Valószínűségszámítás és statisztika

Kehl Dániel

2021

Pécsi Tudományegyetem
Közgazdaságtudományi Kar

Írta: Kehl Dániel

Lektorálta: Szidarovszky Ferenc
Kiadja: Pécsi Tudományegyetem, Közgazdaságtudományi Kar
Felelős kiadó: Schepp Zoltán, egyetemi tanár, dékán
Nyomda: Sz. K. Stúdió Kft.
Borítót tervezte: Kovács Eszter, esztergraphics@gmail.com

ISBN:
print: 978-963-429-787-1
pdf: 978-963-429-788-8
epub: 978-963-429-789-5

Tartalomjegyzék

Ábrák jegyzéke

Táblázatok jegyzéke

Előszó

Ez a tankönyv elsősorban a Pécsi Tudományegyetem Közgazdaságtudományi Karán oktatott Valószínűségszámítás és statisztika tantárgyhoz készült, de bárki más is haszonnal forgathatja. A könyv nyomtatott, online HTML, pdf és epub formátumban is elérhető. Az online elérhetőség lehetővé teszi a folyamatos fejlesztést, ezért különösen örülök az építő visszajelzéseknek, kritikáknak, javaslatoknak, melyeket a kehld@ktk.pte.hu e-mail címre várok. Nem javaslom a tananyag kinyomtatását, formátumának köszönhetően akár okoseszközről is kényelmesen használható. A könyv ingyenesen elérhető és ez – amennyiben a szerzőn múlik – a jövőben is így marad. A technikai szerkesztésben az R környezetben kifejlesztett bookdown csomag volt a segítségemre.

Az anyag szorosan követi az előadások struktúráját, jelölésrendszerét. Az elméleti anyagot gyakran a megértést segítő példák, vagy számítások szakítják meg, melyek a főszövegtől eltérő formában jelennek meg:

> Itt egy rövid példa, néhány számítás olvasható.

A tankönyv olvasása erősen ajánlott, az azonban nem váltja ki az órákon való részvételt, illetve az otthoni gyakorlást sem, ezért ajánlom a kurzushoz tartozó példatár és képletgyűjtemény gondos tanulmányozását is! Valamennyi fejezet végén Excel tippek találhatók, melyek röviden tartalmazzák az adott fejezetben megismerendő függvényeket, funkciókat. Ezek nem helyettesítik a gyakorlatokon való részvételt, illetve a kapcsolódó videók megtekintését!

Köszönetet mondok a PTE-KTK Közgazdaságtan és Ökonometria Intézet múltbeli és jelenlegi statisztika oktatóinak, PhD hallgatóinak a javaslataikért, példáikért, példamutatásukért, támogatásukért, valamint a lektornak a rendkívül gyors és precíz munkájáért.

a szerző

1. fejezet

A statisztika tárgya, alapfogalmak

Big data, data science, data analytics, adatelemzés, mesterséges intelligencia mind-mind olyan kifejezések, melyekkel gyakran és egyre gyakrabban találkozhatunk. A fenti területek mindegyike szorosan kapcsolódik az egyetemeken hagyományosan statisztika néven oktatott klasszikus tananyaghoz, kiegészülve naprakész informatikai, adatbázis, vagy programozási ismeretekkel. A napjainkban keletkező rengeteg adat gyűjtése, rendszerezése, elemzése, az eredmények értelmezése és bemutatása komoly szakértelmet igényel. Ezek a feladatok hagyományosan a statisztika témakörébe esnek, de az adatok keletkezésének gyorsasága és azok mennyisége miatt a korszerű adatfeldolgozási és informatikai ismeretek egyre kevésbé megkerülhetők. Ebben a fejezetben a legfontosabb statisztikai alapfogalmakat ismerjük meg.

1.1. Sokaság, minta, ismérv

Mint minden tudományág, úgy a statisztika is sajátos nyelvezettel bír, amelynek elsajátítása nélkülözhetetlen a tárgy megismerése során. Ezen a helyen csak a legszükségesebb fogalmakat vezetjük be. Hagyományosan két nagy területet különböztetünk meg, a leíró statisztikát és a következtetéses statisztikát. Leíró statisztikáról beszélünk, ha a vizsgálni kívánt emberek, országok, vállalatok stb. elemzendő adatai teljeskörűen rendelkezésünkre állnak. A vizsgálat tárgyát ilyenkor (teljes) sokaságnak hívjuk, az egyes megfigyeléseket pedig általánosan egyednek. Statisztikai sokaságnak nevezzük tehát a statisztikai megfigyelés tárgyát képező egyedek összességét. A sokaság fogalmilag legfontosabb jellemzői által definiálható. Amennyiben nem teljes körű a megfigyelésünk, akkor mintáról, vagy részsokaságról beszélünk.

Amennyiben egy adott ország népességéről szeretnénk információt szerezni, akkor könnyen tudunk sokasági felmérésre és mintavételre is példát hozni. Sokasági adatokat szerzünk, amikor valamennyi egyedet felmérjük, például (a Magyarországon 10 évente tartott) népszámlálás keretében. Mintára vonatkozó adatokat szerzünk egy-egy közvéleménykutatás lebonyolításakor. A két megközelítés fő előnyei és hátrányai ebből a példából is érzékelhetők: a népszámlálás rendkívül lassú és költséges, az eredmények azonban nem függnek a mintavételt terhelő véletlen hatásoktól.

Következtetéses statisztikáról akkor beszélünk, ha a sokaság egésze nem, csak a sokaságot jól bemutató, reprezentáló minta figyelhető meg, amiből a sokaságra szeretnénk következtetéseket levonni. Az ilyen „jól viselkedő” mintát reprezentatív mintának hívjuk, erre a fogalomra a későbbiekben részletesebben visszatérünk. A mintavétel oka jellemzően az, hogy nincs elegendő erőforrás a teljes sokaság megfigyelésére. A mintába kerülés megfelelő mintaválasztás esetén véletlen eseményként kezelhető, így a következtetéses statisztika alkalmazása előtt szükséges némi valószínűségelméleti ismeretre szert tenni. Ez a gondolatmenet adja a kurzus struktúráját: a leíró statisztika után a valószínűségszámítás, majd a következtetéses statisztika alapjai kerülnek tárgyalásra.

A minta mérete nem garantálja annak reprezentativitását. Ennek tipikus példája az 1936-os amerikai elnökválasztás, ahol Alfred Landon és Franklin D. Roosevelt között folyt a versengés. A Literary Digest által végzett közvéleménykutatás során 10 millió amerikai állampolgárt kérdeztek meg, elsősorban telefonon. Végül mintegy 2,4 millió fős mintával rendelkeztek, ami alapján azt jelezték előre, hogy Landon fölényes győzelmet fog aratni. Ez volt minden idők legdrágább közvéleménykutatása. Egy jóval kisebb, mintegy 50,000 fős minta alapján George Gallup sikeresen jelezte előre Roosevelt győzelmét.

A Literary Digest kudarcát két fő hiba, torzítás okozta: akkoriban a telefonos megkeresés főképp a gazdagabb rétegeket érte el, így hiába volt nagy mintájuk, az nem megfelelően mutatta be a sokaságot (szelekciós torzítás), vagyis nem volt reprezentatív. A másik hiba a nemválaszolók figyelmen kívül hagyása volt. Úgy vélték, azok, akik nem válaszoltak, hasonlóan fognak szavazni, mint azok, akik válaszoltak, ez azonban nyilvánvalóan tévedés volt.

A tanulság: inkább egy gondosan kiválasztott kis méretű minta, mint egy rosszul kiválasztott nagy minta! A mai közvéleménykutatások többnyire néhány ezer (jól kiválasztott) főt kérdeznek meg.

A sokaságnak alapvetően két típusát szokás megkülönböztetni. Eszerint beszélhetünk

- álló (stock) sokaságról és
- mozgó (flow) sokaságról.

Amennyiben a sokaság egy adott időpontra (ezt az időpontot szokás ún. eszmei időpontnak is nevezni) vonatkozó állapotát vizsgáljuk, álló sokaságról beszélünk. A mozgó sokaság folyamatot fejez ki, ebből következően időtartamra értelmezhető.

Álló sokaság például a lakásállomány 2016. október 1-jén: a legutóbbi mikrocenzus adatai alapján 4 405 ezer darab lakás volt ebben az időpontban Magyarországon.

Mozgó sokaság például a 2018-ban felépült lakások száma: a statisztikai adatszolgáltatásból ismerjük ezt az adatot is, 17 681 lakás épült az év folyamán hazánkban.

Segíthet annak eldöntésében, hogy álló vagy mozgó sokaságról beszélünk, hogy az adatok összegzése értelmezhető-e. A lakásállomány esetén például a 2015-re és 2016-ra vonatkozó adatok összege nem ad helyes eredményt a halmozódás miatt: a 2016-ban megfigyelt lakások döntő többsége 2015-ben is már létezett. Ezzel szemben az adott évben felépült lakások száma esetén több év összege is értelmezhető.

Az álló- és a mozgó sokaság természetesen nem független egymástól. A folyamatok eredményeit egy adott időpontban mérve már álló sokaságról beszélhetünk. Például az évente épített lakások száma (flow) meghatározza (a megszűnő lakások számával együtt) egy adott időpont lakásállományát (stock).

A sokaság tartalmazhat

- véges és
- végtelen számú egyedet.

A társadalmi-gazdasági vizsgálatok akkor dolgoznak véges számú egyeddel, ha területileg és időben pontosan körülhatárolhatók a sokaságok. A különféle kísérleti statisztikákban, valamint a folyamatok modellezése során azonban találkozhatunk (legalábbis elvben) végtelen számú egyedet tartalmazó sokasággal is.

A sokaságot, vagy mintát leíró jellemzőket változóknak, vagy ismérveknek hívja a statisztika, a két kifejezést az alábbiakban szinonimaként fogjuk használni. A sokaság egységei (egyedei) az ismérvek hordozói. Adott ismérv lehetséges különböző kimenetelei az ismérvváltozatok.

Ismérv lehet például egy adott népesség esetén az életkor, a nemhez való tartozás, a lakóhely, vállalatok esetén a termelés értéke, vagy az elektromos energia fogyasztás időbeli alakulása. A lakóhely ismérv esetén az ismérvváltozatok, változóértékek például Budapest, megyeszékhely, város, község, stb. lehetnek.

Az ismérvek, vagy változók jellegük szerint három nagy csoportba sorolhatók: megkülönböztethetünk kategóriás (minőségi, kvalitatív), numerikus (mennyiségi, kvantitatív) és időbeli ismérveket.

Ismérvek fajtái:

- minőségi: gazdasági társaságok jogi formája (bt, kft, stb.); nem
- mennyiségi: termelési érték; éves jövedelem; életkor
- időbeli: születési év; cég bejegyzésének éve

Amennyiben csak két ismérvváltozat van, alternatív ismérvről beszélhetünk. A szakirodalomban több elnevezés is elterjedt: kétkimenetelű, bináris vagy dummy változóként is találkozhatunk velük. Alternatív ismérvekkel a későbbiekben még foglalkozunk összetettebb módszerek esetén.

Az alternatív ismérvek jellemzően minőségi ismérvek, például nem (férfi-nő); tőzsdei cég (igen-nem). Bármilyen változó alternatív ismérvvé alakítható, a gazdasági társaságok jogi formáját például vizsgálhatjuk úgy is, hogy kft, vagy nem kft az adott társaság.

A numerikus ismérvek esetén megkülönböztetünk továbbá diszkrét és folytonos mennyiségi ismérveket. A diszkrét mennyiségi ismérvek számlálás útján jönnek létre (azaz jellemzően a természetes számok körét ölelik fel az ismérvváltozataik), míg a folytonos mennyiségi ismérvek mérés útján keletkeznek (a valós számok körében lehetségesek az ismérvváltozatok).

