eredményivel. Ehhez definiálnunk kell egy prior eloszláscsaládot, meg kell határoznunk a prior ismereteinknek megfelelő paramétereket, majd a Bayes-tétel segítségével kombinálnunk kell azt a minta likelihoodjával. Az így nyert poszterior összefoglalja a tudásunkat, a feladatunk ennek a sűrűségfüggvénynek az „összefoglalása” és a mintaelemszámra vonatkozó hatások elemzése. A mintavétel nyelvezetére és a harmadik fejezet ajánlásaira lefordítva ez azt jelenti, hogy a kétlépcsős mintavétel bayesi szemléletbe való átültetése gördülékenyen, a szemlélet logikájának megfelelően elvégezhető. Jelen pontban ezt a gondolatmenetet mutatom be.

Az előző alpontban megismert Bernoulli és binomiális eloszlás és béta-eloszlású konjugált prior eloszlás általánosításával oldható meg a jelen probléma. A Bernoulli eloszlás paramétere tulajdonképp két esemény valószínűségére vonatkozó kételemű paramétervektort $(\theta, 1-\theta)$ határoz meg. Ennek kiterjesztése kettőnél több, k kategóriára adja az ún. kategóriás eloszlás (categorical distribution) alapötletét. Bernoulli-kategóriás eloszlások párhuzamához hasonlóan látható be a binomiális-multinomiális eloszlások kapcsolata. Mivel minden kategóriás eloszlású kísérlet végeredménye pontosan egy kategóriába való kerülés, a multinomiális eloszlás n darab ilyen kísérlet végeredményét írja le egy vektor segítségével. További párhuzam figyelhető meg a prior

választásában. A két kimenetelt vizsgáló esetben a konjugált prior a béta eloszlás volt, a több kimenetelt megengedő, most vizsgálandó eset pedig a béta eloszlás általánosított verzióját, az ún. Dirichlet eloszlást (Gelman et al., 2004) alkalmazza. (A Dirichlet eloszlás sűrűségfüggvényét, jellemzőit, valamint néhány rá vonatkozó ábrát lásd a Függelékben.)

Legyen tehát $Y \sim Kat(\boldsymbol{\theta})$, ahol $\boldsymbol{\theta} = (\theta_1 \quad \theta_2 \quad \ldots \quad \theta_k)$ egy k elemű vektor, ahol $\sum_{j=i}^{k} \theta_j = 1$. Ez az eloszlás írja le annak valószínűségét, hogy egy adott egyén a $j.$ kategóriát választja. Amennyiben rendelkezünk egy n elemű mintával, a likelihood arányos az alábbi kifejezéssel:

$$L(\boldsymbol{\theta}) = p(y|\boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{n_j} \qquad \sum_{j=1}^{k} n_j = n,\ 0 \le \theta_j \le 1 \tag{5.3}$$

ahol n_j a $j.$ kategóriába eső mintaelemek száma.

A konjugált prior eloszlás a Dirichlet eloszlás, ami a binomiális eloszlás tapasztalataiból és a sűrűségfüggvény alakjából egyértelműen látszik. A normalizáló konstanssal nem számolva a poszterior eloszlás könnyedén meghatározható a szokásos módon. A $Dir(\underline{\boldsymbol{\alpha}})$ prior (paramétereit alsó vonással jelölve) és a likelihood szorzatával arányos poszterior:

$$p(\boldsymbol{\theta}|y) \propto p(y|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{n_j} \times \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} = \prod_{j=1}^{k} \theta_j^{\alpha_j + n_j - 1} \sim Dir(\overline{\boldsymbol{\alpha}}) \tag{5.4}$$

ahol $\overline{\alpha}_j = \underline{\alpha}_j + n_j$ minden $j = 1, 2, \ldots, k$ esetére. Gyakorlatilag ugyan azt az eredményt kaptuk, mint a binomiális esetben. A Dirichlet eloszlás a multinomiális eloszlás konjugált prior párja, a poszterior egyszerűen meghatározható, a prior paramétervektorához a megfigyelések megfelelő paramétervektorát kell hozzáadnunk.

Sajnálatos módon a poszterior eloszlás sűrűségfüggvényét csak 3 kategória esetéig lehetséges felrajzolni az első két valószínűség koordinátái segítségével, így a poszteriorban rejlő tudást más módon kell bemutatnunk. Vegyük észre, hogy ha nem rendelkezünk mintával, akkor lehetőségünk van csak a priorban lévő információk kinyerésére is az alábbi módon.

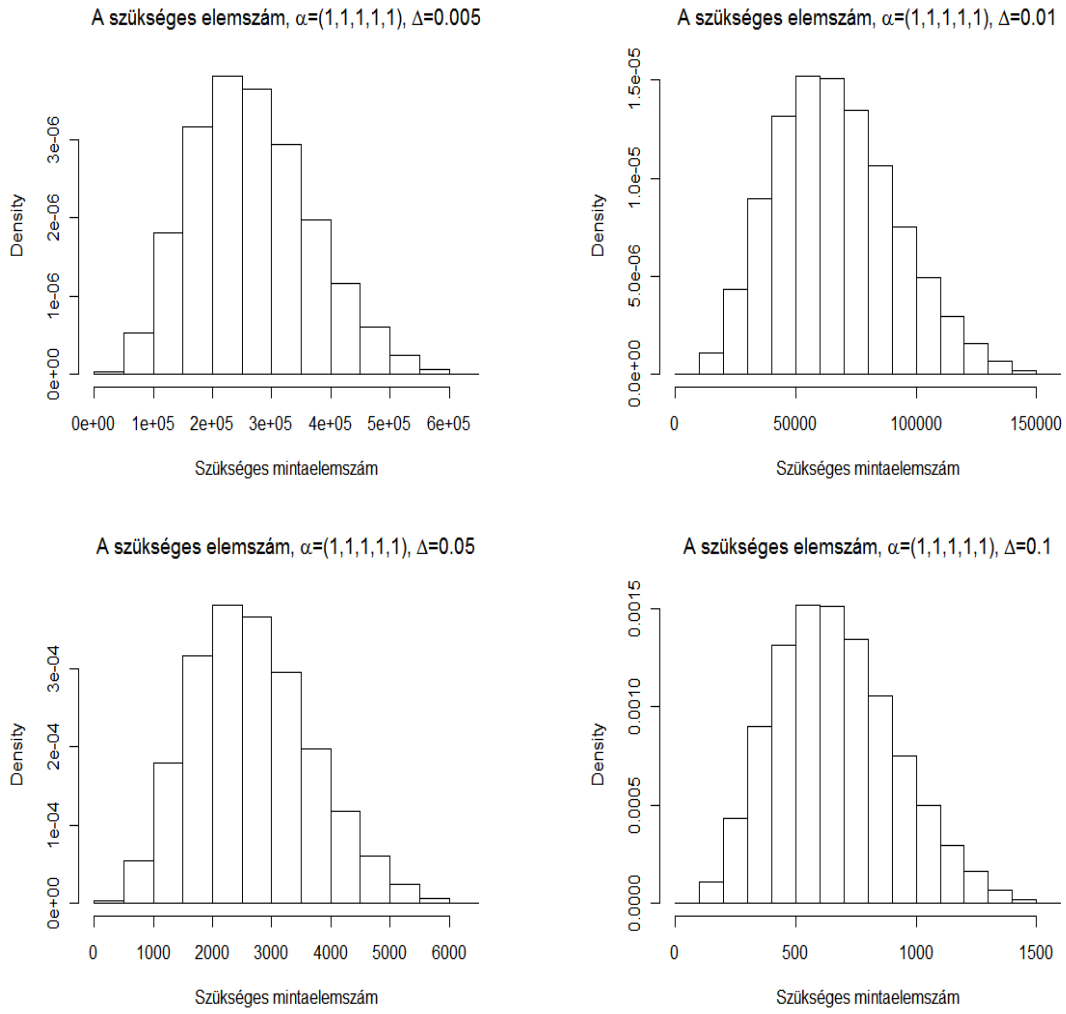
Tegyük fel, hogy a $Dir(\bar{\boldsymbol{\alpha}})$ poszterior eloszlás paramétervektora a fenti jelöléseinknek megfelelően $(\bar{\alpha}_1 \quad \bar{\alpha}_2 \quad \ldots \quad \bar{\alpha}_k)$. Az ebből az eloszlásból nyerhető véletlen érték egy k elemű vektor egységnyi sorösszeggel. Az egyes vektortagok várható értékét a Dirichlet eloszlás tulajdonságai alapján ismerjük, az a j. kategória esetén $\bar{\alpha}_j \Big/ \sum_{j=1}^{k} \bar{\alpha}_j$ értéket vesz fel.

Amennyiben a kategóriákhoz a $1,2,\ldots,k$ számértékeket rendeljük, a poszterior egyetlen realizációjához meghatározható a variancia (hisz a realizáció szolgáltatja a súlyokat), így a szükséges minta elemszám is előáll. A poszterior szimulációja tehát egyben a szükséges mintaelemszám szimulált eloszlását is előállítja.

5.2.1. Nem informatív konjugált prior

A béta eloszláshoz hasonlóan a Dirichlet eloszlás esetén a nem informatív prior a $Dir(\mathbf{1})$ eloszlás. Ez az eloszlás gyakorlatilag egyenletes eloszlást tételez fel az összes lehetséges k elemű vektoron, azt a véleményünket tükrözi, hogy a kimenetek tetszőleges megoszlást felvehetnek egyenletes valószínűséggel (lásd a 3.6. fejezet feltételezésének párhuzamát). Amennyiben nem rendelkezünk mintabeli információkkal és a nem informatív priort használjuk, a klasszikus módon kapott eredményeket kell tapasztalnunk. Az analitikus megoldás nem tűnik célszerűnek, ezért szimulációs módszerrel közelítjük a szükséges mintaelemszám eloszlását, a szükséges program a Függelékben megtalálható.

A kapott eredményeket a 5-4. ábrán foglalom össze. Annak érdekében, hogy az eredmények összehasonlíthatóak legyenek a 3-16. táblázatban közöltekkel, a futtatásokat ötfokozatú skálákra és azonos deltákra végeztem, egyenként 1 000 000 iterációt alkalmazva.



5-4. ábra: Szükséges mintaelemszám különböző hibahatárok és nem informatív prior esetén

A bayesi statisztika logikájának megfelelően az eredmények természetesen (szimulált) eloszlások, amiket valamilyen módon össze kell foglalnunk. Az eloszlások alakjai gyakorlatilag megegyeznek, a különbség a vízszintes tengely értékeiben van csupán. A fenti eloszlások becsült várható értékei sorrendben:

5-2. táblázat: Szükséges mintaelemszámok várható értéke ötfokozatú Likert-skála esetén, előre adott hibahatárok mellett, bayesi közelítésben

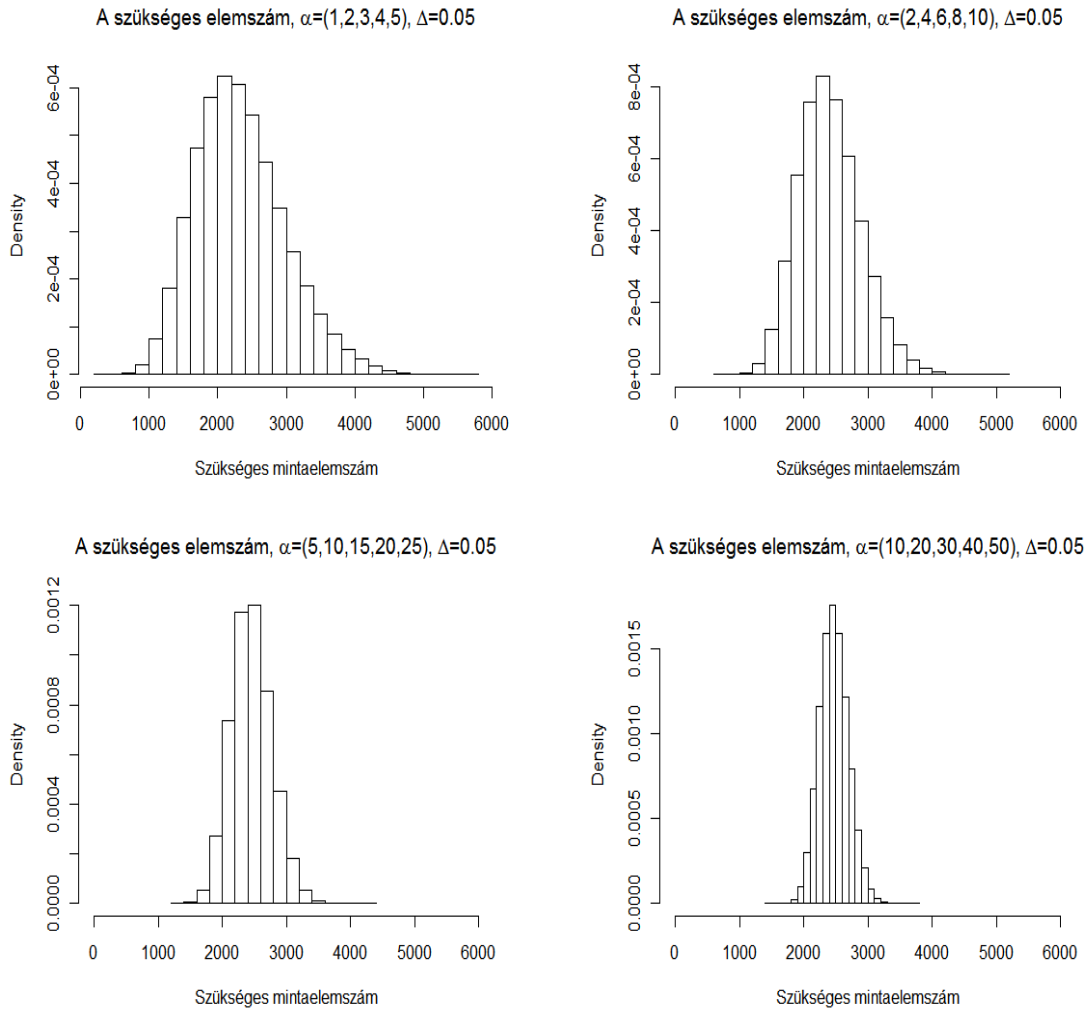
Δ	$E(n)$
0,005	266 616,5
0,010	66 669,5
0,050	2 667,0
0,100	666,5