Diszkrét mennyiségi ismérv például a testvérek száma, ami nem lehet tört, míg folytonos mennyiségi ismérvként kezeljük például a testmagasságot, ami – feltéve hogy elég pontos mérőeszközünk van – végtelen sok értéket felvehet egy adott intervallumban.

A változók jellegének ismerete, azonosítása különösen fontos, hiszen az alkalmazandó elemzési módszer kiválasztásánál ennek döntő jelentősége van.

1.2. Mérési skálák

A változó pontos definíciójához tartozik a mérési skála megállapítása is. A szakirodalom négy mérési skálát különít el hagyományosan, melyeken egyre több művelet értelmezhető:

- minőségi ismérvek esetén
  - nominális: az ismérvváltozatok csak az azonosítást szolgálják, amelyek segítségével elvégezhető a jelenségek, folyamatok osztályozása. Az egyedeket aszerint osztályozzuk ezen a skálán, hogy milyen csoportba, kategóriába tartoznak. A skálán a műveletek közül csak az egyenlőség értelmezhető, amely szerint két megfigyelési egység vagy azonos vagy különböző.

Amennyiben a változókat számokkal kódoljuk (pl. 1 - férfi, 2 - nő), a kódolás gyakorlatilag tetszőleges. A számként kódolt ismérvváltozatok közötti műveletek $=, \neq$.

- ordinális: sorrendiségre vonatkozó relációk alapján rangsorba rendezhetők a megfigyelt egyedek. A sorrendi skálán az egyes egyedek egymástól nem feltétlenül egyenlő távolságban helyezkednek el. A számokkal történő kódolás ebben az esetben már nem tetszőleges, annak figyelembe kell vennie az ismérvváltozatok sorrendiségét (pl. 1 - elégtelen, 2 - elégséges, 3 - közepes, 4 - jó, 5 - jeles). Az elvégezhető műveletek köre $=, \neq, <, >$.

- mennyiségi ismérvek esetén

  - intervallum: a sorrend mellett itt a skála bármely két pontja közötti távolság is értelmezhető. Az intervallum-skálának tulajdonsága, hogy nem rendelkezik igazi zéró ponttal. Ez azt jelenti, hogy a nulla pont meghatározása önkényes, a zéró érték nem jelenti a tulajdonság hiányát. Az elvégezhető műveletek köre $=, \neq, <, >, -, +$, azaz az összeg és különbség is értelmezhető.
  - arány: igazi kvantitatív skála, a numerikus értékek úgy jellemzik az objektumok, egyedek elrendeződését, hogy azok egyértelműen behatárolhatók. A skálának igazi zéró pontja van. Mindez azt is jelzi egyben, hogy a nulla érték a tulajdonság hiányát egyértelműen jelzi. A skála értékei multiplikatív módon transzformálhatók, bármely két pont aránya független a mértékegységtől, valamennyi matematikai, statisztikai művelet elvégezhető az arányskála adataival. Az elvégezhető műveletek köre $=, \neq, <, >, -, +, *, /$, azaz a szorzás és osztás is értelmezhető.

Tipikus példák az egyes mérési skálákra:

- nominális: nem, hajszín, rendszám, irányítószám, állampolgárság
- ordinális: iskolai osztályzat, energiatakarékossági osztály, településtípus
- intervallum: Celsius hőmérséklet, tengerszint feletti magasság
- arány: testsúly, testmagasság, árbevétel

1.3. Adatállományok

Az egyes egyedek valamennyi változó szerinti rendszerezett felsorolását adatállománynak nevezzük. Az adatállományokban jellemzően jóval több az egyed, vagy megfigyelés, mint a változó, ezért konszenzusos alapon a sorokban helyezkednek el a megfigyelések, míg a változók az adatállomány oszlopait alkotják. A változókat nagybetűvel, általában X, Y, Z-vel jelöljük. Az adatállományban tárolt adatok jellemzői alapján három nagy csoportot különböztetünk meg:

- idősoros adatállomány: a változó értékeinek időrendi (általában állandó frekvenciájú - ekvidisztáns) felsorolása
- keresztmetszeti adatállomány: több egyed jellemzőinek egy időszakra, vagy időpontra vonatkozó felsorolása
- panel adatállomány: több egyed jellemzőinek több időszakra, vagy időpontra vonatkozó felsorolása

Jelen tananyag elsősorban keresztmetszeti adatállományok elemzésével foglalkozik, idősoros adatállományokkal csak a legegyszerűbb elemzések erejéig. Az idősoros adatállományokkal a Statisztikai modellezés, majd mesterképzésen a különböző ökonometria tárgyakban találkoznak a hallgatók. A panel adatállományok – melyek ötvözik az idősoros és a keresztmetszeti adatok jellemzőit – tárgyalása a doktori képzés tananyagát képezheti.

A keresztmetszeti adatállomány általánosságban a következőképpen néz ki:

1.1. táblázat: A keresztmetszeti adatállomány általános alakja

egyed sorszáma	X	Y	...	Z
1.	X_1	Y_1	...	Z_1
2.	X_2	Y_2	...	Z_2
...	...	...	...	...
i.	X_i	Y_i	...	Z_i
...	...	...	...	...
N.	X_N	Y_N	...	Z_N

Az 1.1. táblázatban néhány, a tananyagban sokat használt jelölést is bevezettünk. A sokaság elemszámát keresztmetszeti adatállomány esetén N jelöli, míg az általános elemet jellemzően az i indexszel jelöljük.

Az 1.2. táblázat a legnagyobb 100 márkaértékű brandet tartalmazó adatállomány első néhány sorát mutatja be a Millward Brown becslése alapján. A változók közül itt a vállalat fő tevékenységi kategóriáját, a márkaérték hozzájárulását az üzleti sikerhez (mrd USD), illetve a székhely régióját tüntettük fel, a teljes lista a BrandZ_2018.xlsx fájlban érhető el.

1.2. táblázat: TOP 100 márkaérték keresztmetszeti adatállomány (részlet)

	márka	kategória	márkaérték	hozzájárulás	régió
1	Google	Technológia	302,063	magas	É-Amerika
2	Apple	Technológia	300,595	magas	É-Amerika

3	Amazon	Kiskeres.	207,594	magas	É-Amerika
4	Microsoft	Technológia	200,987	magas	É-Amerika
5	Tencent	Technológia	178,990	nagyon magas	Ázsia
6	Facebook	Technológia	162,106	magas	É-Amerika
7	Visa	Pénzügy	145,611	nagyon magas	É-Amerika

Az elemzés megkezdése előtt minden esetben tisztázni kell, hogy az adatállomány sokaság vagy minta. Amennyiben az adatállományunk nem sokaságként, hanem mintaként elemzendő, akkor az adatállomány megjelenése hasonló (például egy táblázatkezelőben), de a jelölések tekintetében változás, hogy minta esetén n lesz az elemszám jele, illetve sok esetben a minta elemeit x jelöli X helyett. A 2. és 3. fejezetekben bemutatott módszerek, mutatók sem pontosan ugyanúgy alkalmazandók a két esetben. A mintából való következtetés módszereit a 8. fejezettől kezdődően tárgyaljuk.

Az idősoros adatállományok ránézésre nem sokban különböznek a keresztmetszeti adatállományoktól, a könnyebb megkülönböztetés miatt sok esetben N helyett T jelöli az elemszámot, illetve i helyett t futóindexet használ sok tankönyv. Fontos különbség ugyanakkor, hogy míg keresztmetszeti adatállományok esetén a megfigyelések sorrendje gyakorlatilag irreleváns, addig az idősoros adatállományokban a megfigyelések sorrendje kötött, jellemzően a legrégebbi megfigyeléstől halad a legújabb felé.

1.3. táblázat: Az idősoros adatállomány általános alakja

időszak	X	Y	...	Z
1.	X_1	Y_1	...	Z_1
2.	X_2	Y_2	...	Z_2
...	...	...	...	...
t.	X_t	Y_t	...	Z_t
...	...	...	...	...
T.	X_T	Y_T	...	Z_T

A t futóindexet és a tényleges dátumot/napot/időpontot jellemzően feltüntetjük az idősori értékek mellett, míg az előbbi az elemzést, utóbbi az értelmezést, azonosítást segíti. Az 1.4. táblázatban a magyarországi vendégéjszakák száma látható 2001-2019 között.

1.4. táblázat: Vendégéjszakák száma Magyarországon

t	év	vendégéjszaka
1	2001	18 648
2	2002	18 450

3	2003	18 611
4	2004	18 899
5	2005	19 737
6	2006	19 652
7	2007	20 129
8	2008	19 974
9	2009	18 710
10	2010	19 554
11	2011	20 616
12	2012	21 805
13	2013	22 968
14	2014	24 434
15	2015	25 888
16	2016	27 629
17	2017	29 769
18	2018	31 011
19	2019	31 538

Forrás: KSH

1.4. Sorok, táblázatok

Statisztikai sornak hívjuk az egyedek egy szempont szerinti jellemzését. A statisztikai sor egyrészt a valamely vizsgálandó ismérv szerinti ismérvváltozatokat, valamint a hozzájuk tartozó statisztikai mutatókat tartalmazza. Attól függően, hogy a kiválasztott statisztikai mutató összegezhető-e, megkülönböztetünk

- összehasonlító és
- csoportosító

sorokat.

A legegyszerűbb ilyen statisztikai mutató a gyakoriság, amit egyszerű leszámlálással nyerünk. A következő fejezetekben újabb statisztikai mutatókat fogunk megismerni (pl. átlag, módusz, medián, szórás, stb.), amik segítségével szintén képezhetünk sorokat, most azonban csak a gyakoriságokat (F_j) tartalmazó sorokat tekintjük át. Mivel a gyakoriságok jellemzően összegezhetőek, ezért csoportosító sorokról beszélünk. Amennyiben J csoportot hozunk létre (azaz J különböző ismérvváltozata van a vizsgálandó változónak, melyeket jelöljünk A_j-vel), úgy a gyakorisági sor általános formája:

1.5. táblázat: A gyakorisági sor általános alakja

ismérvváltozat	gyakoriság
A_1	F_1
A_2	F_2
...	...
A_j	F_j
...	...
A_J	F_J
összesen	N

Vizsgálhatjuk a korábban megismert TOP100 adatállományt a vállalat fő profilja szerint, ekkor a gyakorisági sor az alábbi:

1.6. táblázat: A TOP 100 vállalat kategória szerinti gyakorisági sora

kategória	gyakoriság
Pénzügy	21
Technológia	20
Telekommunikációs szolgáltatás	11
Kiskereskedelem	10
...	...
Babaápolás	1
Dohány	1
Szállítás	1
Összesen	100

A gyakorisági sort jelen esetben a gyakoriságok szerint csökkenő sorban közöltük, ami gyakran kényelmesebb, gyorsabb értelmezést, elemzést tesz lehetővé.

Gyakorisági sorokat jellemzően kevés ismérvváltozattal rendelkező változó esetén készítünk, hiszen ekkor lesz átlátható, informatív a sor. A leggyakrabban ezek tehát kategóriás változók vagy diszkrét mennyiségi ismérvek. Lehetséges folytonos mennyiségi ismérvek esetén is gyakorisági sorokat készíteni, ebben az esetben azonban hasznosabb a mennyiségi ismérvet kategorizálni, erről részletesebben az osztályközös gyakorisági sorokról szóló 3.2. fejezetben szólunk.