A kapott értékek a szimulációból adódó pontatlanságon kívül gyakorlatilag megegyeznek a várható értékkel foglalkozó 3.6. fejezetben analitikusan levezetettekkel. A megközelítés előnye ugyanakkor, hogy a várható érték pontbecslése mellett tetszőleges valószínűségű intervallum meghatározása is egyszerű. Amennyiben nincs előzetes információnk és 95% bizonyosságot szeretnénk abban, hogy elegendő mintát veszünk $\Delta = 0{,}100$ pontosság mellett, ~1114 elemű mintát kell vennünk (az empirikus eloszlás 95. percentilise)[22].

5.2.2. Informatív konjugált prior

Nem informatív konjugált prior felállítása azért nehéz, mert a Dirichlet eloszlás paraméterezése nem triviális. Mivel a paraméterek aránya egyben a kategóriák valószínűségének várható értéke is, célszerű a paraméterek egymáshoz viszonyított arányának meghatározásával kezdeni az informatív prior felállítását. Tegyük fel, hogy k fokozatú skálát szeretnénk alkalmazni és a prior tudásunk az, hogy közelítőleg – a korábbi szóhasználattal élve – egyenletesen növekvő eloszlást követnek a válaszlehetőségek. Mivel tudjuk – a Dirichlet eloszlás jellemzői alapján –, hogy az adott kategória bekövetkezési valószínűségének várható értékeit csak a paraméterek egymáshoz viszonyított arányai határozzák meg, ajánlott $Dir(a, 2a, 3a, \ldots, ka)$ struktúrájú prior alkalmazása, ahol az a paraméter bayesi szóhasználattal a prior feszességét adja meg. Minél nagyobb a értéke, annál biztosabbak vagyunk az egyenletesen növekvő eloszlásban. Annak érdekében, hogy a különböző priorok összehasonlítását ne befolyásolják az adatok, a priorok értékelését elvégezzük úgy, mintha azok poszteriorok lennének, szintén 1 000 000 szimuláció segítségével.

[22] Vegyük észre, hogy a klasszikus statisztika által szolgáltatott formula – szimmetrikus eloszlás esetén – ilyen értelemben csak 50%-os bizonyossággal elegséges méretű mintát eredményez.



5-5. ábra: Szükséges mintaelemszám különböző feszességű, egyenletesen növekvő eloszlást leíró priorok esetén

Az 5-5. ábra a fent bemutatott priorok által implikált mintaelemszámokat ábrázolja az $a = 1, 2, 5, 10$ esetekre. A közel azonos várható értékek mellett a szükséges mintaelemszám szóródása jelentős mértékben csökken a priorban található többlet információnak köszönhetően. A szükséges mintaelemszámok eloszlásának becsült várható értéke kerekítve rendre 2 339, 2 409, 2 456 és 2 473 darab. Ezek az értékek jól láthatóan konvergálnak a 3-10. táblázat megfelelő adatához, ami 2 489. Ezt az értéket $a \to \infty$ esetén kapnánk, ami tulajdonképp azt jelentené, hogy a lépcsős eloszlásba vetett hitünk nagyon erős. Mindez azt mutatja, hogy az előzetes információnkkal együtt azok bizonytalansága is könnyedén építhető be a bayesi keretrendszerben, míg klasszikus esetben ez nem igazán lehetséges.

Amennyiben a nem informatív prior mellé lehetőségünk van adatok begyűjtésére is, akkor a (5.4) formula segítségével előállított Dirichlet-eloszlásban rejlő információk bemutatása a feladatunk.

A fejezetben a bayesi elvek alkalmazását mutattam be a mintaelemszám tervezés kapcsán. A kétváltozós eset konjugált analízise után a Likert-skálás kérdések mellett alkalmazható eljárást is bemutattam. Az eredmények nem informatív prior esetén teljes mértékben összhangban vannak a klasszikus esetben bemutatottakkal. A bayesi megközelítés előnye, hogy a bizonytalanság (miszerint az előzetes feltevésünk az eloszlással kapcsolatban nem teljesen pontos) könnyen kezelhető. Hátrányként fogalmazhatjuk meg, hogy az informatív prior előállítása gyakorlati esetben nem kézenfekvő.

6. Összefoglalás, további kutatási irányok

A doktori disszertációmban a mintavétel egyik fontos részterületével, a mintaelemszám tervezésével foglalkoztam. Az első fejezetben a bevezetés mellett bemutattam a kutatási irányokat és a dolgozat szerkezetét. A második fejezetben a mérési skálák rendszerét, illetve az ehhez kapcsolódó tudományos vitát tekintettem át azzal a céllal, hogy jobban megértsem a statisztikai műveletek alkalmazhatóságának feltételeit Likert-skálás lekérdezések segítségével nyert változók esetén. A harmadik fejezetben a mintaelemszám tervezés hagyományos, aránybecslésre épülő módszerének bemutatása mellett kidolgoztam a Likert-skálák esetére alkalmazható képleteket, eljárásokat. A fejezet konklúziójaként azt találtam, hogy az általam javasolt mintaelemszám tervezés mellett szükség szerint előzetes mintavételt kell végrehajtani. A két információforrás összefogása legegyszerűbben a bayesi keretek között oldható meg, melynek legfontosabb fogalmairól és eljárásairól rövid összefoglalót adtam. A komplex, mindkét információforrást figyelembe vevő bayesi eredmények leírása adja a dolgozat ötödik fejezetét.

Az alábbiakban a bevezetőben már taglalt hipotézisekre adott válaszaimat foglalom össze röviden.

1. **A Likert-skálás lekérdezések segítségével nyert változók esetén a módszerválasztás különös jelentőséggel bír.**

   A mérési skálák elmélete és az azzal kapcsolatos tudományos vita rávilágít arra, hogy a komoly irodalmi csatározás ellenére továbbra is több vélemény, ha úgy tetszik iskola létezik. A klasszikus nominális-ordinális-intervallum-arány skálarendszer hiányosságait többen jelezték, az egyik – témakörünk szempontjából jelentős - kritika az, hogy adott változók bizonyos esetekben nehezen sorolhatók be a fenti kategóriákba. A Likert-skála kapcsán felvetődik a kérdés, hogy ordinális, vagy intervallum skála erősségű eredményeket kapunk-e felhasználásával. Mindet azért fontos, mert a leggyakrabban alkalmazott statisztikai módszertanok csak intervallum, vagy arány skálán értelmezhetőek. A Likert-skálás alkalmazások esetén a leggyakrabban praktikus szempontok érvényesülnek, törekedni kell arra, hogy a

változóértékek valóban ekvidisztansok legyenek. Amennyiben ez a lekérdezés formájával (megfelelő kérdésfeltevés, kategórianevek) nem biztosítható, úgy ordinális skálán mért változóként kell kezelnünk az eredményeinket.

2. **Amennyiben rendelkezünk külső információval a sokasági eloszlásról, az hatékonyan alkalmazható az előzetesen kalkulált szükséges mintaelemszám csökkentésére Likert-skálás kérdések esetén is.**

A hagyományosan alkalmazott – aránybecslés hibahatárából kiinduló – mintaelemszám tervezés esetén is lehetőség van külső információk felhasználására, ekkor a sokasági arányról élhetünk feltételezéssel. Ezzel analóg módon alkalmazható külső információ abban az esetben, ha várható értékre vonatkozó becslést kívánunk végrehajtani. Ebben az esetben a sokasági varianciáról kell tudással rendelkeznünk. Mivel ez a gyakorlatban ritkán fordul elő, azt a megközelítést alkalmaztam, hogy nem a varianciáról, hanem az azt implikáló eloszlásról van sejtésünk. Amennyiben kevés kimenetelű a változónk, úgy definiálhatók olyan tipikus eloszlások, melyek esetén a variancia könnyen meghatározható. A varianciák segítségével számított mintaelemszámok azt mutatják, hogy a szükséges mintaelemszám akár tízszeres is lehet a különböző eloszlások között. Amennyiben tehát rendelkezünk információval, annak tervezésbe való beépítése jelentős megtakarításokat jelenthet a mintaelemszám tekintetében.

3. **Adott kívánt hibahatár és megbízhatósági szint esetén meghatározható a szükséges mintaelemszám várható értéke.**

Abban az esetben, ha nincs előzetes elképzelésünk a válaszlehetőségek közötti megoszlásról, azt feltételezzük, hogy minden lehetőségnek azonos a valószínűsége, úgy a mintaelemszámok várható értékére (és eloszlására) vagyunk kíváncsiak. Egy konkrét eset bemutatása után általános esetben is sikerül meghatároznom a keresett várható értékre vonatkozó – meglepően egyszerű – képletet.

4. **Az előzetes információk és egy esetleges előzetes mintavétel adatainak összesítése a bayesi keretrendszerben probléma nélkül megoldható. A bayesi módszertannal kiszámított szükséges mintaelemszám értékek a klasszikus statisztikai módszerekkel számítható értékekkel összhangot mutatnak.**

A bayesi megközelítés egyrészt prior, másrészt mintabeli információk meglétét feltételezi. A klasszikus eredményekkel való összehasonlítás úgy lehetséges, hogy a priorba beépítjük az arányról, vagy az eloszlás formájáról rendelkezésünkre álló adatokat, majd ezt az eloszlást tekintjük egyben poszteriornak is. A két megközelítés különbsége, hogy a bayesi esetben egy konkrét érték helyett egy eloszlás az eredmény, melynek várható értéke azonban a két vizsgált esetben közelítően megegyezik.

5. **Az előzetes információk bizonytalansága és a mintavételi hiba figyelembe vehető a bayesi gondolatvilágban, ami egyértelműen előnyt jelent a klasszikus megközelítéshez képest.**

Az előzetes információk a bayesi keretek között egy megfelelő eloszláscsaládba tartozó eloszlás paraméterein keresztül építhetők a modellbe. Az ehhez társuló, mintabeli adatokat leíró likelihood segítségével meghatározott poszterior szintén egy eloszlás, azaz nem csupán egy pontbecslés, definíciójánál fogva tartalmazza a variancia és ezzel együtt a szükséges mintaelemszám bizonytalanságát. Az ismeretlen (ténylegesen szükséges mintaelemszám) sokasági paramétert nagy valószínűséggel tartalmazó intervallum meghatározása bayesi értelemben magától értetődő.

További kutatási területként két irányt vázolok fel röviden. Egyrészt a becslés témaköréből kilépve a hipotézisellenőrzés esetén felhasználható eljárások kidolgozása a következő célom. Az ilyen jellegű vizsgálatok figyelembe veszik az elérni kívánt α szignifikancia szint mellett a másodfajú hiba elkövetésének valószínűségét, illetve a próba erejét is. További kutatást indukál az egyszerű véletlen mintavételtől eltérő, például rétegzett mintavétel esetén történő mintaelemszám meghatározás.

A másik fontos kutatási irány a bayesi statisztika és ökonometria alkalmazásainak megismerése és lehetőség szerinti fejlesztése. A klasszikus statisztikai eszköztár valamennyi módszere becsülhető bayesi értelemben is, sok esetben – megfelelő nem informatív priort használva – pedig elkerülhetőek identifikációs problémák is a segítségével. A bayesi ökonometria gazdag nemzetközi irodalma jó alapot nyújt ehhez a jövőbeli kutatási irányhoz.Equation Section (Next)

7. Függelék

Equation Section (Next)7.1. Az első k szám átlagtól mért súlyozatlan eltérés-négyzetösszege:

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

7.2. Az átlagos (relatív gyakoriságokkal súlyozott) eltérés-négyzetösszeg a páratlan szimmetrikus esetben:

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

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

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

7.3. Variancia piramis eloszlás esetén

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

Ismeretes, hogy

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

ebből

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

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

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

az átlagot visszahelyettesítve

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

7.4. Variancia fordított piramis eloszlás esetén

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

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

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

7.5. Variancia extrém egymóduszú eloszlás esetén

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

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

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

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

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

7.6. Variancia egyenletesen növekvő valószínűségek esetén

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

ismeretes, hogy

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

a dolgozatban alkalmazott jelölésekkel

$$MSS^{(k)} = \frac{2\sum_{j=1}^{k} j^3}{k(k+1)} - \left(\frac{2k+1}{3} \right)^2 = \frac{k(k+1)}{2} - \frac{4k^2+4k+1}{9} = \frac{(k-1)(k+2)}{18}$$

7.7. A várható érték meghatározásához szükséges megfontolásokat az alábbiakban mutatom be:

Írjuk fel a varianciák várható értékét a következő módon:

$$E\left(\sigma_{n,k}^2\right) = E\left(\frac{\sum_{i=1}^{k} n_i \cdot i^2}{n} - \left(\frac{\sum_{i=1}^{k} n_i \cdot i}{n} \right)^2 \right) = \frac{E\left(\sum_{i=1}^{k} n_i \cdot i^2 \right)}{n} - \frac{E\left[\left(\sum_{i=1}^{k} n_i \cdot i \right)^2 \right]}{n^2} \qquad (7.1)$$

Ezután az egyenletben szereplő két várható érték meghatározása a feladat. Írjuk fel előbb az

$$E\left(\sum_{i=1}^{k} n_i \cdot i^2\right) \tag{7.2}$$

kifejezést. Mivel tudjuk, hogy $\binom{n+k-1}{k-1}$ kimenetel várható értékét szeretnénk kiszámolni, méghozzá az egyenlő súlyozás miatt „súlyozatlanul", ezért az alábbi, kibontott alakban írhatjuk fel az összeget. A felírásban a jobb átláthatóság kedvéért valamennyi kimenetelt külön zárójelezem.

$$\frac{\left(0\cdot 1^2+0\cdot 2^2+\ldots+0\cdot(k-1)^2+n\cdot k^2\right)+\left(0\cdot 1^2+0\cdot 2^2+\ldots+1\cdot(k-1)^2+(n-1)\cdot k^2\right)+\ldots}{\binom{n+k-1}{k-1}}$$

$$\frac{\ldots+\left((n-1)\cdot 1^2+1\cdot 2^2+\ldots+0\cdot(k-1)^2+0\cdot k^2\right)+\left(n\cdot 1^2+0\cdot 2^2+\ldots+0\cdot(k-1)^2+0\cdot k^2\right)}{\binom{n+k-1}{k-1}}=$$

Vegyük észre, hogy valamennyi kimenetből kiemelhetjük 1^2-t. Az 1^2 kifejezés $\binom{n+k-1}{k-1}$ alkalommal szerepel, amiből $\binom{n+k-2}{k-2}$ alkalommal 0-val „van párban"[23], $\binom{n+k-3}{k-2}$ alkalommal 1-essel, végül $\binom{k-2}{k-2}=1$ (jelölésünk szerint az utolsó) alkalommal n-nel. Tekintsük tehát kifejtve csupán az 1^2 kiemelése után megmaradó összeget:

$$=\frac{1^2\cdot\left(0+0+\ldots+0+1+1+\ldots+(n-2)+(n-1)+(n-1)+n\right)+\ldots}{\binom{n+k-1}{k-1}}=$$

[23] Amennyiben bármely válaszlehetőség 0-val van párban, az azt jelenti, hogy az adott lehetőséget senki sem választotta. Ez természetesen azt jelenti, hogy annyi ilyen eset van, ahányféleképp $k-1$ elem n-ed osztályú ismétléses kombinációja képezhető. Amennyiben 1-gyel van párban a válaszlehetőség, úgy $k-1$ elem $n-1$-ed osztályú ismétléses kombinációiról van szó, általánosságban amennyiben az adott válaszlehetőségre $m\left(0\le m\le n\right)$ válasz érkezett, azaz m-mel van párban, úgy $k-1$ elem $n-m$-ed osztályú ismétléses kombinációiról beszélhetünk.

Végezzük el a kiemelést $2^2, 3^2, \ldots, k^2$-re is! A kiemeléseket elvégezve azt vehetjük észre, hogy a négyzetszámok együtthatói azonosak, hisz teljesen mindegy, hogy melyik számot (válaszlehetőséget) vizsgáljuk, a „párok" ugyan azok maradnak, így a teljes összegre az alábbiakat írhatjuk fel:

$$= \frac{\left(0+0+\ldots+0+1+1+\ldots+(n-2)+(n-1)+(n-1)+n\right)\cdot\left(1^2+2^2+\ldots+k^2\right)}{\binom{n+k-1}{k-1}} =$$

Mivel tudjuk, hogy a fenti kifejezés első tényezőjében hány darab 0, hány darab 1-es stb. szerepel, valamint kihasználva az első k négyzetszám összegéről ismert összefüggést, a következő zárt alakot írhatjuk fel:

$$= \frac{\sum_{j=0}^{n}\left[\binom{n+k-2-j}{k-2}\cdot j\right]\cdot\frac{k(k+1)(2k+1)}{6}}{\binom{n+k-1}{k-1}}$$

A következő lépéshez szükséges összefüggés igazolását terjedelmi okok és a gondolat folytonossága miatt a Függelék 7.8. pontjában mutatom be.

A bizonyított $\sum_{j=0}^{n}\binom{n+k-2-j}{k-2}\cdot j = \binom{n+k-1}{k}$ összefüggés felhasználásával, valamint egyszerűsítések elvégzése után kapjuk a szükséges részeredményt:

$$E\left(\sum_{i=1}^{k} n_i \cdot i^2\right) = \frac{n(k+1)(2k+1)}{6} \tag{7.3}$$

A fentiek után meg kell határoznunk a következő, (7.1)-ben szereplő kifejezést:

$$E\left[\left(\sum_{i=1}^{k} n_i \cdot i\right)^2\right] \tag{7.4}$$

A (7.2) kifejezés meghatározásánál alkalmazott módszerhez hasonlóan most is felírhatjuk az $\binom{n+k-1}{k-1}$ különböző hatványozandó összegét kifejtve:

$$= \frac{\left(0\cdot 1+0\cdot 2+\ldots+0\cdot(k-1)+n\cdot k\right)^2+\left(0\cdot 1+0\cdot 2+\ldots+1\cdot(k-1)+(n-1)\cdot k\right)^2+\ldots}{\binom{n+k-1}{k-1}}$$

$$\frac{\ldots+\left((n-1)\cdot 1+1\cdot 2+\ldots+0\cdot(k-1)+0\cdot k\right)^2+\left(n\cdot 1+0\cdot 2+\ldots+0\cdot(k-1)+0\cdot k\right)^2}{\binom{n+k-1}{k-1}} =$$

Emeljük négyzetre valamennyi n tagú összeget. A négyzetre emelés után külön csoportosíthatjuk a négyzetes (A), illetve a vegyes tagokat (B).

Az áttekinthetőség érdekében tekintsük elsőként a négyzetes tagokat:

$$A = \frac{\left[\left(0^2\cdot 1^2\right)+\left(0^2\cdot 2^2\right)+\ldots+\left(n^2\cdot k^2\right)\right]+\left[\left(0^2\cdot 1^2\right)+\left(0^2\cdot 2^2\right)+\ldots+\left(1^2\cdot(k-1)^2\right)+\left((n-1)^2\cdot k^2\right)\right]+\ldots}{\binom{n+k-1}{k-1}}$$

$$\frac{\ldots+\left[\left((n-1)^2\cdot 1^2\right)+\left(1^2\cdot 2^2\right)+\ldots+\left(0^2\cdot(k-1)^2\right)+\left(0^2\cdot k^2\right)\right]+\left[\left(n^2\cdot 1^2\right)+\left(0^2\cdot 2^2\right)+\ldots+\left(0^2\cdot k^2\right)\right]}{\binom{n+k-1}{k-1}} =$$

A négyzetes tagok esetén az előző várható érték meghatározásánál alkalmazott sorozatos kiemelést $\left(1^2, 2^2, \ldots, k^2\right)$ alkalmazva struktúrájában hasonló eredményre juthatunk. A különbség csupán az, hogy szorzótényezőként a futóindex négyzetesen szerepel:

$$= \frac{\left[\sum_{j=0}^{n}\binom{n+k-2-j}{k-2}\cdot j^2\right]\cdot\frac{k(k+1)(2k+1)}{6}}{\binom{n+k-1}{k-1}} =$$

A következő lépésben felhasználom a $\sum_{j=0}^{n}\binom{n+k-2-j}{k-2}\cdot j^2 = \binom{n+k-1}{k}+2\cdot\binom{n+k-1}{k+1}$ összefüggést, valamint egyszerű ekvivalens átalakításokat végzek el. A felhasznált egyenlőség könnyedén belátható a Függelék 7.8. pontjának instrukciói alapján. A négyzetes tagok összege a fentiek alapján a következő alakban írható:

$$A = \frac{(k+1)+2(n-1)}{6} \cdot (2k+1) \cdot n \tag{7.5}$$

A négyzetes tagok összegének meghatározása után tekintsük a vegyes tagokat. A vegyes tagok esetén is kiemeléseket kell elvégeznünk.

$$B = \frac{2\left[\{0\cdot1\cdot0\cdot2+0\cdot1\cdot0\cdot3+\ldots+0\cdot1\cdot n\cdot k\}+\{0\cdot2\cdot0\cdot3+0\cdot2\cdot0\cdot4+\ldots+0\cdot2\cdot n\cdot k\}+\ldots+\{0\cdot(k-1)\cdot n\cdot k\}\right]}{\binom{n+k-1}{k-1}}+$$

$$\frac{2\left[\{0\cdot1\cdot0\cdot2+0\cdot1\cdot0\cdot3+\ldots+0\cdot1\cdot(n-1)\cdot k\}+\{0\cdot2\cdot0\cdot3+\ldots+0\cdot2\cdot(n-1)\cdot k\}+\ldots+\{1\cdot(k-1)\cdot(n-1)\cdot k\}\right]}{\binom{n+k-1}{k-1}}+$$

$$\vdots$$

$$\frac{2\left[\{n\cdot1\cdot0\cdot2+n\cdot1\cdot0\cdot3+\ldots+n\cdot1\cdot0\cdot k\}+\{0\cdot2\cdot0\cdot3+0\cdot2\cdot0\cdot4+\ldots+0\cdot2\cdot0\cdot k\}+\ldots+\{0\cdot(k-1)\cdot0\cdot k\}\right]}{\binom{n+k-1}{k-1}}=$$

Elsőként az összes sor[24] első tagjának első tagjából emeljünk ki $1\cdot2$-t. A kiemelés után $\binom{n+k-3}{k-3}$ alkalommal marad $0\cdot0$, $\binom{n+k-4}{k-3}$ alkalommal $0\cdot1$, $\binom{k-3}{k-3}$ alkalommal $0\cdot n$, $\binom{n+k-4}{k-3}$ alkalommal $1\cdot0$, végül $\binom{k-3}{k-3}$ alkalommal $n\cdot0$[25]. Végezzük el a kiemelést $1\cdot3$-ra (minden sor első tagjának második tagjából), $1\cdot4$-re, végül $1\cdot k$-ra. Ezt követően emeljünk ki $2\cdot3$-at, $2\cdot4$-et, legvégül $(k-1)\cdot k$-t. Valamennyi kiemelés után ugyanaz az összeg áll elő, így ezt az összeget kiemelve kapjuk a következőket:

---

[24] Egy sor egyetlen átlag kibontásából származó vegyes tagoknak felel meg, így összesen $\binom{n+k-1}{k-1}$ darab sorból áll a fenti összeg.