A statisztikai sorok között szokás megemlíteni a leíró sort, ami különböző, de összefüggő statisztikai adatok egyszerű felsorolását jelenti. Ezek a sorok sok esetben egy elemzés eredményét jelenthetik, ami természetesen egy újabb elemzés bemenő adataként is szolgálhat, például intenzitási viszonyszámokat számíthatunk leíró sorok alapján (1.5.4. fejezet).

Az alábbi táblázat Magyarország néhány makrogazdasági adatát tartalmazza a 2020. évre vonatkozóan. Az adatok forrása a KSH.

1.7. táblázat: Leíró sor Magyarország néhány makrogazdasági adatával

mutató	érték
Házasságkötések száma az év első felében	26793
Élveszületések száma az év első felében	44302
Halálozások száma az év első felében	64657
A GDP (nyers) volumenindexe a II. negyedévben	86,4%
Munkanélküliségi ráta a II. negyedévben	4,6%
Foglalkoztatási ráta a II. negyedévben	68,7%

Statisztikai táblázatnak nevezzük több statisztikai sor összefüggő rendszerét. A statisztikai táblázatok sokkal komplexebb elemzésekre adnak lehetőséget a statisztikai sorokhoz képest. Attól függően, hogy hány sort tartalmaznak, azaz hány változó szerinti információt tartalmaznak, beszélünk dimenziószámról (ilyen értelemben a statisztikai sor egy egydimenziós statisztikai táblázat). A gyakorlatban az átláthatóság miatt leginkább két- és háromdimenziós táblázatokat alkalmazunk. A táblázat típusa a csoportosító sorok száma szerint lehet:

- egyszerű (nincs csoportosító sor);
- csoportosító (pontosan egy csoportosító sor);
- kombinációs (kettő, vagy több csoportosító sor).

A sorokhoz hasonlóan gyakran használunk gyakoriságokat a táblázatok esetén is (azaz a gyakoriságokat nem egy, hanem több szempont együttes figyelembevételével határozzuk meg), a kétdimenziós gyakorisági táblázat általános sémája a 1.8. táblázatban látható (ahol az A változónak J, míg a B változónak M darab különböző ismérvváltozata van):

1.8. táblázat: A gyakorisági táblázat általános alakja

ismérvváltozat	B_1	B_2	...	B_M	összesen
A_1	F_{11}	F_{12}	...	F_{1M}	$F_{1.}$
A_2	F_{21}	F_{22}	...	F_{2M}	$F_{2.}$
...	...	...	...	...	...
A_j	F_{j1}	F_{j2}	...	F_{jM}	$F_{j.}$
...	...	...	...	...	...
A_J	F_{J1}	F_{J2}	...	F_{JM}	$F_{J.}$
összesen	$F_{.1}$	$F_{.2}$	...	$F_{.M}$	N

A gyakoriságokat tartalmazó táblázatokat kontingenciatáblázatnak nevezzük. Az utolsó sorban és oszlopban az összesített gyakoriságokat tüntetjük fel.

A korábbi példánkban a TOP100 brandet csak a vállalati kategória alapján vizsgáltuk. Amennyiben a régiót is figyelembe kívánjuk venni, gyakorisági sor helyett gyakorisági táblázatot kapunk.

1.9. táblázat: A TOP100 vállalat kategória és régió szerinti gyakorisági táblázata

kategória/régió	Ausztrália	Ázsia	É-Amerika	Európa	Összesen
Pénzügy	2	6	11	2	21
Technológia		4	16		20
Telekom.		1	6	4	11
Kisker.		2	6	2	10
...	...	...	...	...	...
Babaápolás			1		1
Dohány			1		1
Szállítás			1		1
Összesen	2	21	58	19	100

Érdekes megfigyelni az észak-amerikai vállalatok túlsúlyát, vagy Európa lemaradását a technológiai szektorban Észak-Amerikával és Ázsiával szemben.

1.5. Viszonyszámok

A viszonyszámok a legegyszerűbb, de egyben fontos elemzési eszközeink, melyek az összehasonlításban, összehasonlíthatóságban segítenek. Az alábbiakban négy fontos viszonyszámmal ismerkedünk meg.

1.5.1. Dinamikus viszonyszámok

A dinamikus viszonyszámok az időbeli összehasonlítások eszközei, tehát alapvetően idősoros adatállományokhoz kötődnek. Két dinamikus viszonyszámot különböztetünk meg.

- A bázisviszonyszám (B_t) valamennyi ($t = 1, 2, ..., T$) idősori értéket egy kitüntetett idősori értékhez hasonlítja, ami a leggyakrabban egyszerűen a legelső vizsgálatba vont megfigyelés, így a tankönyvben a továbbiakban B_t alatt ezt a számítási módot értjük. Megjegyezzük azonban, hogy a bázisviszonyszám kötődhet más megfigyeléshez is.

$$B_t = \frac{Y_t}{Y_1} \qquad \left(B_t = \frac{Y_t}{Y_{\text{bázis}}}\right) \tag{1.1}$$

- A láncviszonyszám (L_t) valamennyi ($t = 2, 3, \ldots, T$) idősori értéket az eggyel megelőző idősori értékhez hasonlítja. Az első megfigyeléshez természetesen nem tudunk láncviszonyszámot számítani, így L_1 nem határozható meg.

$$L_t = \frac{Y_t}{Y_{t-1}} \tag{1.2}$$

A vendégéjszakák számának elemzése történhet bázis- és láncviszonyszámok segítségével. Tekintsük az első, 2001-es évet bázisnak. Valamennyi más évhez kiszámíthatjuk a bázis- és láncviszonyszámokat, álljon itt két év példaként.

A 2009. évhez ($t = 9$) tartozó dinamikus viszonyszámok:

$$B_9 = \frac{Y_9}{Y_1} = \frac{18710}{18648} = 1{,}003 \quad L_9 = \frac{Y_9}{Y_8} = \frac{18710}{19974} = 0{,}937$$

Illetve a 2018. évhez ($t = 18$) tartozó dinamikus viszonyszámok:

$$B_{18} = \frac{Y_{18}}{Y_1} = \frac{31011}{18648} = 1{,}663 \quad L_{18} = \frac{Y_{18}}{Y_{17}} = \frac{31011}{29769} = 1{,}042$$

Elmondhatjuk tehát, hogy a 2009. évben a vendégéjszakák száma gyakorlatilag nem változott 2001-hez képest (0,3%-os növekedés), sőt, elsősorban a válság hatására az előző évhez képest jelentős, mintegy 6,3%-os visszaesést tapasztalunk. Ezzel ellentétben 2001 és 2018 között mintegy kétharmadával, 66,3%-kal nőtt a vendégéjszakák száma. 2017-ről 2018-ra a növekedés mértéke 4,2%-os volt.

Ne feledjük, hogy a 0,937 viszonyszám három különböző módon is interpretálható: beszélhetünk 6,3%-os csökkenésről, 93,7%-ra csökkenésről, esetleg 0,937-szeresre csökkenésről. Hasonlóan egy 1,042 értékű hányados esetén beszélhetünk 4,2%-os növekedésről, 104,2%-ra növekedésről, vagy 1,042-szeresére növekedésről. Gyakran okoz gondot a 2 feletti viszonyszám értelmezése, például a 2,14-es viszonyszám 114%-os növekedést jelent!

A két dinamikus viszonyszám egymással szoros kapcsolatban áll, egyrészt a láncviszonyszám kiszámítható a megfelelő bázisviszonyszámok hányadosaként:

$$L_t = \frac{B_t}{B_{t-1}} \tag{1.3}$$

Másrészt – amennyiben a bázis az első megfigyelt időszak – a bázisviszonyszám kiszámítható a láncviszonyszámok szorzataként:

$$B_t = \prod_{u=2}^{t} L_u \tag{1.4}$$

1.5.2. Megoszlási viszonyszámok

Megoszlási viszonyszámokat jellemzően a csoportosított sokaságban megfigyelt gyakoriságokból számolunk, az adott csoport gyakoriságát hasonlítjuk a sokaság elemszámához, azaz részt hasonlítunk az egészhez. A megoszlási viszonyszámokat jellemzően gyakorisági sorokból, vagy gyakorisági táblázatokból számítjuk. Az előbbi esetben a használt képlet:

$$G_j = \frac{F_j}{\sum_{j=1}^{J} F_j} = \frac{F_j}{N} \tag{1.5}$$

Gyakorisági táblázatok esetén többféle megoszlási viszonyszám is számítható, attól függően, hogy a sorösszesenhez, az oszlopösszesenhez, esetleg a teljes sokasághoz viszonyítunk. Erről részletesebben a későbbiekben lesz szó, hiszen mélyebb elemzések alapjairól van szó.

A TOP100 brand példája esetén különösen könnyű megoszlási viszonyszámokat számítani, hiszen a sokaság elemszáma $N = 100$. Az észak-amerikai régióhoz tartozó megoszlási viszonyszám például

$$G_{\text{ÉA}} = \frac{F_{\text{ÉA}}}{N} = \frac{58}{100}$$

Azaz a 100 legnagyobb márkaértékkel bíró brandek 58%-a észak-amerikai székhellyel rendelkezik. Hasonlóan kiszámítható, hogy a brandek 21%-a elsősorban a pénzügyi szektorban tevékenykedik.

Magyarország népessége a legutóbbi, 2011-es népszámlálás alapján 9 937 628 volt, ebből a nők száma 5 219 149. A nők megoszlási viszonyszáma tehát

$$G_{\text{nő}} = \frac{F_{\text{nő}}}{N} = \frac{5\,219\,149}{9\,937\,628} = 0{,}5252$$

Eszerint a nők a teljes népesség 52,52%-át tették ki 2011-ben, amiből az is következik – mivel a nem alternatív ismérv – hogy a férfiak részaránya 47,48%.

1.5.3. Koordinációs viszonyszámok

Koordinációs viszonyszámokat jellemzően a csoportosított sokaságban megfigyelt gyakoriságokból számolunk, két kiemelt csoport gyakoriságát hasonlítjuk egymáshoz, azaz részt hasonlítunk a részhez. A megoszlási viszonyszámokhoz hasonlóan szintén jellemzően gyakorisági sorokból számítjuk:

$$G_j^k = \frac{F_j}{F_k} \tag{1.6}$$

A koordinációs viszonyszámokra lehet példa az egy kiskereskedelmi cégre jutó technológiai vállalatok száma a 100 legértékesebb brand esetén.

$$G_{\text{tech}}^{\text{kisker}} = \frac{F_{\text{tech}}}{F_{\text{kisker}}} = \frac{20}{10} = 2$$

Azaz minden kiskereskedelmi vállalatra két technológiai vállalat jut.

A koordinációs viszonyszámot gyakran alkalmazza a demográfia, a népesség, népesedés tudománya, amikor a férfi és női népesség arányát elemzi. Az előző példában kiszámítottuk a nők (és a férfiak) megoszlási viszonyszámait, álljon itt most egy példa a koordinációs viszonyszámra.

$$G_{\text{nő}}^{\text{férfi}} = \frac{F_{\text{nő}}}{F_{\text{férfi}}} = \frac{5\,219\,149}{4\,718\,479} = 1{,}106$$

A koordinációs viszonyszám alapján tehát egy férfira 1,106 nő jutott 2011-ben Magyarországon. A leggyakrabban említett példa az ezer férfira jutó nők száma, ami természetesen a kapott viszonyszám 1000-rel való szorzásával adódik. Magyarországon tehát 1000 férfira 1106 nő jutott 2011-ben, azaz nőtöbbletet figyelhettünk meg. Érdemes a számításokat korcsoportonként is elvégezni a KSH adatai alapján!