[25] Általánosságban (bármely szorzat kiemelése után) $\binom{n+k-3-f-g}{k-3}$ alkalommal marad $f\cdot g$, $0\le f+g\le n$.

$$2\cdot\frac{\left[\begin{array}{l}\left\{\binom{n+k-3}{k-3}0\cdot 0+\binom{n+k-4}{k-3}0\cdot 1+\ldots+\binom{k-2}{k-3}0\cdot(n-1)+\binom{k-3}{k-3}0\cdot n\right\}+\\ +\left\{\binom{n+k-4}{k-3}1\cdot 0+\binom{n+k-5}{k-3}1\cdot 1+\ldots+\binom{k-3}{k-3}1\cdot(n-1)\right\}+\\ \vdots\\ +\left\{\binom{k-2}{k-3}(n-1)\cdot 0+\binom{k-3}{k-3}(n-1)\cdot 1\right\}+\\ +\left\{\binom{k-3}{k-3}n\cdot 0\right\}\end{array}\right]}{\binom{n+k-1}{k-1}}\cdot$$

$$\cdot\left[\{1\cdot 2+1\cdot 3+\ldots+1\cdot k\}+\{2\cdot 3+2\cdot 4+\ldots+2\cdot k\}+\ldots+\{(k-1)\cdot k\}\right]$$

A fenti kifejezés első tényezőjének soraiból rendre 0-t, 1-et, …, $n-1$-et kiemelhetünk. A kapcsos zárójelekben ekkor $\sum\binom{n+k-3-j}{k-3}\cdot j$ típusú összegek szerepelnek a kiemelt számokon kívül. Amennyiben azt a sort tekintjük, ahol m-et emelhetünk ki[26], a következő összeget kapjuk: $m\cdot\sum_{j=0}^{n-m}\binom{n+k-3-m-j}{k-3}\cdot j$

A fenti kifejezés első tényezőjének számlálója tehát az alábbi zárt alakra hozható, amit a már jól ismert – függelékben bemutatott – összefüggés kétszeri alkalmazásával egyszerűen kiszámíthatunk:

$$\sum_{m=0}^{n}\left(m\cdot\sum_{j=0}^{n-m}\binom{n+k-3-m-j}{k-3}\cdot j\right)=\sum_{m=0}^{n}\left(m\cdot\binom{n+k-2-m}{k-1}\right)=\binom{n+k-1}{k+1}.$$

A kifejezés második tényezője, melyben a kiemelt szorzatok összege szerepel, megfelelő kiemelések után[27] a következő alakra hozható:

---

[26] Az $m+1.$ sorról van szó. Az összeg összesen $n+1$ sorból áll.

[27] Elsőként k-t, majd $k-1$-et stb. emelhetünk ki. A kiemelések után rendre a kiemelt számnál eggyel kisebb számig terjedően a számjegyek összege marad tényezőként; ezek után zárt alakra hozható az összeg.

$$\frac{(k-1)k(k+1)(3k+2)}{24}.$$

A vegyes tagok összege tehát a következőképp írható fel:

$$B=\frac{2\binom{n+k-1}{k+1}(k-1)k(k+1)(3k+2)}{24\binom{n+k-1}{k-1}}=\frac{n(n-1)(k-1)(3k+2)}{12} \tag{7.6}$$

Felhasználva a négyzetes tagok összegéről és a vegyes tagok összegéről megállapítottakat, némi átalakítást elvégezve a következő eredményt kapjuk (7.5) és (7.6) felhasználásával:

$$E\left[\left(\sum_{i=1}^{k} n_i \cdot i\right)^2\right]=A+B=\frac{k(k-1)n+(3k+1)(k+2)n^2}{12} \tag{7.7}$$

Ne felejtsük el, hogy ez csak részeredmény. A megoldást a variancia várható értékére keressük. A felírt és bizonyított (7.3) és (7.7) összefüggések alapján tehát:

$$E\left(\sigma_{n,k}^2\right)=\frac{E\left(\sum_{i=1}^{k} n_i \cdot i^2\right)}{n}-\frac{E\left[\left(\sum_{i=1}^{k} n_i \cdot i\right)^2\right]}{n^2}=\frac{(k+1)(2k+1)}{6}-\frac{k(k-1)}{12n}-\frac{(3k+1)(k+2)}{12}$$

7.8. A várható érték meghatározásához szükséges tétel és bizonyítása:

<u>Tétel:</u> $\sum_{j=0}^{n}\binom{n+k-j-2}{k-2}\cdot j=\binom{n+k-1}{k}$

<u>Bizonyítás:</u> Tekintsük a Pascal-féle háromszög egy kiválasztott darabját. A bizonyítandó állítás bal oldalán látható összeg binomiális tagjai jól kivehetően a háromszög egy „átlóját” alkotják. Például $j=n$ esetre az összeg tagja $\binom{k-2}{k-2}\cdot n$, $j=n-1$ esetre $\binom{k-1}{k-2}\cdot(n-1)$ stb. Az egyes binomiális kifejezések szorzótényezőként szerepelnek a futóindex mellett. Jelöljük az ábrán kis indexekkel azt, hogy „hány darabot” kell összeadnunk az adott kifejezésből. Az alábbi ábrán tehát a fenti összeg egy darabja látható.

$$
\begin{array}{ccccccc}
\ddots & \binom{k-3}{k-4} & & \binom{k-3}{k-3} & & & \\
\binom{k-2}{k-4} & & \binom{k-2}{k-3} & & \binom{k-2}{k-2}_{n} & & \\
& \binom{k-1}{k-3} & & \binom{k-1}{k-2}_{n-1} & & \binom{k-1}{k-1} & \\
\binom{k}{k-3} & & \binom{k}{k-2}_{n-2} & & \binom{k}{k-1} & & \binom{k}{k} \\
& \binom{k+1}{k-2}_{n-3} & & \binom{k+1}{k-1} & & \binom{k+1}{k} & \\
\binom{k+2}{k-2}_{n-4} & & \binom{k+2}{k-1} & & \binom{k+2}{k} & & \ddots \\
& \vdots & & \vdots & & \vdots &
\end{array}
$$

Az összegzés megkönnyítése érdekében használjuk ki, hogy $\binom{k-2}{k-2}=\binom{k-1}{k-1}$, valamint, hogy $(n-1)\cdot\binom{k-1}{k-2}+(n-1)\cdot\binom{k-1}{k-1}=(n-1)\cdot\binom{k}{k-1}$. A fenti összegzések elvégzése után vegyük még észre, hogy $\binom{k-1}{k-1}=\binom{k}{k}$. Ekkor az összeg ábrázolható a következő módon, kis indexekkel újra az összegzendő mennyiséget láthatjuk:

$$
\begin{array}{ccccccc}
\ddots & \binom{k-3}{k-4} & & \binom{k-3}{k-3} & & & \\
\binom{k-2}{k-4} & & \binom{k-2}{k-3} & & \binom{k-2}{k-2} & & \\
& \binom{k-1}{k-3} & & \binom{k-1}{k-2} & & \binom{k-1}{k-1} & \\
\binom{k}{k-3} & & \binom{k}{k-2}_{n-2} & & \binom{k}{k-1}_{n-1} & & \binom{k}{k}_{1} \\
& \binom{k+1}{k-2}_{n-3} & & \binom{k+1}{k-1} & & \binom{k+1}{k} & \\
\binom{k+2}{k-2}_{n-4} & & \binom{k+2}{k-1} & & \binom{k+2}{k} & & \ddots \\
& \vdots & & \vdots & & \vdots &
\end{array}
$$

Ezután a következő algoritmus alapján kiszámíthatjuk a következő lépés eredményét: párosítsuk az $n-2$ darab összeadandó $\binom{k}{k-2}$-t $n-2$ darab $\binom{k}{k-1}$-gyel. A maradék 1 darab $\binom{k}{k-1}$-et pedig $\binom{k}{k}$-val. Ekkor az alábbi ábrán szemléltethetjük az eredményt:

$$
\begin{array}{ccccccc}
\ddots & \binom{k-3}{k-4} & & \binom{k-3}{k-3} & & & \\
\binom{k-2}{k-4} & & \binom{k-2}{k-3} & & \binom{k-2}{k-2} & & \\
& \binom{k-1}{k-3} & & \binom{k-1}{k-2} & & \binom{k-1}{k-1} & \\
\binom{k}{k-3} & & \binom{k}{k-2} & & \binom{k}{k-1} & & \binom{k}{k} \\
& \binom{k+1}{k-2}_{n-3} & & \binom{k+1}{k-1}_{n-2} & & \binom{k+1}{k}_{1} & \\
\binom{k+2}{k-2}_{n-4} & & \binom{k+2}{k-1} & & \binom{k+2}{k} & & \ddots \\
& \vdots & & \vdots & & \vdots &
\end{array}
$$

A fenti lépéseket addig végezzük, míg az alábbi összegzendőket nem kapjuk:

$$
\begin{array}{ccccccc}
\vdots & \vdots & \vdots & \binom{n+k-5}{k-2} & \vdots & \vdots & \vdots \\
& & \binom{n+k-4}{k-2}_{2} & & \binom{n+k-4}{k-1}_{3} & & \binom{n+k-4}{k}_{1} \\
& \binom{n+k-3}{k-2}_{1} & & \binom{n+k-3}{k-1} & & \binom{n+k-3}{k} & \\
\binom{n+k-2}{k-2}_{0} & & \binom{n+k-2}{k-1} & & \binom{n+k-2}{k} & & \ddots
\end{array}
$$

A fenti, egyszerű lépéseket tovább alkalmazva eljuthatunk $\binom{n+k-2}{k-1}$ és $\binom{n+k-2}{k}$ összegéig, ami $\binom{n+k-1}{k}$. Ezzel a bizonyítandó állítást beláttuk.

A $\sum_{j=0}^{n}\binom{n+k-2-j}{k-2}\cdot j^2=\binom{n+k-1}{k}+2\cdot\binom{n+k-1}{k+1}$ összefüggést hasonló módszerrel bizonyíthatjuk. Az összegzésnél mindig balról indulunk, és sorról sorra haladunk lefelé.

Itt kell megjegyezni, hogy hasonló összefüggés j bármely pozitív egész hatványára belátható, a fentiekkel megegyező módon. Minderre akkor lenne szükségünk, ha magasabb rendű momentumokat is vizsgálnánk.

7.9. Kiegészítő számítások a kontroll változós varianciacsökkentő módszerhez

$$f(x)=\frac{x-2}{2}$$

$$g(x)=e^{-x}$$

$$E\left[f(x)\right]=\int_2^4\frac{x-2}{2}\frac{1}{2}=\left[\frac{x^2}{8}-\frac{x}{2}\right]_2^4=\frac{1}{2}$$

$$E\left[f^2(x)\right]=\int_2^4\frac{x^2-4x+4}{4}\frac{1}{2}=\left[\frac{x^3}{24}-\frac{x^2}{4}+\frac{x}{2}\right]_2^4=\frac{1}{3}$$

$$Var\left[f(x)\right]=\frac{1}{3}-\frac{1}{4}=\frac{1}{12}$$

$$E\left(g(X)\right)=\frac{e^{-2}-e^{-4}}{2}$$

$$E\left[f(X)g(X)\right]=\int_2^4 e^{-x}\frac{x-2}{2}\frac{1}{2}=\left[-e^{-x}\frac{x-2}{4}\right]_2^4+\int_2^4 e^{-x}\frac{1}{4}=\frac{e^{-2}-3e^{-4}}{4}$$

$$Cov\left[f(X)g(X)\right]=\frac{e^{-2}-3e^{-4}}{4}-\frac{e^{-2}-e^{-4}}{4}=-\frac{e^{-4}}{2}$$

$$c^*=-\frac{-\frac{e^{-4}}{2}}{\frac{1}{12}}=6e^{-4}$$

$$Var\left(g(X)+c^*\left(f(X)-\eta\right)\right)=Var\left[g(X)\right]-\frac{\left[Cov\left(g(X),f(X)\right)\right]^2}{Var\left[f(X)\right]}=$$

$$=\frac{e^2-1}{2e^8}-\frac{\frac{e^{-8}}{4}}{\frac{1}{12}}=\frac{e^2-1}{2e^8}-3e^{-8}$$

$$\frac{Var(\theta_1)-Var(\theta_2)}{Var(\theta_1)}=\frac{\left[Cov\left(g(X),f(X)\right)\right]^2}{Var\left[f(X)\right]Var\left[g(X)\right]}=\frac{\frac{e^{-8}}{4}}{\frac{1}{12}\frac{e^2-1}{2e^8}}=\frac{6}{e^2-1}$$

7.10. Fontos eloszlások rövid bemutatása

Multinomiális eloszlás

A multinomiális eloszlás a binomiális eloszlás általánosításaként értelmezhető. Tegyük fel, hogy a kísérlet c lehetséges kimenetellel rendelkezik, és n alkalommal végezzük el azt, természetesen itt is a kísérletek egymástól való függetlenségét kell feltételeznünk. Legyen $x_{ij}=1$, amennyiben az $i.$ kísérlet $(i=1,2,\ldots,n)$ a $j.$ kategóriát $(j=1,2,\ldots,c)$ eredményezi, egyébként $x_{ij}=0$. Ekkor az $x_i=(x_{i1}\quad x_{i2}\quad\ldots\quad x_{ic})$ vektor egy kísérletet jellemez, ahol $\sum_{j=1}^{c}x_{ij}=1$. Jelölje továbbá $n_j=\sum_{i=1}^{n}x_{ij}$ a $j.$ kategória összes bekövetkezésének gyakoriságát. A $(n_1\quad n_2\quad\ldots\quad n_c)$ gyakoriságok multinomiális eloszlást követnek. Az egyes vektorokhoz tartozó valószínűségek meghatározhatók az alábbi függvény segítségével:

$$P\left(X_1=n_1,X_2=n_2,\ldots,X_c=n_c\right)=\left(\frac{n!}{n_1!n_2!\ldots n_c!}\right)p_1^{n_1}p_2^{n_2}\ldots p_c^{n_c} \qquad (7.8)$$

ahol p_j a $j.$ kategória bekövetkezésének valószínűsége. Vegyük észre, hogy a probléma szabadságfoka $c-1$, hisz bármelyik kategória gyakorisága kifejezhető n és a többi gyakoriság lineáris kombinációjaként. Multinomiális eloszlás $X \square Multin(n,\mathbf{p})$ esetén $E(n_j)=np_j, Var(n_j)=np_jq_j, Cov(n_j,n_k)=-np_jp_k$. Természetesen a binomiális és multinomiális eloszlások szoros kapcsolatban állnak, a binomiális a több kategóriát is lehetővé tevő eloszlás speciális, $c=2$ esete.