1.5.4. Intenzitási viszonyszámok

Az intenzitási viszonyszám általában különböző, de egymással kapcsolatban álló statisztikai adatok hányadosa, kiszámítása a nagyon egyszerű

$$I = \frac{A}{B} \tag{1.7}$$

módon történik. Az intenzitási viszonyszámokat több szempont szerint tipizálhatjuk:

- azonos, vagy különböző mértékegységű,

- egyenes, vagy fordított,
- nyers, vagy tisztított

intenzitási viszonyszámok. Az első csoportosítás nem szorul különösebb magyarázatra, pusztán arról van szó, hogy a két adat azonos (ekkor gyakran százalékos, vagy ezrelékes formában értelmezhető a viszonyszám), vagy különböző mértékegységű (ezekben az esetekben könnyebb felfedezni, hogy viszonyszámokról van szó). Egyenesnek nevezünk egy viszonyszámot, ha annak növekedése kedvezőnek tekinthető, míg ellenkező esetben fordítottnak. Természetesen ennek megítélése nem minden esetben egyszerű, erősen szubjektív is lehet.

Egy adott viszonyszám nyers, vagy tisztított mivolta viszonyítás kérdése. Amennyiben található olyan viszonyítási alap (B), amely az adott jelenség pontosabb vizsgálatát teszi lehetővé, és ez az új viszonyítási alap (b) az eredetinek egy részhalmaza, akkor egy nyers intenzitási viszonyszám felbontható egy tisztított és egy megoszlási viszonyszám szorzatára:

$$I = \frac{A}{B} = \frac{A}{b} \cdot \frac{b}{B}$$

ahol tehát $\frac{A}{b}$ a tisztított intenzitási viszonyszám, $\frac{b}{B}$ pedig egy megoszlási viszonyszám (vegyük észre, hogy ugyan más jelöléseket használtunk, mint az előző alpontban, de itt is részt hasonlítunk az egészhez).

A különböző mértékegységű intenzitási viszonyszámra példa lehet az országok népsűrűsége, vagy az egy főre jutó GDP. Az egy főre jutó GDP-re jellemzően úgy tekintünk, hogy minél nagyobb a mutató értéke, annál jobb az adott ország helyzete, ilyen értelemben egyenes intenzitási viszonyszámról beszélünk. Ugyanígy egyenes intenzitási viszonyszám az Egyesült Államokban használt gépkocsi fogyasztását jellemző mérőszám, a MPG (miles per gallon). A mutató azt méri, hogy egy gallon (~3,8 liter) üzemanyag hány mérföldre (~1,6 km) elegendő. Európában a hasonló jelenség mérésére fordított intenzitási viszonyszámot használunk, a fogyasztást liter/100 km mértékegységgel számítjuk, ahol természetesen a minél alacsonyabb érték a kedvező. Ez utóbbi példa azt is mutatja, hogy nem minden esetben 1 a nevezőben lévő mértékegység természetes egysége.

Egy ország népesedését mérhetjük az adott évben születő gyermekek számával, ami önmagában is értelmes mutató, ha például idősoros adatokon csak az adott ország adatait elemezzük (vegyük észre, hogy az országon belüli összehasonlítás is csak akkor korrekt, ha a népesség száma nem változik jelentősen az adott időszakban). Amennyiben azonban országok közötti összehasonlítást kívánunk végezni, önmagában a születő gyermekek száma semmi esetre sem alkalmas, hiszen az összehasonlítandó országok népessége akár teljesen más is lehet, ekkor fordulunk az intenzitási viszonyszámok felé:

$$I = \frac{A}{B} = \frac{\text{gyermekek száma}}{\text{népesség száma}}$$

Ez a mutató már lehetővé teszi az összehasonlítást, de bizonyos jellemzőket nem vesz figyelembe. Néhány országban egészen eltérő a nemek közötti arány (pl. Kína), ezért pontosabb mutatót kapunk, ha csak a nők számával hasonlítjuk össze a születő gyermekek számát. Ekkor már tisztított intenzitási viszonyszámról beszélhetünk:

$$I = \frac{A}{B} = \frac{A}{b} \cdot \frac{b}{B} = \frac{\text{gyermekek száma}}{\text{népesség száma}} = \frac{\text{gyermekek száma}}{\text{nők száma}} \cdot \frac{\text{nők száma}}{\text{népesség száma}}$$

ahol tehát a nyers intenzitási viszonyszámot felbontottunk egy tisztított viszonyszám és egy megoszlási viszonyszám (nők aránya a népességen belül) szorzatára.

Hasonló logika mentén képezhető egy további tisztított intenzitási viszonyszám, ami esetén nem az összes nővel, hanem az ún. szülőképes korú nők számával osztjuk a születő gyermekek számát. Ez a mutató egy fontos demográfiai mutató. Más gazdasági események esetén is érdemes elgondolkodni, hogy mi a viszonyítás megfelelő alapja.

1.6. Grafikus ábrák

A grafikus ábra az elemzések és közlések fontos eszköze. A grafikus ábrák felhívják a figyelmet a statisztikai adatok által reprezentált jelenségek fő vonásaira, a főbb arányokra, tendenciákra, összefüggésekre. A statisztikai munka különböző fázisaiban fontos szerepet töltenek be a grafikus ábrák. Az ábrázolás célja lehet a jelenségek kapcsolatait, okait kereső vagy leíró célú alkalmazás, döntés-előkészítés alátámasztása, közlés és a statisztikai munka belső eszközeként történő alkalmazás.

A grafikus ábrák óriási fejlődésen mentek keresztül az informatika térhódításával, elegendő csak a gyakran használt infografikákra, egzotikus grafikonokra gondolni. Jelen tananyagban csupán a legfontosabb és könnyen elkészíthető (figyelembe véve az alkalmazott szoftvert) ábrák említésére, illetve a legfontosabb alapelvek bemutatására szorítkozunk.

Az egyszerűbb ábratípusok közé tartoznak:

- oszlop- és szalagdiagramok: jellemzően összehasonlításra, vagy az oszlopdiagramot tartam idősorok ábrázolására használhatjuk (ritkábban).
- vonaldiagramok: idősorok ábrázolására alkalmazzuk.

- kördiagramok: gyakoriságok és/vagy megoszlási viszonyszámok ábrázolását teszik lehetővé, de alkalmazásuk nem javasolt, mert az adott kategória nagyságát egy szög reprezentálja, amit az emberi szem nem érzékel olyan pontosan, mint pl. a hosszúságot.
- pont- vagy xy-diagramok: két mennyiségi ismérv kapcsolatának vizsgálatára alkalmazzuk.

A későbbiekben több speciálisabb ábratípust is megismerünk, ezeket itt csak megemlítjük, a későbbi fejezetekben részletesen bemutatjuk őket:

- boxplot (3.1. fejezet)
- hisztogram (3.2. fejezet)
- Lorenz-görbe (3.4. fejezet)

Az ábrakészítés legfontosabb alapelvei szerint az ábra legyen:

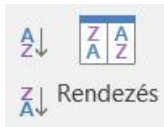
- áttekinthető
- célorientált és homogén
- a lehető legegyszerűbb
- rekonstruálható
- optikailag semleges

1.7. Excel tippek

Hasznos Excel függvények:

- DARAB
- DARAB2
- DARABTELI

Hasznos Excel funkciók:

- Abszolút és relatív hivatkozások =A$7
- Sorbarendezés 

- Szűrő 

Ha numerikus adatokon végzünk szűrést ajánlott a Számszűrők használata. Fontos, hogy a Szűrő alkalmazása csak elrejti a kiszűrt sorokat, nem törli azokat! Ha szűrt adatállományra szeretnénk Excel függvényt alkalmazni, célszerű és nagyon fontos a szűrt adatállományt új munkalapra másolni, amelyen már nem szerepelnek elrejtve a kiszűrt sorok.

- Kimutatás 

A Kimutatás funkció kiválóan alkalmazható gyakorisági sorok és gyakorisági táblázatok elkészítésére. Amennyiben a ismérvváltozatonkénti gyakoriságok helyett más statisztikákat szeretnénk megkapni, ezt az Értékmező-beállítások opcióval tudjuk beállítani.

- Ábrák beszúrása 

2. fejezet

Sokaság leírása egy ismérv alapján

Az előző fejezetben megismertük a statisztika alapvető fogalmait. Ebben (és a következő) fejezetben a sokaság leírásának módszereit ismerjük meg. Mivel a minőségi ismérvek esetén jelenleg nem túl széles az eszköztárunk (a gyakorisági sorok, gyakorisági táblák képzésén túl), ezért a könnyebben kezelhető mennyiségi ismérvek felé fordulunk. Egyelőre célunk egyetlen változó szerinti feldolgozás, a változók közötti kapcsolatok vizsgálata a későbbiekben ezekre az alapokra épülhet.

A fejezetben tehát célunk egy (potenciálisan nagy, akár több millió megfigyeléssel rendelkező) sokaság egyetlen változójának néhány mutatóval történő leírása. Természetesen amennyiben rengeteg számot néhány számban próbálunk tömöríteni, adatvesztés fog történni. Azokat a mutatókat keressük, melyek ennek ellenére jó képet nyújtanak a sokaságunkról.

2.1. Statisztikai műveletek

Ahogy azt már láttuk, a változókat nagybetűkkel jelöljük, amennyiben csupán egyetlen változóról van szó, mint ebben a fejezetben is, azt jellemzően X jelöli. Az alábbiakban néhány egyszerű műveletet, vagy ami fontosabb, azok jelölését mutatjuk be:

- felsorolás (a sokaság i. elemének jele: X_i)
- leszámlálás (a sokaság elemszámának jele: N)
- rangsorolás (a sokaság i. legkisebb elemének jele: $X_{(i)}$)
- rangszám készítése: új változó létrehozása, mely az adott ismérvváltozat emelkedő rangsorban elfoglalt helyét mutatja (R)

- összegzés (az összegzés művelete: $\sum_{i=1}^{N} X_i$, $\sum_{i=1}^{N} X_i$, vagy $\sum X_i$)
- szorzás (a szorzás művelete: $\prod_{i=1}^{N} X_i$, $\prod_{i=1}^{N} X_i$, vagy $\prod X_i$)

2.2. Középértékek

A középértékek az adatok tendenciáját, elhelyezkedését próbálják megragadni. A középértékekkel kapcsolatos lehetséges elvárások:

- közepes: ez alatt azt értjük, hogy a sokaság közepére jellemző, nem pedig szélsőséges értéket keresünk,
- tipikus: legyen a középérték jellemző a sokaság elemeire,
- jól értelmezhető: jelentése legyen intuitív, könnyen megérthető,
- egyszerűen meghatározható legyen,
- robusztus: a középérték a sokaság kis megváltozására ne legyen túl érzékeny, azaz stabil legyen.

A középértékek egyike sem felel meg minden kritériumnak, az egyes mutatók egy vagy több szempont szerint jobb, más szempontok szerint gyengébb jellemzőkkel bírhatnak. Az alfejezet végén visszatérünk az elvárásokra az egyes konkrét középértékekkel kapcsolatosan, értékeljük, melyikre melyik a jellemző.

A középértékeket két nagy csoportra szokás bontani:

- számított (számtani, harmonikus, mértani, négyzetes közép) és
- helyzeti (módusz, medián)

középértékeket különböztetünk meg. Az alábbi alfejezetekben a fenti hat középértéket mutatjuk be részletesebben.

2.2.1. Számtani közép

A számított középértékek közül a számtani közép a legismertebb, gyakran nevezzük számtani átlagnak, vagy egyszerűen átlagnak. A számtani közepet μ-vel jelöljük, a mindenki által ismert és alkalmazott formula segítségével számítjuk ki alapadatokból:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{2.1}$$

azaz (ha a sokaság elemei felsorolásként adottak) a sokasági megfigyelések összege, osztva a sokaság elemszámával. A számtani átlag definíciója tulajdonképpen az, hogy a számtani átlagot az eredeti (átlagolandó) értékek helyébe írva az összeg állandó marad. Azaz ha mind az N megfigyelés μ értéket venne fel, akkor az összeg $N\mu$ értékű, ennek kell megegyeznie az eredeti megfigyelések $\sum X_i$ összegével. Az egyenlőség felírásából egyszerűen adódik a (2.1) egyenlet.