Béta eloszlás

A béta eloszlás egy két paraméterű (α,β) folytonos eloszlás, mely a $(0,1)$ intervallumon értelmezett. Az értelmezési tartomány lehetővé teszi, hogy az eloszlás értékeit egy ismeretlen valószínűség leírására használjuk a későbbiekben. Az eloszlás sűrűségfüggvénye leírható az alábbi formula segítségével:

$$f(x,\alpha,\beta)=\begin{cases}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}=\dfrac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} & 0\le x\le 1\\ 0 & \text{egyébként}\end{cases} \tag{7.9}$$

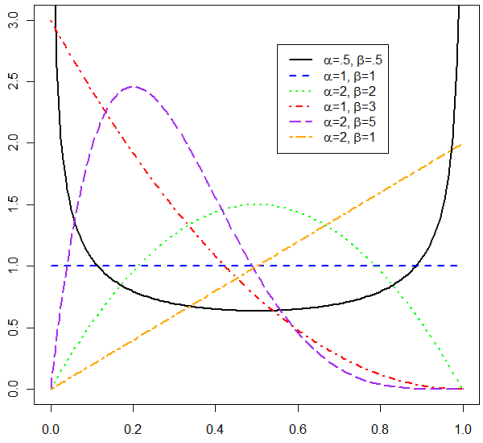
ahol Γ a gamma függvény, B pedig a béta függvény (nem összekeverendő a béta eloszlással). A béta függvény szerepe a normalizálás, azaz annak biztosítása, hogy a sűrűségfüggvény alatti terület egységnyi legyen. Gyakran hívjuk az ilyen jellegű szorzótényezőt normalizáló konstansnak. A normalizáló konstans értéke természetesen csak a paraméterektől függ, az x értékektől nem, azaz az értelmezési tartományon belül állandó értéket vesz fel.

Amennyiben $X \sim Beta(\alpha,\beta)$, úgy

$$E(X)=\frac{\alpha}{\alpha+\beta}, Var(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Mivel a béta eloszlás nem annyira széleskörűen alkalmazott, mint az előzőekben említett diszkrét eloszlások, néhány paraméterkombinációra bemutatjuk a sűrűségfüggvény képét az alábbi, 7-1. ábrán. A különböző paraméterkombinációk igen eltérő eloszlásokat, illetve sűrűségfüggvényeket alkotnak, ami a 0 és 1 közé eső értékek rugalmas modellezését teszi lehetővé.

A 7-1. ábra alapján jól látható, hogy a különböző paraméterkombinációk igen különböző alakú eloszlásokat eredményezhetnek. Amennyiben mindkét paraméter egy alatti értéket vesz fel, a görbe U-alakú. Speciális eset az $\alpha=\beta=1$ paraméterű béta eloszlás, ami egyenletes eloszlást eredményez. Elképzelhető monoton növekvő, illetve csökkenő, lineáris, szimmetrikus és aszimmetrikus görbe is.



7-1. ábra: A béta eloszlás sűrűségfüggvénye néhány kiválasztott paraméterpárral

Dirichlet eloszlás

A Dirichlet eloszlás gyakorlatilag a béta eloszlás többváltozós általánosítása. Viszonya a béta eloszláshoz hasonló a diszkrét esetben látott binomiális-multinomiális viszonyhoz. Legyen k pozitív egész és $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ a paramétervektor. Ekkor, ha $\mathbf{X} = (X_1, X_2, \ldots, X_k) \sim Dir(\boldsymbol{\alpha})$, a sűrűségfüggvény a következő alakban írható:

$$f(\mathbf{x}, \boldsymbol{\alpha}) = \begin{cases} \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{k} x_i^{\alpha_i - 1}, \text{ ha } \quad \forall x_i > 0 \quad \text{és} \quad \sum_{i=1}^{k-1} x_i < 1 \\ 0 \qquad\qquad\qquad\qquad \text{egyébként} \end{cases} \tag{7.10}$$

ahol $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ a változók vektora és $\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}$ normalizáló konstans.

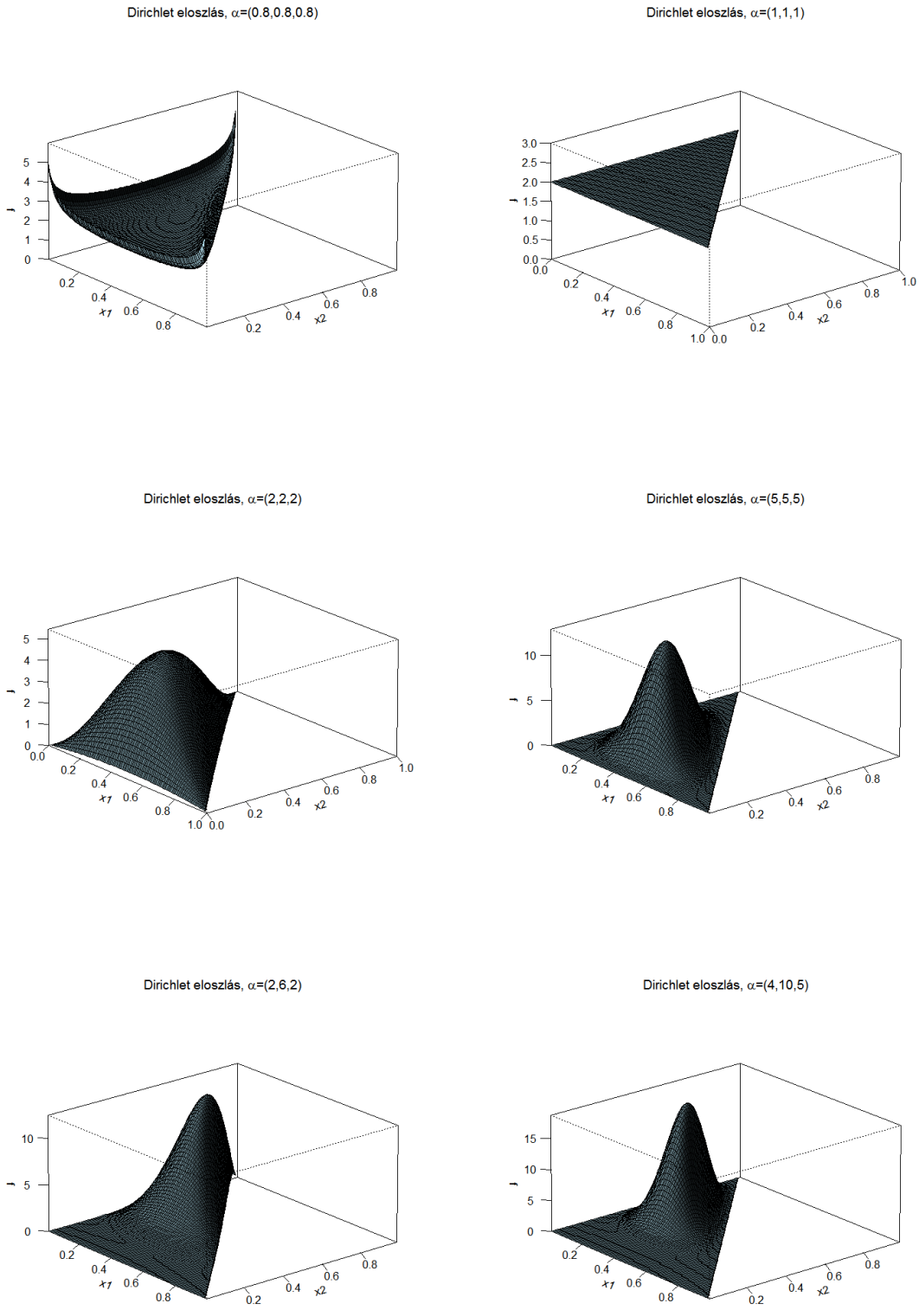
A sűrűségfüggvény ábrázolására a $k = 3$ eset ad még lehetőséget. Ebben az esetben a függvény értelmezési tartománya egy háromszög. A sűrűségfüggvényből adódó felületeket R segítségével[28] ábrázolhatjuk. A dirfelulet() függvény alkalmazásával meg-

[28] A Dirichlet eloszlás ábrázolására szolgáló, általam implementált dirfelulet() függvény a Függelékben megtalálható.

rajzoltunk néhány, különböző paraméterkombinációkhoz tartozó esetet, melyeket az alábbi, 7-2. ábrán mutatunk be. A bal felső felület tipikus az egy érték alatti paraméterkombinációkra (lásd béta eloszlás, $\alpha < \beta < 1$). Amennyiben valamennyi paraméter 1 értéket vesz fel, úgy az eloszlás egyenletes az értelmezési tartomány felett. A gyakorlatban gyakran fordul elő, hogy az azonos $(\alpha_1 = \alpha_2 = \ldots = \alpha_k = \alpha)$ paraméterű eloszlásokat alkalmazzuk, ekkor α jellemző megnevezése koncentrációs paraméter. Erre mutat példát a következő két felület. Amennyiben α növekszik, az eloszlás egyre jobban koncentrálódik a várható érték környékén. Amennyiben a változókat valószínűségekként fogjuk fel, úgy ez azt jelenti, hogy egyre biztosabban vagyunk az értékekben. Az utolsó két sűrűségfüggvény olyan eloszlásokat mutatunk be, melyek esetén a paramétervektor elemei különböző értékeket vesznek fel, ami hatással van a felületek elhelyezkedésére és alakjára is.

Amennyiben $\mathbf{X} = (X_1, X_2, \ldots, X_k) \sim Dir(\boldsymbol{\alpha})$, úgy az eloszlás fő jellemzői: $E(X_i) = \frac{\alpha_i}{\alpha_0}, Var(X_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$, ahol $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.

A marginális eloszlások béta eloszlást követnek, azaz $X_i \sim Beta(\alpha_i, \alpha_0 - \alpha_i)$. A kovariancia pedig a következő módon számítható: $Cov(X_i, X_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$.



7-2. ábra: A Dirichlet eloszlás sűrűségfüggvénye néhány kiválasztott paraméterhármassal

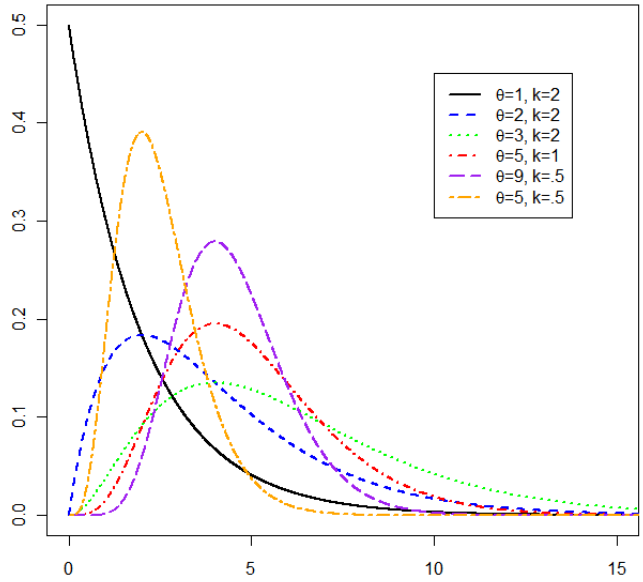
Gamma eloszlás

A gamma eloszlás egy két paraméterrel (k,θ) rendelkező, folytonos eloszlás. Amennyiben k egész, speciális esetként az ún. Erlang eloszlást kapjuk, azaz k független, azonos, θ^{-1} paraméterű exponenciális eloszlás összegét.