Amennyiben a sokaság nem az elemei felsorolásával adott, hanem például egy diszkrét numerikus ismérv egy gyakorisági táblázata segítségével, akkor a (2.1) formula ún. súlyozott alakját kell alkalmaznunk, ahol a súlyokat a gyakoriságok adják és a súlyok összege a sokaság elemszámát adja vissza.

$$\mu = \frac{\sum_j F_j X_j}{N} \quad \text{ahol } \sum_j F_j = N$$

Számítsuk ki az átlagos brand értéket a TOP100 márka esetén!

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{302{,}063 + 300{,}595 + \cdots + 12{,}456}{100} = 43{,}835$$

Azaz az átlagos márkaérték 2018-ban a TOP100 márkaérték esetén 43,8 milliárd dollár. Más megközelítésben, ha mind a 100 márka esetén 43,8 milliárd dollár lenne a márkaérték, a TOP100-ra összesített márkaérték akkor lenne egyenlő a valóságban megfigyelt összeggel.

2.2.2. Mértani közép

A mértani, vagy geometriai közép azokban az esetekben használatos, amikor nem a megfigyelések összegét szeretnénk, hogy állandó legyen, hanem a szorzatuk, azaz olyan adatok esetén, amikor az összegzés helyett a szorzatszerű összefüggésnek van relevanciája. Elegendő az előző fejezetben megismert láncviszonyszámokra gondolni, ahogy a (1.4) formula esetén is láttuk, szorzatszerűen kapcsolódva a bázisviszonyszámot adják vissza. Logikusan a láncviszonyszámok átlagát mértani középpel számítjuk. A geometriai közép formulája:

$$\mu_g = \sqrt[N]{\prod_{i=1}^{N} X_i} \quad X_i > 0 \tag{2.2}$$

azaz valamennyi pozitív megfigyelésünket összeszorozzuk, majd N-dik gyököt vonunk. A mértani közép esetén fennáll, hogy a mértani közepet az átlagolandó értékek helyébe írva a szorzat állandó marad. Ebben az esetben is előfordulhat, hogy a súlyozott formulára van szükségünk, ha gyakoriságok segítségével állnak rendelkezésre az adatok:

$$\mu_g = \sqrt[N]{\prod_j X_j^{F_j}}$$

A napi, vagy éves hozamok szorzatszerűen kapcsolódnak egymáshoz, hiszen tulajdonképpen az előző időszaki adathoz hasonlító láncviszonyszámokról van szó: ha a mai hozam 2%-os volt, akkor ez azt jelenti, hogy a pénzügyi eszköz árfolyama 2 százalékkal emelkedett a tegnapi értékhez képest, így a mai láncviszonyszám együtthatós alakja 1,02. Legyen egy adott részvény elmúlt 5 évben realizált éves hozama -10%, 15%, 7%, 3%, 14%. Ekkor az átlaghozam

$$\mu_g = \sqrt[N]{\prod_{i=1}^{N} X_i} = \sqrt[5]{0{,}9 \cdot 1{,}15 \cdot 1{,}07 \cdot 1{,}03 \cdot 1{,}14} = 1{,}05393$$

Vegyük észre, hogy a hozamok együtthatós formájával számoltunk, az átlagos hozamra 5,39% adódott. A számtani átlagra ugyanezen adatokból 1,058, azaz 5,8% adódna. Melyik megközelítés ad pontosabb képet? Ennek eldöntésére tegyük fel, hogy 100 dollár értékben vásároltunk részvényt az első év elején. A kezdeti veszteség után emelkedett a részvényünk értéke és az ötödik év végén 130,0368 dollárt ért a befektetésünk. Amennyiben 100 dollárt 5,393%-os éves kamatra helyeztünk volna el a bankban, majd lejáratkor újra és újra befektettünk volna azonos kamatra, pontosan ennyi pénzünk lenne. 5,8%-os éves hozam esetén ez nem áll fenn, a számtani közép itt tehát nem ad pontos képet. A két közép formulával számított eredmény között annál kisebb a különbség, minél közelebb vannak a hozamok a 0-hoz, illetve a láncviszonyszámok az 1-hez. A pénzügy területén gyakran számítanak ún. loghozamokat az elméletileg helyes mértani közép számítás megkerülésére.

2.2.3. Harmonikus közép

Ahogyan azt láttuk, a számtani közép esetén az összeg, míg a mértani közép esetén a szorzat állandó. A harmonikus közép esetén ez a kifejezés a reciprokösszeg, aminek első ránézésre nincs nagy gyakorlati haszna, azonban sok esetben ez a megfelelő átlagformula, ha viszonyszámokból kívánunk középértéket meghatározni. Nagyon gyakori hiba ebben az esetben is a számtani közép alkalmazása, amely akár komoly torzításokat is eredményezhet. A pozitív valós számokra értelmezett harmonikus közép formula:

$$\mu_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \quad X_i > 0 \tag{2.3}$$

illetve súlyozott formájában:

$$\mu_h = \frac{N}{\sum_j \frac{F_j}{X_j}}$$

Tegyük fel, hogy egy gépkocsi egy órán keresztül 10 km/h, egy órán keresztül 20 km/h, egy órán keresztül pedig 30 km/h sebességgel halad. Mekkora az átlagsebessége? Mivel azonos ideig halad mindhárom különböző sebességgel, súlyozás nélkül, egyszerű számtani közepet számítva válaszolhatunk, hogy az átlagsebesség 20 km/h. Számításunkat ellenőrizhetjük is, az első órában 10, a másodikban 20, a harmadikban 30 km-t tett meg, azaz összesen 60 km-t, 3 óra alatt.

Mi a helyzet, ha a feladatot némileg módosítjuk: egy gépkocsi 10 km-t tesz meg 10 km/h, 10 km-t 20 km/h, újabb 10 km-t 30 km/h sebességgel? Ebben az esetben nem alkalmazhatjuk a számtani közepet, a helyes megoldás nem 20 km/h, hanem a harmonikus közép formuláját kell alkalmaznunk:

$$\mu_h = \frac{N}{\sum_j \frac{F_j}{X_j}} = \frac{30}{\frac{10}{10} + \frac{10}{20} + \frac{10}{30}} = \frac{30}{1{,}8333} = 16{,}3636$$

Talán intuitívebb, ha úgy gondoljuk végig: összesen 30 km-t tett meg a gépkocsi, az első 10 km-t 1 óra alatt, a másodikat fél óra alatt, a harmadikat pedig 20 perc alatt, összesen tehát 1 óra 50 percet (1,8333 órát) vett igénybe az út, azaz az átlagsebesség 16,3636 km/h.

A gazdasági életben gyakran van szükség viszonyszámok, hányadosok átlagának meghatározására. A fenti példa emlékeztesse az olvasót arra, hogy ne automatikusan a számtani középre gondoljon! Amennyiben a súlyok a viszonyszám számlálójának mértékegységében vannak megadva, a harmonikus közép a megfelelő átlagformula!

2.2.4. Négyzetes közép

A négyzetes, vagy kvadratikus közép esetén a négyzetösszeg állandó. Inkább csak a teljesség kedve miatt említjük meg ezen a helyen, önmagában ritkán alkalmazott közgazdasági területen, a 2.3. fejezetben azonban alkalmazni fogjuk egy fontos szóródási mérőszám esetén, ahol vissza fogunk utalni rá. A kvadratikus közép formulája:

$$\mu_q = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}} \tag{2.4}$$

amiben tehát a megfigyelések négyzete szerepel, innen származik a közép elnevezése. A súlyozott formula az alábbi:

$$\mu_q = \sqrt{\frac{\sum_j F_j X_j^2}{N}}$$

2.2.5. Helyzeti középértékek

A módusz az egyik helyzeti középérték, definíciója szerint a leggyakrabban előforduló sokasági megfigyelés. A módusz leginkább a kevés ismérvváltozattal rendelkező változók esetén alkalmazható. Ez a középérték kivételes olyan szempontból, hogy minőségi ismérvek esetén is alkalmazható. Folytonos mennyiségi ismérvek esetén gyakran előfordul, hogy minden megfigyelésünk egyedi, ilyen esetekben a módusz nem értelmezhető. Más esetekben előfordulhat, hogy több olyan ismérvváltozat van, amihez ugyanaz a legnagyobb gyakoriság tartozik, ilyen esetekben a módusz nem egyedi, azaz több módusz is elképzelhető.

A medián definíciója szerint középső értéket jelent a sorba rendezett sokaságban. A középső elem megtalálásához más-más képletet használunk, ha N páros (bal oldal), vagy páratlan (jobb oldal).

$$\text{Me} = \frac{X_{(\frac{N}{2})} + X_{(\frac{N}{2}+1)}}{2} \quad \text{Me} = X_{(\frac{N+1}{2})} \tag{2.5}$$

Amennyiben N páros, valójában minden, $X_{(\frac{N}{2})}$ és $X_{(\frac{N}{2}+1)}$ közötti szám medián, csupán matematikai „egyezség”, hogy a két „középső” szám átlagát hívjuk mediánnak.

A medián az eredeti adatok mértékegységében azt az értéket adja meg, amelynél a sokaság elemeinek fele nagyobb, fele kisebb. Újfent fel szeretnénk hívni rá a figyelmet, hogy – a jelöléseknek megfelelően – a sorba rendezett megfigyelések közül kell kiválasztanunk a középső(ke)t a medián meghatározásához!

A TOP100 brand érték esetén a sorba rendezéssel nincs már dolgunk, hiszen rangsorolva látjuk a megfigyeléseket, pont abból a szempontból (márkaérték), ami alapján mediánt szeretnénk számítani. Mivel $N = 100$ páros, ezért

$$\text{Me} = \frac{X_{(\frac{N}{2})} + X_{(\frac{N}{2}+1)}}{2} = \frac{X_{(50)} + X_{(51)}}{2} = \frac{23{,}633 + 22{,}958}{2} = 23{,}2955$$

Azaz az 50. helyezett HSBC és az 51. helyezett YouTube márkaértékének az átlagát vesszük, a medián mintegy 23,3 milliárd dollár. Ez azt jelenti, hogy a TOP100 márkák felének 23,3 milliárd dollár alatti, felének pedig ennél több a márkaértéke.

2.2.6. Középértékek tulajdonságai

Ebben a pontban áttekintjük a leggyakrabban használt középértékek néhány fontos tulajdonságát. A pozitív megfigyelésekre számított középértékek esetén érvényes a

$$\mu_h \leq \mu_g \leq \mu \leq \mu_q \tag{2.6}$$

egyenlőtlenség, azaz azonos adatokon számított harmonikus közép a legkisebb, a kvadratikus pedig a legnagyobb. Természetesen az adatok jellege határozza meg, hogy melyik középértéket alkalmazzuk, ahogy ezt korábban láttuk. Az egyenlőség abban és csak abban az esetben teljesül, ha az átlagolandó értékek megegyeznek.

A számított közepek közül a számtani közép az, amit a legtöbb gazdasági életben előforduló ismérv esetén alkalmazunk. A fejezet elején említett szempontok szerint vizsgáljuk meg tehát a számtani közepet, valamint a két helyzeti középértéket (az értelmezhetőség és a könnyű kiszámíthatóság követelményének mindhárom mutató megfelel). Definíció szerint a medián közepes, a másik két középértékről ez nem feltétlenül mondható el. A módusz tipikus, de a számtani közép és a medián akár olyan értéket is felvehet, ami a sokaságban nem található meg. A két helyzeti középérték robusztus, egy, vagy néhány kiugró érték nem, vagy csak alig befolyásolja értéküket, ami a számtani középről nem mondható el. A fentieket foglalja össze az alábbi táblázat.

2.1. táblázat: Középértékek jellemzése néhány szempont szerint

	közepes	tipikus	robusztus
számtani közép	nem feltétlenül	nem feltétlenül	nem
módusz	nem feltétlenül	igen	igen
medián	igen	nem feltétlenül	igen

Összességében tehát úgy tűnik, hogy a számtani közép, vagy rövidebben átlag nem teljesít túl jól a vizsgált szempontokból, mégis ez a leggyakrabban alkalmazott középérték, amit talán az alábbiakban bemutatott jellemzők némileg magyaráznak:

1. az átlagtól mért eltérések összege zérus: könnyen belátható, hogy ha minden sokasági érték távolságát előjelesen megmérjük az átlagtól, akkor nullát kapunk

$$(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_N - \mu) = \sum_{i=1}^{N} X_i - N\mu = 0 \tag{2.7}$$

2. négyzetes minimum: ez a tulajdonság a következő tényt takarja: ha azt a számot keressük, amelyik (négyzetes értelemben) egyszerre van a legközelebb minden sokasági megfigyeléshez, akkor az pontosan a számtani közép

$$(X_1 - A)^2 + (X_2 - A)^2 + \cdots + (X_N - A)^2 \to \min \Rightarrow A = \mu_X \tag{2.8}$$

3. lineáris transzformálhatóság: amennyiben az X változót lineárisan transzformáljuk (konstanssal eltoljuk (b) és szorozzuk (a)), akkor az új Y változó számtani közepe kiszámítható azonos transzformált segítségével

$$\mu_Y = a\mu_X + b \tag{2.9}$$

Az utolsó tulajdonság a számtani középre, valamint a helyzeti középértékekre is igaz, de a többi számított középértékre nem!

Itt jegyezzük meg, hogy amennyiben a négyzetes távolság helyett az abszolút távolságot minimalizáljuk a (2.8) képlethez hasonlóan, azaz a

$$|X_1 - A| + |X_2 - A| + \cdots + |X_N - A| \to \min \Rightarrow A = \mathrm{Me} \tag{2.10}$$

minimumot keressük, akkor kiderül, hogy a kifejezést minimalizáló érték pontosan a medián.

Összefoglalva: a középértékekkel kapcsolatosan meg kell jegyeznünk, hogy a leggyakrabban használt középérték a számtani közép, vagy átlag, elsősorban kedvező matematikai tulajdonságai miatt. Nem szabad azonban elfeledkeznünk a mediánról sem, amely az átlaggal ellentétben nem érzékeny a kiugró értékekre, így olyan sokaságok esetén, ahol ez elképzelhető, érdemes az átlag mellett a medián közlése is. A két alapvető helyzeti középérték mutató egymáshoz viszonyított elhelyezkedése is fontos jellemzőit mutatja meg a sokaságnak, ahogy azt a későbbiekben látni fogjuk.

2.3. Szóródási mérőszámok

A legfontosabb középértékek áttekintése után figyelmünket egy másik jelenség, a szóródás fogalma felé fordítjuk, ami egyszerűen azt jelenti, hogy a sokasági elemek, megfigyelések egymástól eltérnek, nem azonosak, ennek az eltérésnek a mértékét (a homogenitást, vagy épp a heterogenitást) pedig a statisztika mérni szeretné.

A szóródás mérésére szolgáló néhány mutatót tekintünk át az alábbiakban. A mutatók körét a 3.1. fejezetben tovább bővítjük.

2.3.1. Terjedelem

A legegyszerűbb szóródási mérőszám a terjedelem (R), ami a sokaság legnagyobb és legkisebb értéke közti távolságot méri:

$$R = X_{(N)} - X_{(1)} \tag{2.11}$$

ahol a már megismert jelöléseket alkalmaztuk. A mutató gyorsan meghatározható, azonban kevéssé robusztus, egyetlen kiugró érték nagy hatással van értékére.

A terjedelem könnyedén számítható:

$$R = X_{(N)} - X_{(1)} = 302{,}063 - 12{,}456 = 289{,}607$$

A legkisebb és legnagyobb márkaérték különbsége tehát mintegy 289,6 milliárd dollár, vagy úgy is fogalmazhatunk, hogy bármely két márkaérték közötti a különbség nem nagyobb mint 289,6 milliárd dollár a TOP100 brand esetén.

2.3.2. Átlagos abszolút eltérés

Az átlagos abszolút eltérés (δ) mutatója a szóródás jelenségét már nem az értékek egymástól vett távolsága, hanem egy kitüntetett középértéktől, az átlagtól vett távolság alapján méri, méghozzá ahogy a neve is mutatja, a távolságot abszolút értékként kezelve. Ahogy azt a (2.7) formula alapján láttuk, az abszolút érték nélküli távolságok összege zérus lenne! Az így kialakított képlet:

$$\delta = \frac{1}{N} \sum |X_i - \mu| \tag{2.12}$$

A mutató mértékegysége a megfigyelt adatok mértékegységével egyezik meg, intuitív, könnyen értelmezhető, nem ideális matematikai tulajdonságai (az abszolút érték függvény nehezen kezelhető sok esetben) miatt azonban nem terjedt el a gyakorlatban.

Az átlagos abszolút eltérés kiszámításához valamennyi sokasági értékből ki kell vonnunk az átlagot, majd a különbségek abszolút értékeinek átlagát kell vennünk:

$$\begin{aligned}\delta &= \frac{1}{N} \sum |X_i - \mu| = \\ &= \frac{1}{100}(|302{,}063 - 43{,}83513| + |300{,}595 - 43{,}83513| + \cdots + \\ &+ |12{,}456 - 43{,}83513|) = 33{,}1659\end{aligned}$$

Az egyes márkaértékek tehát átlagosan 33,17 milliárd dollárral térnek el az átlagos márkaértéktől a TOP100 vállalatot tekintve 2018-ban (abszolút értelemben).

Az átlagos abszolút eltérés esetén jogosan vetődik fel (lásd (2.10)), hogy miért az átlagtól, miért nem a mediántól vett eltérést vizsgáljuk. A statisztikában létezik természetesen ez a mutatószám is, ahogy az átlagos eltéréseket bármely középértéktől,

akár a módusztól is vizsgálhatnánk, ezek azonban jóval ritkábban alkalmazott mutatószámok, mint az átlagos abszolút eltérés.

2.3.3. Szórás, variancia

Az abszolút érték függvény helyett az átlagtól való eltérések előjelének kezelésére a négyzetre emelés is hatékony. A mutatót hívhatnánk átlagos négyzetes eltérésnek is, de olyan gyakran alkalmazott mutató (sokak szerint a statisztika nem az átlagok, hanem a szórások tudománya), hogy rövidebb nevet kapott: szórás (σ). A képlete:

$$\sigma = \sqrt{\frac{1}{N}\sum (X_i - \mu)^2} \tag{2.13}$$

A négyzetre emelések, majd a gyökvonás miatt a mutató mértékegysége szintén megegyezik az eredeti sokasági változó mértékegységével.

Gyakran dolgozik a statisztika a szórás négyzetével, a varianciával (σ^2), ami önmagában nem értelmezhető, a gyökvonás hiánya miatt a mértékegysége sem releváns. Sok összefüggésben azonban a varianciák szerepelnek, így a fogalom megismerése már most fontos.

A variancia kiszámításához valamennyi sokasági értékből ki kell vonnunk az átlagot, majd a különbségek négyzeteinek átlagát kell vennünk. A szórás kiszámításához gyököt kell vonnunk a varianciából.

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum (X_i - \mu)^2 = \\ &= \frac{1}{100}((302{,}063 - 43{,}83513)^2 + (300{,}595 - 43{,}83513)^2 + \cdots + \\ &+ (12{,}456 - 43{,}83513)^2) = 2911{,}81\end{aligned}$$

A variancia értéke tehát 2911,81, amit nem értelmezünk, a gyöke $\sigma = 53{,}9612$.

Az egyes márkaértékek tehát átlagosan 53,96 milliárd dollárral térnek el az átlagos márkaértéktől a TOP100 vállalatot tekintve 2018-ban (négyzetes értelemben). Szembetűnő az átlagos abszolút eltérés és a szórás nagyon hasonló értelmezése (de különböző a számszerű érték). A két mutató közötti különbséget a távolság mérésének módja adja.

Már ezen a helyen ki szeretnénk emelni, hogy a (2.13) képlet csak az alapsokaságból számított szórás esetén helytálló. Ahogyan azt a 8.2.3. fejezetben látni fogjuk, a mintából számított szórás képletének logikája a fentitől némileg eltér. A legtöbb szoftverben ezért a szórás kiszámításához két különböző képlet is tartozik.

2.3.4. Relatív szórás

Amint azt láttuk, a szórás mértékegysége megegyezik az eredeti adatokéval, így ha két sokaságot, vagy egy sokaságot két változó heterogenitása, szóródása alapján szeretnénk összehasonlítani, akkor az eltérő mértékegységek, vagy egyszerűen a változók különböző nagyságrendje miatt ezt nem tehetjük meg. A relatív szórás éppen arra szolgál, hogy összehasonlíthatóvá tegyük ezeket a mutatókat. A relatív szórás a sokaság szórását a sokasági átlaghoz viszonyítja, általában százalékos formában értelmezzük, illetve feltesszük, hogy az átlag nem 0.

$$V = \frac{\sigma}{\mu} \tag{2.14}$$

A relatív szórás mutatóját a korábbi eredmények alapján számítjuk ki.

$$V = \frac{\sigma}{\mu} = \frac{53{,}9612}{43{,}83513} = 1{,}231$$

A relatív szórás mutatója 1,23, vagy 123%. Az egyes márkaértékek az átlagos márkaértéktől tehát átlagosan 123%-kal térnek el. Amint azt a példa is mutatja, a mutató értéke akár 1 feletti is lehet, ebben az esetben nagyon erős szóródást, azaz heterogenitást figyelünk meg.

2.3.5. Szórás tulajdonságai

Ebben az alpontban a szórás – mint a leggyakrabban alkalmazott szóródási mutató – néhány tulajdonságát, illetve a szóráshoz és a varianciához kapcsolódó összefüggést, fogalmat mutatunk be.

1. A variancia számlálóját eltérés-négyzetösszegnek nevezzük, angol neve sum of squares, gyakori rövidítése SS. Az eltérés-négyzetösszeg fontos szerepet fog játszani a későbbi tanulmányok során, itt csupán annyit jegyzünk meg, hogy a variancia számlálójában lévő zárójelek felbontásával az alábbi formulát kapjuk:

$$SS = \sum_{i=1}^{N} (X_i - \mu)^2 = \sum X_i^2 - N\mu^2 \tag{2.15}$$

2. A (2.15) összefüggésből egyszerűen adódik az ún. variancia átlagfelbontás képlete. A variancia tehát kifejezhető a megfigyelések négyzetes közepének és számtani közepének segítségével, méghozzá a két közép négyzetének különbségeként. A képlet gyakorlati jelentőségét többek közt az adja, hogy számításigénye elmarad a (2.13) formuláétól, így a variancia (és a szórás) meghatározása gyorsabb.

$$\sigma^2 = \frac{SS}{N} = \frac{\sum_{i=1}^{N} X_i^2}{N} - \mu^2 = \mu_q^2 - \mu^2 \tag{2.16}$$

3. Vizsgáljuk meg az átlaghoz hasonlóan a változó lineáris transzformációjának hatását a varianciára! Mivel a b konstanssal való eltolás a számtani átlagot is eltolja, pontosan b-vel (lásd (2.9)), az átlagtól való átlagos eltérések (szórás) nem változnak. Ezzel szemben az a-val való szorzás a^2-szeresére változtatja a varianciát és $|a|$-szeresére a szórást. Képletszerűen:

$$\text{ha } Y = aX + b, \text{ akkor } \sigma_Y^2 = a^2\sigma_X^2 \quad (!) \tag{2.17}$$

2.4. Sokasági elhelyezkedés

Gyakori feladat, hogy egy adott megfigyelés sokasági elhelyezkedését kell vizsgálnunk, vagy egy megfigyelés két különböző változó szerinti pozícióját összehasonlítanunk. A leggyakrabban alkalmazott két módszert mutatjuk be röviden.