Amennyiben $X \sim Gamma(k,\theta)$, úgy a sűrűségfüggvény az alábbi formában írható:

$$f(x,k,\theta) = x^{k-1}\frac{e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}, \qquad k,\theta > 0; x \geq 0 \tag{7.11}$$

ahol Γ továbbra is a gamma függvényt jelöli. A béta eloszláshoz hasonlóan a 7-3. ábra segítségével néhány paraméterkombináció esetén bemutatjuk a gamma eloszlás sűrűségfüggvényét is.



7-3. ábra: A gamma eloszlás sűrűségfüggvénye néhány kiválasztott paraméterpárral

A gamma eloszlást gyakran alkalmazzák például a várakozási idő modellezésére. Amennyiben $X \sim \Gamma(k,\theta)$, úgy $E(X) = k\theta,\ Var(X) = k\theta^2$.

8. A dolgozatban használt fontosabb R programok

1. Equation Section (Next)A mintaátlag eloszlásának szimulációja R segítségével

```
set.seed(1)
iter <- 10000
n <- 100
multi <- c(22,2,2,2,22)
scores <- c(1:length(multi))
mu <- (scores %*% multi)/sum(multi)
sigma <- sqrt((scores^2 %*% multi)/sum(multi)-mu^2)
rawdata <- rmultinom(iter,n,multi)
mu_hat <- (scores %*% rawdata)/n
hist(mu_hat, freq = FALSE, nclass=30, main="", ylab="", xlab="", xlim=c(1,length(multi)),
      ylim=c(0,2.5))
curve(dnorm(x,mu,sigma/sqrt(n)), add = TRUE)
```

2. A variancia változása egyenletes esetben

```
varfelulet <- function (k = 5, x = 3){
m <- seq(1, k)
epsz <- seq(-0.2, .2, by=.01)

varf <- function(m, epsz){
    term1 <- (k*(k+1)*(2*k+1-6*x)-6*k*(m^2-2*m*x))/(6*(k-1))*epsz
    term2 <- ((k+1-2*m)^2*k^2)/(4*(k-1)^2)*epsz^2
    term1 + term2
    }
f <- outer(m, epsz, varf)
nrz <- nrow(f)
ncz <- ncol(f)
jet.colors <- colorRampPalette( c("lightblue", "green") )
nbcol <- 100
color <- jet.colors(nbcol)
zfacet <- f[-1, -1] + f[-1, -ncz] + f[-nrz, -1] + f[-nrz, -ncz]
facetcol <- cut(zfacet, nbcol)
persp(m, epsz, f,
    ylab = "Epszilon",
    zlab = "Variancia hibája",
    xlab = "m",
    col=color[facetcol],
    theta = 50,
    phi = 20,
    r = 50,
    d = 0.1,
    expand = 0.5,
    ltheta = 90,
    lphi = 180,
```

```
    shade = 0.75,
    ticktype = "detailed",
    nticks = 5)
          }
```

3. Cauchy eloszlás az inverz eloszlásfüggvény módszerrel

```
cauchyvel <- function(n, mu, sigma){
          u <- runif(n)
          x <- mu+sigma*tan(pi*(u-0.5))
          return(x)
          }
```

4. Standard normális eloszlás Cauchy eloszlás és elfogadás-elutasítás módszer segítségével

```
n <- 10000
k <- 0
j <- 0
y <- numeric(n)
while (k < n) {
  u <- runif(1)
  j <- j + 1
  x <- qcauchy(runif(1))
  if (dnorm(x)/((sqrt(2*pi)*exp(-.5))*dcauchy(x)) > u) {
    k <- k + 1; y[k] <- x  }     }
```

5. Egyszerű Monte-Carlo integrál (exponenciális eloszlásra)

```
n <- 10000
a <- 2; b <- 4
x <- runif(n, a, b)
MC <- (b-a)*mean(exp(-x))
```

6. Egyszerű Monte-Carlo integrál (normális eloszlás eloszlásfüggvényére)

```
set.seed(999)
n <- 10000
a <- 2
x <- rnorm(n, 0, 1)
MC <- length(x[x<a])/n
```

7. Az MC becslés varianciája és a variancia csökkentése antitetikus változóval

```
iter <- 10000
MCs <- numeric(iter)
MCs_a <- numeric(iter)
n <- 10000
a <- 2; b <- 4

for (i in 1:iter){
```

```
 x <- runif(n, a, b)
 MCs[i] <- (b-a)*mean(exp(-x))
 }

var_g <- (exp(-2*a)-exp(-2*b))/(2*(b-a)) - ((exp(-a)-exp(-b))/(b-a))^2
hist(MCs, freq=FALSE, main="", breaks = 20, xlim=c(exp(-2)-exp(-4)-0.003,exp(-2)-exp(-
      4)+0.003), xlab="")
curve(dnorm(x,exp(-2)-exp(-4),sqrt((b-a)^2*var_g/n)), add = TRUE)

for (i in 1:iter){
 x <- runif(n/2, a, b)
 xx <- a + b - x
 x <- c(x,xx)
 MCs_a[i] <- (b-a)*mean(exp(-x))
 }

hist(MCs_a, freq=FALSE, main="", breaks = 20, xlim=c(exp(-2)-exp(-4)-0.003,exp(-2)-
      exp(-4)+0.003), xlab="")
curve(dnorm(x,exp(-2)-exp(-4),sqrt((b-a)^2*var_g/n)), add = TRUE)
(var(MCs)-var(MCs_a))/var(MCs)
```

8. Az MC becslés varianciájának csökkentése kontroll változó segítségével

```
n <- 10000
a <- 2; b <- 4

x <- runif(n, a, b)
g <- (exp(-x))
g_c <- (exp(-x)+6*exp(-4)*((x-2)/2-.5))
MCs <- (b-a)*g
MCs_c <- (b-a)*g_c

(var(g)-var(g_c))/var(g)
(var(MCs)-var(MCs_c))/var(MCs)
```

9. Fontossági mintavétel (Importance sampling)

```
# Importance sampling
m <- 10000
theta.hat <- var_fg <- nr0 <- numeric(5)

g <- function(x) {
   x^2/sqrt(2*pi)*exp(-x^2/2) * (x >= 1)
   }

#f1
rayleighrand <- function(n = 1, sigma = 1) {
   vel <- runif(n)
   x <- sigma*sqrt(-2*log(vel))
```

```
   return(x)
   }
x <- rayleighrand(m)
fg1 <- g(x) / (x*exp(-x^2/2))
theta.hat[1] <- mean(fg1)
var_fg[1] <- var(fg1)
nr0[1] <- length(fg1[fg1==0])

#f2
x <- rnorm(m,1,1)
fg2 <- g(x) / dnorm(x,1,1)
theta.hat[2] <- mean(fg2)
var_fg[2] <- var(fg2)
nr0[2] <- length(fg2[fg2==0])

#f3
x <- rexp(m)
fg3 <- g(x) / dexp(x)
theta.hat[3] <- mean(fg3)
var_fg[3] <- var(fg3)
nr0[3] <- length(fg3[fg3==0])

#f4
x <- rexp(m)+1
fg4 <- g(x) / dexp(x-1)
theta.hat[4] <- mean(fg4)
var_fg[4] <- var(fg4)
nr0[4] <- length(fg4[fg4==0])

#f5
x <- abs(rnorm(m,0,1))+1
fg5 <- g(x) / (dnorm(x-1,0,1)*2)
theta.hat[5] <- mean(fg5)
var_fg[5] <- var(fg5)
nr0[5] <- length(fg5[fg5==0])

#grafikon
colors <- c("black", "blue", "green", "red", "purple", "orange")
plot(g, ylim=c(0,1), xlim=c(1,6), ylab="", xlab="", lty=1, lwd=4, col=colors[1])
curve(x*exp(-x^2/2), add=TRUE, lwd=2, lty=2, col=colors[2])
curve(dnorm(x,1,1), add=TRUE, lwd=2, lty=3, col=colors[3])
curve(dexp(x), add=TRUE, lwd=2, lty=4, col=colors[4])
curve(dexp(x-1), add=TRUE, lwd=2, lty=5, col=colors[5])
curve(dnorm(x-1,0,1)*2, add=TRUE, lwd=2, lty=6, col=colors[6])
legend(x=4.8,     y=1,     legend=c("g(x)","Rayleigh","N(1,1)","Exp(1)","Exp(1)*","N(0,1)*"),
       lty=c(1:6), col=colors, cex=1, lwd=2)

colors <- c("blue", "green", "red", "purple", "orange")
```

```
curve(g(x)/(x*exp(-x^2/2)), ylim=c(0,2), xlim=c(1,6), ylab="", xlab="", lty=2, lwd=2,
      col=colors[1])
curve(g(x)/dnorm(x,1,1), add=TRUE, lwd=2, lty=3, col=colors[2])
curve(g(x)/dexp(x), add=TRUE, lwd=2, lty=4, col=colors[3])
curve(g(x)/dexp(x-1), add=TRUE, lwd=2, lty=5, col=colors[4])
curve(g(x)/(dnorm(x-1,0,1)*2), add=TRUE, lwd=2, lty=6, col=colors[5])
legend(x=4.8, y=1.5, legend=c("Rayleigh","N(1,1)","Exp(1)","Exp(1)*","N(0,1)*"),
      lty=c(2:6), col=colors, cex=1, lwd=2)

theta.hat
var_fg
nr0
```

10. Rétegző mintavétel négy réteg segítségével

```
n <- 10000
a <- 2; b <- 4
Iter <- 1000
MC <- Int <- rep(0,Iter)

for (i in 1:Iter){

x <- runif(n, a, b)
MC[i] <- (b-a)*mean(exp(-x))

Int1 <- exp(-runif(n/4,2,2.5))
Int2 <- exp(-runif(n/4,2.5,3))
Int3 <- exp(-runif(n/4,3,3.5))
Int4 <- exp(-runif(n/4,3.5,4))
Int[i] <- mean(Int1+Int2+Int3+Int4)/2
}
```