A z-score az adott egyed átlagtól vett távolságát méri szórásokban, gyakran standardizált értéknek, vagy magyarul sztenderdizált, vagy z-értéknek hívjuk, kiszámítása

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{2.18}$$

módon történik. Amennyiben értéke

- 0, akkor a megfigyelésünk épp átlagos,
- pozitív, akkor a megfigyelésünk átlag feletti,
- 1, akkor a megfigyelésünk épp egy szórásnyival van az átlag felett,
- negatív, akkor a megfigyelésünk átlag alatti,
- -1, akkor a megfigyelésünk épp egy szórásnyival van az átlag alatt.

A sokaság összes megfigyelését sztenderdizálva (a hozzájuk tartozó z-értékeket kiszámítva) egy olyan új, sztenderdizált változót kapunk, amelynek átlaga 0, szórása 1, ezért a sztenderdizált változók különösen alkalmasak arra, hogy két sokaságban található értékek elhelyezkedését hasonlítsuk össze. A z-értékeket gyakran alkalmazzuk kiugró, szokatlan értékek keresésére is. Hüvelykujjszabályként a -3 alatti és 3 feletti értékeket kiugró értéknek tekinthetjük.

Egy másik gyakran alkalmazott transzformáció a minmax normalizálás, azt mutatja meg, hogy az adott érték mennyire esik közel a minimumhoz, illetve a maximumhoz, más szóval a sokaság teljes terjedelmén belül hol helyezkedik el. Kiszámítása a

$$Y_i = \frac{X_i - \min(X)}{\max(X) - \min(X)} = \frac{X_i - X_{(1)}}{X_{(N)} - X_{(1)}} \tag{2.19}$$

formulával lehetséges. A formula valamennyi megfigyelést a 0-1 zárt intervallumra transzformál, az 1 közeli értékek a maximumhoz vannak közel, míg a 0 közeliek a minimumhoz.

Vizsgáljuk meg, hogy az Apple márkaértéke hol helyezkedik el a TOP100-on belül. A hozzá tartozó z-érték

$$Z_i = \frac{X_i - \mu}{\sigma} = \frac{300{,}595 - 43{,}83513}{53{,}9612} = 4{,}758$$

azaz az Apple márkaértéke mintegy 4,76 szórásnyival az átlagos márkaérték felett van. Ez azt jelenti, hogy kiugró, nem szokásos értékek közé sorolhatjuk. A minmax normalizált értéke

$$Y_i = \frac{X_i - \min(X)}{\max(X) - \min(X)} = \frac{300{,}595 - 12{,}456}{302{,}063 - 12{,}456} = 0{,}9949$$

azaz nagyon közel van az egyes, maximális értékhez.

2.5. Excel tippek

Hasznos Excel függvények:

- Középértékek: ÁTLAG, MÉRTANI.KÖZÉP, HARM.KÖZÉP, MÓDUSZ.EGY, MEDIÁN

A MÓDUSZ.TÖBB Excel függvény képes a móduszt abban az esetben is megtalálni, ha az nem egyedi, azaz több módusz is van. A MÓDUSZ.EGY függvényt egyszerűbb használni, azonban csak azt a móduszt adja vissza, amelyik az adatállományban „előbb” fordul elő, azaz az eredmény függ a megfigyelések sorrendjétől. A MÓDUSZ.TÖBB egy ún. tömbfüggvény, amit az ENTER helyett a CTRL+SHIFT+ENTER billentyűkkel kell meghívni.

- Szóródás: MIN, MAX, ÁTL.ELTÉRÉS, SZÓR.S, VAR.S
- Sokasági elhelyezkedés: NORMALIZÁLÁS (a függvény a z-score számítását végzi el)

Hasznos Excel funkciók:

- Az Adatelemzés menü Leíró statisztika eszköze 

Az Adatelemzés menüben több hasznos elemző eszköz található. Ezt az Adatok fülön érhetjük el, ehhez a „Beállítások/Bővítmények” opciót választva aktiválnunk kell az Analysis ToolPak bővítményt. A Leíró Statisztika eszköz Összesítő Statisztika beállítása több ebben és a következő fejezetben tárgyalt mutató értékét közli. A szórásnál és a varianciánál nem pontosan a sokasági értékeket adja meg, így alkalmazása csak kellő körültekintéssel ajánlott!

3. fejezet

Sokasági eloszlás alakja

Az előző fejezetben megismertük a főbb középérték és szóródás mutatókat. Most azt vizsgáljuk meg, hogy ezeken a mutatókon túl milyen más módszerek vannak a sokaság leírására, az egyes megfigyelések hogyan helyezkednek el az átlag körül, az átlagnál alacsonyabb, vagy magasabb értékek a jellemzőbbek, hogyan helyezkednek el az adatok a minimum és a maximum között. A hasonló kérdéseket összefoglalóan a sokasági eloszlás alakjának vizsgálatával végezzük el.

Bizonyos sokaságok szimmetrikusak, abban az értelemben, hogy az átlag alatt és felett a megfigyelések közel hasonló módon helyezkednek el. Ez alatt azt értjük, hogy ha egy populációban az átlagos testmagasság 175 centiméter, akkor a 165 és 185 centiméter körüli egyedek körülbelül azonos számban fordulnak elő, de ugyanez igaz a 160 és 190 centiméter körüli egyedekre is. Ezt a tulajdonságot az eloszlás szimmetriájaként fogjuk említeni. A szimmetria egyik jellemzője, hogy az átlag és a medián közel helyezkednek el egymáshoz.

Más sokaságok esetén a fenti megfigyelés nem igaz, például a jövedelem esetén van egy erős alsó korlát (pl. a mindenkori minimálbér), de a felső határ nem ilyen erős, néhányan extrém (nagy) jövedelemmel rendelkeznek. Ahogy azt a 2.2.6. fejezetben láttuk, a néhány magas megfigyelésre az átlag érzékeny, míg a medián nem, ezért az ilyen eloszlások esetén az átlag jelentősen meghaladja a medián értékét. Ezt az esetet pozitív ferdeségnek, vagy jobboldali aszimmetriának nevezzük ebben a tankönyvben. Természetesen a másik irányú aszimmetria is elképzelhető, amikor a szimmetria hiányát alacsony, vagy lefelé kiugró értékek okozzák. Ezt negatív, vagy baloldali ferdeségnek nevezzük.

A sokaságokat jellemezhetjük az alapján is, hogy a megfigyelések mennyire tömörülnek az átlag közelében, vagy esetleg mennyire egyenletes az megoszlás a minimum és a maximum között. A jelenséget csúcsosságnak/lapultságnak nevezzük.

A sokasági eloszlás alakját három különböző megközelítéssel vizsgáljuk a 3.1., 3.2. és 3.3. alfejezetekben, ahol részletesebben kitérünk arra, hogy az adott módszerrel hogyan

mérhető a szimmetria és a csúcsosság. Ehhez előre definiálunk öt különböző sokaságot, amiket az alfejezetekben tanult módszerekkel vizsgálunk:

1. szimmetrikus, nem túl lapos, nem túl csúcsos
2. szimmetrikus, lapult eloszlás
3. szimmetrikus, csúcsos eloszlás
4. jobboldali aszimmetria, csúcsos eloszlás
5. baloldali aszimmetria, csúcsos eloszlás

A fejezetet egy speciális témakör, a koncentráció jelensége, mérési módszerei zárják.

3.1. Kvantilisek

A kvantilisek egyik tagjával, a mediánnal már találkoztunk a 2.2.5. fejezetben, amit sokasági felezőpontként definiáltunk. A sokaságot azonban nem csak két, hanem akár több egyenlő részre tudjuk bontani, ezeket az általános osztópontokat hívjuk kvantiliseknek. Attól függően, hogy hány részre osztják a sokaságot

- tercilis (harmadoló - jele $T_t \quad t = 1, 2$),
- kvartilis (negyedelő - jele $Q_q \quad q = 1, 2, 3$),
- kvintilis (ötödölő - jele $K_k \quad k = 1, 2, 3, 4$),
- decilis (tizedelő - jele $D_d \quad d = 1, 2, \ldots, 9$),
- percentilis (századoló - jele $P_p \quad p = 1, 2, \ldots, 99$)

névvel illetjük a kvantiliseket. A tercilis tehát például három egyenlő részre bontja a sokaságot, abban az értelemben, hogy mind a három harmadban a sokaság (közel) azonos számú eleme található. A sorba rendezett értékek közül tehát egy adott algoritmus alapján azokat az értékeket határozzuk meg, melyeknél a sokasági egyedek harmada kisebb, illetve nagyobb. Tercilisből tehát kettő van, jelük T_1 és T_2. A leggyakrabban alkalmazott kvantilisek talán a kvartilisek, azaz a negyedelőpontok. Definíció szerint Q_1 azt az értéket jelöli, amelynél a megfigyelések negyede kisebb, háromnegyede pedig nagyobb, elnevezése első, vagy alsó kvartilis. A második kvartilis egyben a medián, a felső, vagy harmadik kvartilis jele Q_3. Hasonlóan definiálhatunk ötödölő és tizedelő osztópontokat, amiből rendre 4, illetve 9 értelmezhető. A legáltalánosabb (hiszen valamennyi fent említett kvantilis előállítható speciális eseteként) nevesített kvantilis a percentilis, azaz századoló pont, ennek segítségével tetszőleges százalékos felosztás készíthető a sokaságról.

Fontos tehát, hogy két szomszédos kvantilis között mindig a sokaság adott hányada (harmada, negyede, ötöde, stb. található), az azonban nem feltétlenül igaz (sőt, szinte soha nem igaz!), hogy a szomszédos kvantilisek közötti távolság azonos. Ne felejtsük el, a fejezetben a cél a sokasági értékek elhelyezkedésének vizsgálata, az osztópontok elhelyezkedése, távolsága pontosan az az információ, ami segít ezt megérteni.

Amennyiben „kézzel" szeretnénk például kvartiliseket meghatározni, úgy nagyon hasonlóan járunk el, mint a medián (lásd 2.2.5. fejezet) esetén, az alsó kvartilis esetén a sorba rendezett sokaság $\frac{N+1}{4}$-edik, míg a felső kvartilis esetén a $\frac{3\,(N+1)}{4}$. értéket keressük meg. Amennyiben a képletek egész sorszámot adnak meg, úgy meg is találtuk a keresett értékeket. Ahogy már a medián esetén is megkülönböztettük a páros és páratlan elemszámú sokaságokat, úgy ebben az esetben a néggyel való osztási maradék alapján négy esetet is megkülönböztethetnénk. A helyzet tovább bonyolódik, ha például percentilist szeretnénk meghatározni.