11. Standard normális értékek generálása véletlen bolyongás MCMC módszerrel

```
vbnormal <- function(n=10000, x0=0, a){
x <- numeric(n)
k <- 0
x[1] <- x0
u <- runif(n)
for (i in 2:n) {
  y <- runif(1, x[i-1]-a, x[i-1]+a)
  if (u[i] <= (dnorm(y)/dnorm(x[i-1])))
  x[i] <- y else {
    x[i] <- x[i-1]
    k <- k+1
    }
  }
  return(list(x=x,k=k))
  }
```

```
# plot
set.seed(999)
x1 <- vbnormal(20000,30,.1)
x2 <- vbnormal(20000,30,1)
x3 <- vbnormal(20000,30,10)
burnin <- 2000

plot(x1$x, type="l", ylab="")
plot(x2$x, type="l", ylab="")
plot(x3$x, type="l", ylab="")

plot(x1$x[4501:5500], type="l", ylab="", ylim=c(-4,4))
plot(x2$x[4501:5500], type="l", ylab="", ylim=c(-4,4))
plot(x3$x[4501:5500], type="l", ylab="", ylim=c(-4,4))

# elutasítások
print(c(x1$k,x2$k,x3$k))

# hisztogram
hist(x1$x[(burnin+1):length(x1$x)], freq=FALSE, xlab="", ylab="", main="", xlim=c(-4,4),
      ylim=c(0,0.5), breaks=30)
curve(dnorm(x,0,1),add=TRUE)
hist(x2$x[(burnin+1):length(x1$x)], freq=FALSE, xlab="", ylab="", main="", xlim=c(-4,4),
      ylim=c(0,0.5), breaks=30)
curve(dnorm(x,0,1),add=TRUE)
hist(x3$x[(burnin+1):length(x1$x)], freq=FALSE, xlab="", ylab="", main="", xlim=c(-4,4),
      ylim=c(0,0.5), breaks=30)
curve(dnorm(x,0,1),add=TRUE)

# ACF
acf(x1$x, main="")
acf(x2$x, main="")
acf(x3$x, main="")

# Kvantilisek
z <- -1
int1                                                                                    <-
      length(x1$x[(burnin+1):length(x1$x)][x1$x[(burnin+1):length(x1$x)]<z])/(length(x1$x)
      -burnin)
int2                                                                                    <-
      length(x2$x[(burnin+1):length(x2$x)][x2$x[(burnin+1):length(x2$x)]<z])/(length(x2$x)
      -burnin)
int3                                                                                    <-
      length(x3$x[(burnin+1):length(x3$x)][x3$x[(burnin+1):length(x3$x)]<z])/(length(x3$x)
      -burnin)
print(c(pnorm(z,0,1),int1,int2,int3))
```

```
# inicializálás
n <- 30
k <- 10
prior <- c(1,1) # a béta prior paraméterei
# Prior
curve(dbeta(x,prior[1], prior[2]), xlim=c(0,1), ylim=c(0,8), xlab="", ylab="", lty=1, lwd=3)
# Likelihood
curve(dbeta(x,k+1,n-k+1),add=TRUE, lty=2, lwd=3)
# Poszterior
curve(dbeta(x,prior[1]+k,prior[2]+n-k),add=TRUE, lty=3, lwd=3)
# Jelmagyarázat
legend(x=0.1, y=7.5, c("Prior", "Likelihood", "Poszterior"), lty=c(1:3), lwd=c(3,3,3))

# Intervallum
alfa <- 0.05
# Egyenlő valószínűségek
CS <- qbeta(c(alfa/2,1-alfa/2), prior[1]+k,prior[2]+n-k)
# HPD
require(TeachingDemos)
HPDCS <- hpd(qbeta, shape1=prior[1]+k, shape2=prior[2]+n-k, conf=1-alfa)

# HPD
CS[2]-CS[1]
HPDCS[2]-HPDCS[1]
# A sűrűségfüggvény értéke az intervallumok határain
dbeta(CS,prior[1]+k,prior[2]+n-k)
dbeta(HPDCS,prior[1]+k,prior[2]+n-k)
# Az eloszlásfüggvény értéke az intervallumok határán
pbeta(CS,prior[1]+k,prior[2]+n-k)
pbeta(HPDCS,prior[1]+k,prior[2]+n-k)
# Előrejelzés
m <- 10
y <- seq(0,m)
py <- choose(m,y)*beta(prior[1]+k+y,prior[2]+n-k+m-y)/beta(prior[1]+k, prior[2]+n-k)
plot(y, py, type="h", xlab="y*", ylab="Valószínüség")
# Valószínűségek összege (1-től 5-ig)
sum(py[2:6])
# Arány-béta eloszlás
N <- 10000
delta <- 0.05
alf <- 5
bet <- 800
p <- rbeta(N,alf,bet)
nszuks <- 4*p*(1-p)/delta^2
hist(nszuks, freq=FALSE, xlim=c(0,1/delta^2), xlab="Szükséges mintaelemszám",
    main = bquote(paste("A szükséges elemszám, Beta(",
                    .(alf),",",.(bet),"), ", Delta,"=",.(delta))))
```

13. A Dirichlet eloszlás felületét ábrázoló R függvény

```
dirfelulet <- function (a1 = 1, a2 = 1, a3 = 1){
x1 <- x2 <- seq(0, 1, by=.01)

dirf <- function(x1, x2){
    term1 <- gamma(a1+a2+a3)/(gamma(a1)*gamma(a2)*gamma(a3))
    term2 <- x1^(a1-1)*x2^(a2-1)*(1-x1-x2)^(a3-1)
    term3 <- (x1 + x2 < 1)
    term1 * term2 * term3
    }

f <- outer(x1, x2, dirf)
f[f<=0] <- NA
f[is.infinite(f)] <- NA

persp(x1, x2, f,
    zlim = c(0, max(f, na.rm = TRUE)+1),
    main = bquote(paste("Dirichlet eloszlás, ", alpha ,"=(",.(a1),",",.(a2),",",.(a3),")")),
    col = "lightblue",
    theta = 50,
    phi = 20,
    r = 50,
    d = 0.1,
    expand = 0.5,
    ltheta = 90,
    lphi = 180,
    shade = 0.75,
    ticktype = "detailed",
    nticks = 5)
        }
```

14. Szükséges mintaelemszám eloszlása Likert-skála és különböző poszteriorok és hibahatárok mellett

```
likertn <- function(postvec){
  x <- seq(1:length(postvec))
  (sum(x*x*postvec)-(postvec %*% x)^2)*4/delta^2
}
Reps <- 100000
postvec <- c(1,1,1,1,1)
Dirrn <- rdirichlet(Reps,postvec)
delta <- 0.05
ns <- apply(Dirrn,1,likertn)
hist(ns, main = bquote(paste("A szükséges elemszám, ",
                    alpha
      ,"=(",.(postvec[1]),",",.(postvec[2]),",",.(postvec[3]),",",.(postvec[4]),",",.(postvec[5]),")"
      )),
    freq=FALSE, xlab="Szükséges mintaelemszám")
```

Irodalomjegyzék

- Adams, E. W., Fagot, R. F., Robinson, R. E. (1965): A theory of appropriate statistics, Psychometrika, 30, 99-127
- Agresti, A. (2002): Categorical Data Analysis, Wiley
- Albert, J. H. (2009): Bayesian Computation with R, 2nd edition, Use R!, Springer
- Anderson, N. H. (1961): Scales and Statistics: Parametric and Non-parametric, Psychological Bulletin, 1961, 58, 305-316
- Angus, J. E. (1994): The probability integral transform and related results, SIAM Review, 36, 4, 652-654
- Anwer, K., Hardeo S. (1993): Scales of measurements: An introduction and a selected bibliography, Quality and Quantity, 27, 3, 303-324
- Armstrong, G. D. (1981): Parametric statistics and ordinal data: A pervasive misconception, Nursing Research, 30, 60-62
- Armstrong, G. D. (1984): Letter to the editor, Nursing Research, 33, 54
- Baker, B. O., Hardyck, C. D., Petrinovich, L. F. (1966): Weak Measurements vs. Strong Statistics: An Empirical Critique of S. S. Stevens' Proscriptions on Statistics, Educational and Psychological Measurement, 26, 291-309
- Balogh L. Dániel, Kehl Dániel (2009): Effects of Rounding on Descriptive Statistical Measures, Challenges for Analysis of the Economy, the Businesses, and Social Progress, Szeged
- Bartlett, M. S. (1947): The Use of Transformations, Biometrics, 3, 39-52
- Bayes, T. (1763): An Essay Toward Solving a Problem in the Doctrine of Chances, Philosophical Transactions, 53, 370-418
- Bearden, W. O., Netemeyer, R. G. (1999): Handbook of Marketing Scales, Sage Publications Inc.
- Behan, F. L., Behan, R. A. (1954): Football numbers (continued), American Psychologist, 9, 262-263
- Blair, R. C., Higgins, J. J. (1980): A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions, Journal of Educational Statistics, 5, 309-335
- Blair, R. C., Higgins, J. J. (1985): Comparison of the power of the paired samples t test to that of Wilcoxon's signed-rank test under various population shapes, Psychological Bulletin, 97, 119-128
- Blalock, H. M. (1968): The measurement problem: A gap between the languages of theory and research, In: Methodology in social research, McGraw-Hill, New York, 5-27
- Bogdan, R. J. (1979) (editor): Patrick Suppes, Stanford University, Reidel Publishing, Holland

- Boneau, C. A. (1960): The Effects of Violations of Assumptions Underlying the t-test, Psychological Bulletin, 57, 49-64
- Boneau, C. A. (1961): A note on measurement scales and statistical tests, American Psychologist, 16, 260-261
- Box, G. E. P. (1953): Non-normality and Tests on Variances, Biometrika, 40, 318-335
- Box, G. E. P., Müller, M. E. (1958): A note on the generation of random normal deviates, The Annals of Mathematical Statistics, 29, 610-611
- Büning, H., Trenkler, G. (1978): Nichtparametrische statistische Methoden, Walter de Gruyter, Berlin, New York
- Burke, C. J. (1953): Additive scales and statistics, Psychological Review, 60, 73-75
- Campbell, N. R. (1920): Physics, the elements, London, Cambridge University Press
- Casella, G., Berger, R. L. (2002): Statistical Inference (2nd edition), Duxbury
- Chib, S., Greenberg, E. (1995): Understanding the Metropolis-Hastings Algorithm, The American Statistician, 49, 327-335
- Cochran, W. G. (1947): Some Consequences When the Assumptions for the Analysis of Variance are not Satisfied, Biometrics, 3, 22-38
- Coelho, P. S., Esteves, S. P. (2007): The choice between a five-point and a ten-point scale in the framework of customer satisfaction measurement, International Journal of Market Research, 49, 3, 313-339
- Congdon, P. (2005): Bayesian Models for Categorical Data, Wiley Series in Probability and Statistics
- Coombs, C. H., Raiffa, H., Thrall, R. M. (1954): Some views on mathematical models and and measurement theory, Psychological Review, 61, 2, 132-144
- Cox, E. P. (1980): The Optimal Number of Response Alternatives for a Scale: A Review, 17, 407-422
- Eisenhart, C. (1947): The assumptions underlying the analysis of variance, Biometrics, 3, 1-21
- Falmagne, J. C., Narens, L. (1983): Scales and meaningfulness of quantitative laws, Synthese, 55, 287-325
- Gaito, J. (1960): Scale classification and statistics, Psychological Review, 67, 277-278
- Gaito, J. (1972): An Index of Estimation to Ascertain the Effect of Unequal n on ANOVA F Tests, American Psychologist, 27, 1081-1082
- Gaito, J. (1980): Measurement Scales and Statistics: Resurgence of an Old Misconception, Psychological Bulletin, 87, 564-567
- Galambosné Tiszberger Mónika (2011): A rétegzett mintavételről, Statisztikai Szemle, 89, 9, 909-929

- Gardner, P. L. (1975): Scales and Statistics, Review of Educational Research, 45, 1, 43-57
- Gayen, A. K. (1949): The Distribution of Students's t in Random Samples of Any Size Drawn from Non-normal Universes, Biometrika, 36, 353-369
- Gelfand, A. E., Smith, A. F. M. (1990): Sampling-based approaches to calculating marginal densities, Journal of the American Statistician Association, 85, 398-409
- Gelman, A. (1992): Iterative and Non-Iterative Simulation Algorithms, Computing Science and Statistics, 24, 433-438
- Gelman, A., Carlin, J. B., Stern H. S., Rubin, D. B. (2004): Bayesian Data Analysis, Chapman & Hall/CRC
- Geman, S., Geman, D. (1984): Stochastic relaxation, Gibbs Distributions and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741
- Geweke, J. (2005): Contemporary Bayesian Econometrics and Statistics, Wiley, New Jersey
- Glass, G. V. (1972): Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance, Review of Educational Research, 42, 3, 237-288
- Godard, R. H., Lindquist, E. F. (1940): An Empirical Study of the Effect of Heterogeneous Within-groups Variance upon Certain F-tests of Significance in Analysis of Variance, Psychometrika, 5, 263-274
- Hajdu Ottó, Pintér József, Rappai Gábor, Rédey Katalin (1994): Statisztika I, JPTE-KTK
- Hajdu Ottó (2003): Sokváltozós statisztikai számítások, Központi Statisztikai Hivatal, Budapest
- Hand, D. J. (1996): Statistics and the Theory of Measurement, Journal of the Royal Statistical Society, Series A, 159, 3, 445-492
- Hand, D. J. (2004): Measurement theory and practice: the world through quantification, London, Arnold
- Hartley, S. L., MacLean, W. E. Jr. (2006): A review of the reliability and validity of Likert-type scales for people with intellectual diability, Journal of Intellectual Disability Research, 50, 2, 813-827
- Hastings, W. K. (1970): Monte Carlo sampling methods using Markov chains and their application, Biometrika, 57, 97-109
- Heine, S. J., Lehmann, D. R., Peng, K. (2002): What's Wrong With Cross-Cultural Comparisons of Subjective Likert Scales? The Reference-Group Effect, Journal of Personality and Social Psychology, 82, 6, 903-918
- Helmholtz, H. von (1887): Numbering and measuring from an epsitemological viewpoint, in P. Hertz, M. Schlick: Hermann von Helmholtz: Epistemological writings (72-114), Dordrecht, Holland: Reidel

- Herman Sándor, Pintér József, Rappai Gábor, Rédey Katalin (1999): Statisztika II, JPTE-KTK
- Hitchcock, D. B. (2003): A History of the Metropolis-Hastings Algorithm, The American Statistician, 57, 4, 254-257
- Horváth, J-né. (2001): Bayes tétel alkalmazása a conjoint analízisben, Szakmai Füzetek 10, BGF KKFK, 27-45
- Hölder, O. (1901): Die Axiome der Quantität und die Lehre vom Mass, Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse, 53, 1-64
- Hunyadi László (2001): Statisztikai következtetéselmélet közgazdászoknak, KSH, Budapest
- Hunyadi László, Vita László (2008): Statisztika I-II., Aula Kiadó
- Hunyadi László (2011a): Bayesi gondolkodás a korai modern detektívtörténetekben: Monsieur Lecoq, C. Auguste Dupin és Sherlock Holmes (Kadane, J. B.), Statisztikai Szemle, 89, 2, 238-240
- Hunyadi László (2011b): Bayesi gondolkodás a statisztikába, Statisztikai Szemle, 89, 10-11, 1150-1171
- Kehl Dániel (2007): A szükséges mintaelemszám várható értéke Likert-skálás lekérdezések esetén, Egy életpálya három dimenziója: tanulmánykötet Pintér József emlékére, 63-77, PTE-KTK
- Kehl Dániel (2009): On the History of Measurement, Challenges for Analysis of the Economy, the Businesses, and Social Progress, Szeged
- Kehl Dániel (2011): Skálák és statisztikák: a méréselméletről és történetéről, Statisztikai Szemle, 89, 10-11, 1057-1080
- Kehl Dániel (2012a): Szemelvények a Markov-lánc Monte-Carlo módszerek történetéből (cikk ismertetés), Statisztikai Szemle, 90, 4, 352-354
- Kehl Dániel (2012b): Monte-Carlo módszerek a statisztikában, Statisztikai Szemle, 90, 6, 521-543
- Kehl Dániel, Rappai Gábor (2006): Mintaelemszám tervezése Likert-skálát alkalmazó lekérdezésekben, Statisztikai Szemle, 84, 9, 848-875
- Kish, L. (1989): Kutatások statisztikai tervezése, Statisztikai Kiadó, Budapest
- Knapp, T. R. (1984): Letter to the editor, Nursing Research, 33, 54
- Knapp, T. R. (1990): Treating Ordinal Scales as Interval Scales: An Attempt To Resolve the Controversy, Nursing Research, 39, 2, 121-124
- Komlósi Sándor (2008): Gazdaságmatematika I (második kiadás), PTE KTK, Pécs
- Koop, G. (2003): Bayesian Econometrics, Wiley
- Koop, G., Poirier, D. J., Tobias, J. L. (2007): Bayesian Econometric Methods, Econometric Exercises 7, Cambridge University Press

- Kőrösi Gábor, Mátyás László, Székely István (1990): Gyakorlati ökonometria, Tankönyvkiadó, Budapest
- Kotz, S., Read, C. B., Balakrishnan, N., Vidakovic B. (editors) (2006): Encyclopedia of Statistical Sciences, 16 Volume Set, 2nd edition, Wiley
- Kovács, S., Balogh P. (2009): Bayesi statisztikával becsült nem stacionárius idősorok a sertésárak előrejelzésében, Statisztikai Szemle, 87, 10-11, 1058-1077
- Krantz, D. H., Luce, R. D., Suppes, P., Tversky, A. (1971): Foundations of Measurement, Vol. I: Additive and polynomial representations, New York: Academic Press
- Labovitz, S. (1967): Some Observations on Measurement and Statistics, Social Forces, Vol. 46, Number 2, 151-160
- Labovitz, S. (1970): The assignment of numbers to rank order categories, American Sociologist Review, 35, 515-524
- Laerhoven, H. van, Zaag-Loonen, H. J. van der, Derkx, B. H. F. (2004): A comparison of Likert scale and visual analogue scales as response options in children's questionnaires, Acta Paediatr, 93, 830-835
- Lénárt Imre, Rappai Gábor (2001): Néhány gondolat a varianciabecslés hibahatáráról, Statisztikai Szemle, 79, 7, 613-621
- Likert, R. (1932): A Technique for the Measurement of Attitudes, New York, McGraw-Hill
- Lord, F. M. (1953): On the Statistical Treatment of Football Numbers, The American Psychologist, 8, 750-751
- Luce, R. D. (1996): The Ongoing Dialog between Empirical Science and Measurement Theory, Journal of Mathematical Psychology, 40, 78-98
- Luce, R. D. (2005): Measurement analogies: comparisons of behavioral and physical measures, Psychometrika, 70, 2, 227-251
- Luce, R. D., Krantz, D. H., Suppes, P., Tversky, A. (1990): Foundations of Measurement, Vol. III: Representation, axiomatization and invariance, New York: Academic Press
- Luce, R. D., Narens, L. (1994): Fifteen Problems Concerning the Representational Theory of Measurement, Patrick Suppes: Scientific Philosopher, Ed: Humpreys, P.
- Luce, R. D., Tukey, J. W. (1964): Simultaneous conjoint measurement: a new scale type of fundamental measurement, Journal of Mathematical Psychology, 1, 1–27
- Marcus-Roberts, H. M., Roberts, F. S. (1987): Meaningless Statistics, Journal of Educational and Behavioral Statistics, 12, 4, 383-394
- Matsumoto, M., Nishimura, T. (1998): Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, ACM Transactions on Modeling and Computer Simulation, 8, 1, 3–30

- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. (1953): Equations of state calculations by fast computing machines, Journal of Chemical Physics, 21, 6, 1087-1092
- Michell, J. (1986): Measurement Scales and Statistics: A Clash of Paradigms, Psychological Bulletin, 100, 3, 398-407
- Michell, J. (1986): Measurement Scales and Statistics: A Clash of Paradigms, Psychological Bulletin, 100, 3, 398-407
- Michell, J. (1994): Numbers as Quantitative Relations and the Traditional Theory of Measurement, British Journal for the Philosophy of Science, 45, 389-406
- Michell, J. (1999): Measurement in psychology: critical history of a methodological concept, Cambridge University Press
- Michell, J. (2005): The logic of measurement: a realist overview, Measurement, 38, 285-294
- Michell, J. (2008): Is Psychometrics Pathological Science?, Measurement: Interdisciplinary Research & Perspective, 6, 7-24
- Michell, J., Ernst, C. (1996, 1997): The Axioms of Quantity and the Theory of Measurement, Journal of Mathematical Psychology, 40, 235-252 ill. 41, 345-356
- Narens, L. (1981a): A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness, Theory and Decision, 13, 1-70
- Narens, L. (1981b): On the scales of measurement, Journal of Mathematical Psychology, 24, 249-275
- Narens, L. (2002): Theories of Meaningfulness, Scientific Psychology Series, Mahwah, Lawrence Erlbaum Associates
- Narens, L. (2007): Introduction to the Theories of Measurement and Meaningfulness and the Use of Invariance in Science, Mahwah, Lawrence Erlbaum Associates
- Nelder, J., Wedderburn, R. W. M. (1972): Generalized linear models, Journal of the Royal Statistical Society, Series A (General), 135, 3, 370-384
- Nicholls, M. E. R., Orr, C. A., Okubo, M., Loftus, A. (2006): Satisfaction Guaranteed – The Effect of Spatial Biases on Responses to Likert Scales, Psychological Science, 17, 12, 1027-1028
- O'Brien, R. M. (1979): The Use of Pearson's r with Ordinal Data, American Sociological Review, 44, 851-857
- Peskun, P. H. (1973): Optimum Monte Carlo Sampling Using Markov Chains, Biometrika, 60, 607-612
- Pintér József, Rappai Gábor (2001): A mintavételi tervek készítésének néhány gyakorlati megfontolása, Marketing & Menedzsment, 35, 4, 4-10

- Preston, C. C., Colman, A. M. (2000): Optimal number of response categories in rating scales: reliability, validity, discriminating power and respondent preferences, Acta Psychologica, 104, 1-15
- R Development Core Team (2011): R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/
- Rappai Gábor (2001): Üzleti statisztika Excellel, KSH
- Rappai Gábor, Pintér József (szerk.) (2007): Statisztika, PTE KTK Kiadó, Pécs
- Rizzo, L. M. (2008): Statistical Computing with R, Chapman & Hall/CRC
- Robert, C. P., Casella, G. (2004): Monte Carlo Statistical Methods (2nd edition), New York, Springer
- Robert, C. P., Casella, G. (2010): Introducing Monte Carlo Methods with R, Use R!, Springer
- Robert, C. P., Casella, G. (2011): A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data, Statistical Science, 26, 1, 102-115
- Roberts, G. O., Gelman, A., Gilks, W. R. (1997): Weak convergence and optimal scaling of random walk Metropolis algorithms, Annals of Applied Probability, 7, 110-120
- Russel, B. (1903): The principles of mathematics, London, Cambridge University Press
- Scholten, A. Z., Borsboom, D. (2009): A reanalysis of Lord's statistical treatment of football numbers, Journal of Mathematical Psychology, 53, 69-75
- Siegel, S. (1956): Nonparametric Statistics for the Behavioral Sciences, New York, McGraw-Hill Book Co.
- Stevens, S. S. (1946): On the Theory of Scales of Measurement, Science, 103, 677-680
- Stevens, S. S. (1955): On the Averaging of Data, Science, 121, 113-116
- Student (1908): The Probable Error of a Mean, Biometrika, 6, 1, 1-25
- Suppes, P. (1951): A set of independent axioms for extensive quantities, Portugaliae Mathematica, 10, 163-172
- Suppes, P. (1959): Measurement, empirical meaningfulness and three-valued logic, New York, Wiley, Chapter 6, 129-143
- Suppes, P., Krantz, D. M., Luce, R. D., Tversky, A. (1989): Foundations of Measurement, Vol. II: Geometrical, Threshold and Probabilistic Representations, New York: Academic Press
- Suppes, P., Zinnes, J. L. (1963): Basic Measurement Theory, Handbook of Mathematical Psychology, Vol. I, 3-76, Wiley, New York
- Surányi Bálint, Vita László (1972): A mérési szintek elmélete és értéke a társadalomstatisztikában, Statisztikai Szemle, 7, 731-743

- Szidarovszky, F., Yakowitz, S. (1978): Principles and Procedures of Numerical Analysis, Plenum Press, New York/London
- Theiss E. (1971): A Bayes-módszertan és a statisztikai döntéselmélet alkalmazásai a gazdaságpolitikai modellekben, Statisztikai Szemle, 11, 1087-1106
- Thomas, H. (1982): IQ Interval Scales, and Normal Distributions, Psychological Bulletin, 91, 198-202
- Tierney, L. (1994): Makov Chains for exploring posterior distributions, The Annals of Statistics, 22, 4, 1701-1728
- Townsend, J. T., Ashby, F. G. (1984): Measurement scales and statistics: The misconception misconceived, Psychological Bulletin, 96, 394-401
- Tversky, A. (1967): A General Theory of Polynomial Conjoint Measurement. Journal of Mathematical Psychology, 4, 1–20
- Vargha András (2003): Robusztussági vizsgálatok az egymintás t-próbával, Statisztikai Szemle, 81, 10, 872-890
- Vargha András (2004): A kétszempontos sztochasztikus összehasonlítás modellje, Statisztikai Szemle, 82, 1, 67-82
- Vargha András (2008): Matematikai statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal, Pólya Kiadó
- Varga József (1991): Autoregresszív folyamatok előrejelzése Bayes-módszerrel, Szigma, 23, 1-4, 75-83
- Varga József (2001): A valószínűség-elmélet alapjai, PTE, Pécs
- Vargo, L. G. (1971): Comment on the assignment of numbers to rank order categories, American Sociological Review, 36, 517-518
- Várpalotai, V. (2008): Modern Bayes-i ökonometriai elemzések, Simasági priorok alkalmazása az üzleti ciklusok szinkronizációjának mérésére és az infláció előrejelzése, PhD értekezés, Budapest
- Velleman, P. F., Wilkinson, L. (1993): Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading, The American Statistician, 47, 1, 65–72
- Wickmann, D. (1999): Bayes-statisztika, ELTE Eötvös Kiadó, Budapest
- Wiley, D. A., Bunderson, C. V., Olsen, J. A. (2000): An exploratory study of the statistical and educational implications of violations of the assumptions of parametric analysis techniques, http://opencontent.org/docs/parametric.pdf (2010. február 22.)
- Zimmerman, D. W., Zumbo, B. D. (1989): A note on rank transformations and comparative power of the Student t-test and Wilcoxon-Mann-Whitney test, Perceptual and Motor Skills, 68, 1139-1146
- Zimmerman, D. W., Zumbo, B. D. (1990): The relative power of the Wilcoxon-Mann-Whitney test and Student t-test under simple bounded transformations, Journal of General Psychology, 117, 425-436
- Zumbo, B. D., Zimmerman, D. W. (1993): Is the selection of statistical methods governed by level of measurement?, Canadian Psychology, 34, 390-400