A legegyszerűbb, ha az alábbi (némileg önkényes) elvet követjük tetszőleges kvantilis közelítő kiszámítása esetén:

- amennyiben a keresett érték sorszáma egész, úgy egyszerűen válasszuk ki az adott értéket a sorba rendezett sokaságban,
- amennyiben a keresett sorszám ,5-re végződik, vegyük a két szomszédos megfigyelés átlagát (ahogy a mediánnál is tettük),
- minden más esetben kerekítsük a sorszámot a legközelebbi egészre és az annak megfelelő elemet válasszuk a sokaságból.

Egy másik lehetőség az lenne, ha arányosítanánk a két szóban forgó megfigyelés közötti távolságot. Látható, hogy több különböző elv, algoritmus van a kvantilisek kiszámítására, ennek megfelelően Excelben két, míg az R-ben 9, némileg eltérő eredményt adó függvény került implementálásra, a köztük lévő különbség taglalása meghaladja tananyagunk kereteit.

Tegyük fel, hogy a kvartiliseket keressük egy $N = 11$ elemű sokaságban. Ekkor az alsó kvartilis sorszáma $\frac{N+1}{4} = 3$, azaz a sorba rendezett sokaság 3. eleme, a 6. eleme a medián, és minthogy $\frac{3\,(N+1)}{4} = 9$, a 9. elem a felső kvartilis. Újra nyomatékosítani szeretnénk, hogy nem a kvartiliseket és a mediánt számoltuk ki, csupán a keresett sokasági elemek sorszámát!

Legyen most a sokaság $N = 20$ elemű. Ekkor a fenti szabályok szerint $\frac{N+1}{4} =$ 5,25, ami kerekítve 5, azaz az ötödik elemet keressük, míg $\frac{3\,(N+1)}{4} = 15{,}75$, azaz a 16. elemet keressük és azonosítjuk őket kvartilisként.

Ne felejtsük el, hogy az egyszerű kerekítés helyett más szabályok is vannak, azaz ha szoftverrel számítjuk a kvantiliseket, némileg eltérő eredményeket kaphatunk!

Amint azt már említettük, gyakran alkalmazzuk a kvartiliseket, egyrészt azért mert a kvartilisek távolságára épül egy szóródási mérőszám, az interkvartilis terjedelem, másrészt egy gyakran használt statisztikai ábra, a boxplot is felhasználja az értékeket.

Az interkvartilis terjedelem kiküszöböli a terjedelem mutató azon hibáját, hogy túl érzékeny a kiugró adatokra, az interkvartilis terjedelem ugyanis csak a sokaság középső 50%-ának terjedelmét vizsgálja:

$$\mathrm{IQR} = Q_3 - Q_1 \tag{3.1}$$

Az interkvartilis terjedelem mutatójának mértékegysége megegyezik az eredeti adatok mértékegységével. Amennyiben a sokaság nem szóródik, azaz minden eleme megegyezik, az IQR mutató értéke 0 (az állítás fordítva nem feltétlenül igaz).

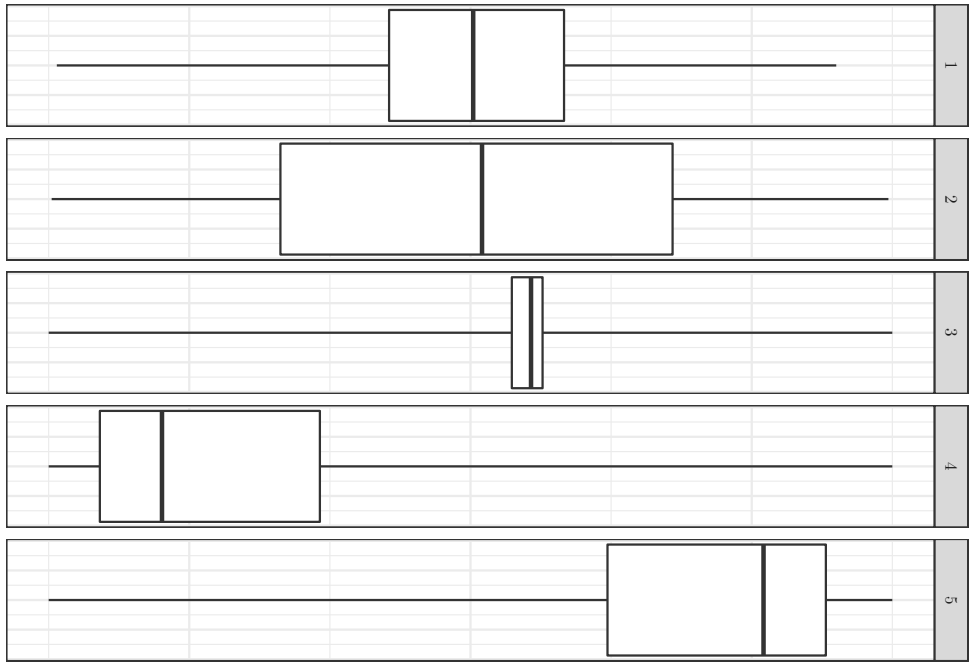
A boxplot (a magyar doboz diagram kifejezés nem igazán terjedt el) a sokaság öt jellemző értékének (five number summary) grafikus megjelenítése, amely alkalmas különböző sokaságok, vagy csoportok eloszlásának összehasonlítására. Az ábrának több, kissé eltérő változata is létezik, de az valamennyi esetben közös, hogy az alábbi öt érték feltüntetésre kerül:

- $X_{(1)}$
- Q_1
- Me
- Q_3
- $X_{(N)}$

Az ábra felépítése egy Q_1 és Q_3 közötti dobozból (innen az elnevezés), a mediánnál egy ezen áthaladó vonalból, valamint a dobozból kinyúló „bajszokból” (angolul box-and-whisker plot néven is ismert) áll a minimum és a maximum irányában. Néhány szoftver a medián mellett az átlagot is jelöli egy ponttal, vagy kereszttel, illetve sok esetben a kiugró értékekként érzékelt megfigyelések (lásd 2.4. fejezet, bár az outliereket általában nem a z-score, hanem az IQR segítségével azonosítják) is különálló pontokként szerepelnek, ami miatt a bajusz rövidebb lesz. A boxplotokat vízszintesen és függőlegesen is rajzolhatjuk. A 3.1. ábrán a fejezet elején definiált öt sokaság boxplot ábráját tüntettük fel, ezek emlékeztetőül:

1. szimmetrikus, nem túl lapos, nem túl csúcsos
2. szimmetrikus, lapult eloszlás
3. szimmetrikus, csúcsos eloszlás

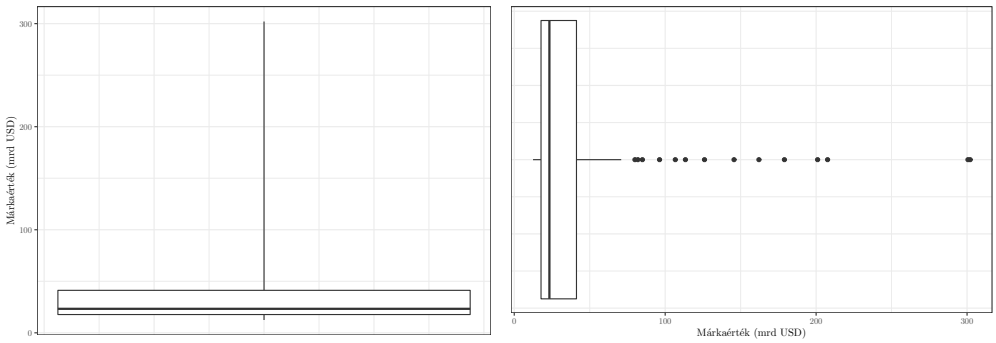
4. jobboldali aszimmetria, csúcsos eloszlás
5. baloldali aszimmetria, csúcsos eloszlás



3.1. ábra. Különböző eloszlással rendelkező sokaságok alapján készített boxplotok

A TOP100 brandérték példánkra visszatérve: azt már korábban láttuk, hogy az átlag (43,835) jelentősen meghaladja a mediánt (23,2955), tehát várhatóan jobboldali aszimmetriáról fog tanúskodni a boxplot. A terjedelem kiszámításánál (2.3.1. fejezet) láttuk, hogy a minimum 12,456, a maximum pedig 302,063, azaz a medián jóval közelebb helyezkedik el a minimumhoz, mint a maximumhoz. Az alsó kvartilis értéke mintegy 17,6 (a pontos érték az alkalmazott képlettől függ), míg a felső kvartilis mintegy 41,4 milliárd dollár. Azaz a TOP100 márkaértékkel rendelkező vállalatok negyedének értéke 17,6 milliárd dollár alatti, háromnegyedüknek ezt az értéket meghaladja, illetve a vállalatok negyedének haladja meg az értéke a 41,4 milliárd dollárt. A TOP100 brandérték alapján készült boxplotot mutatja be a 3.2. ábra. Míg bal oldali ábra egy kiugró értékeket külön nem jelölő, függőlegesen elhelyezkedő boxplot, addig a jobb oldalon a kiugrónak ítélt megfigyeléseket pontokkal jelöltük a vízszintes tengelyen helyeztük el az ábrázolandó értékeket.

Összefoglalóan tehát a kvantilisek tetszőleges osztópontokat jelölnek, melyek önmagukban is alkalmasak elemzésre, a leggyakrabban a kvartiliseket használja a



3.2. ábra. A márkaértékek boxplotja

statisztika, elsősorban a boxplot megalkotására. A boxplot alakjából következtetünk a sokasági megfigyelések elhelyezkedésére, a jobboldali aszimmetria esetén hosszú jobb oldali (vagy fenti) bajuszt és akár sok kiugró értéket látunk. A sokaság lapult, ha mind a doboz, mind pedig a bajszok viszonylag nagyok, illetve csúcsos, ha a box (vagyis az interkvartilis terjedelem), vagy valamelyik bajusz nagyon rövid. Ne feledjük, bármely két szomszédos kvartilis között a sokaság elemeinek negyede található meg!

3.2. Osztályközös gyakorisági sor

Arra, a fejezet elején feltett kérdésünkre, hogy milyen struktúrában helyezkednek el az adataink, egy másik megközelítés, és egy másik statisztikai ábra is választ adhat. A megközelítést osztályközös gyakorisági sornak, az ábrát pedig hisztogramnak hívjuk.

Míg a 3.1. fejezetben taglalt kvantilisek esetén a két szomszédos kvantilis között található sokasági elemek száma állandó, és az információt az ő távolságuk hordozza, addig az osztályközös gyakorisági sorok esetén épp fordított a megközelítés: (alapesetben) azonos távolságra lévő pontokat jelölünk ki, majd azt vizsgáljuk, hogy hány darab sokasági megfigyelés esik az így kialakított kategóriákba, vagy más szóval osztályközökbe. Az így kialakított statisztikai sort osztályközös gyakorisági sornak nevezzük. Tananyagunkban egyenlő hosszúságú osztályközös gyakorisági sorokra koncentrálunk, de elvileg készíthetünk egymástól eltérő hosszúságú osztályközöket is.

Az osztályközös gyakorisági sor általános sémája a 3.1. táblázatban látható, J osztályköz feltételezésével. Az első osztályköz alsó és az utolsó osztályköz felső határa azért került zárójelbe, mert ezek gyakran nyitottak, azaz nincsenek megadva, például az $X_1^{\text{felső}}$ értékkel egyenlő, vagy attól kisebb valamennyi megfigyelés az első osztályközbe tartozik.

3.1. táblázat: Osztályközös gyakorisági sor

osztályköz	gyakoriság
$(X_1^{\text{alsó}})$ - $X_1^{\text{felső}}$	F_1
$X_2^{\text{alsó}}$ - $X_2^{\text{felső}}$	F_2
...	...
$X_J^{\text{alsó}}$ - $(X_J^{\text{felső}})$	F_J
Összesen	N

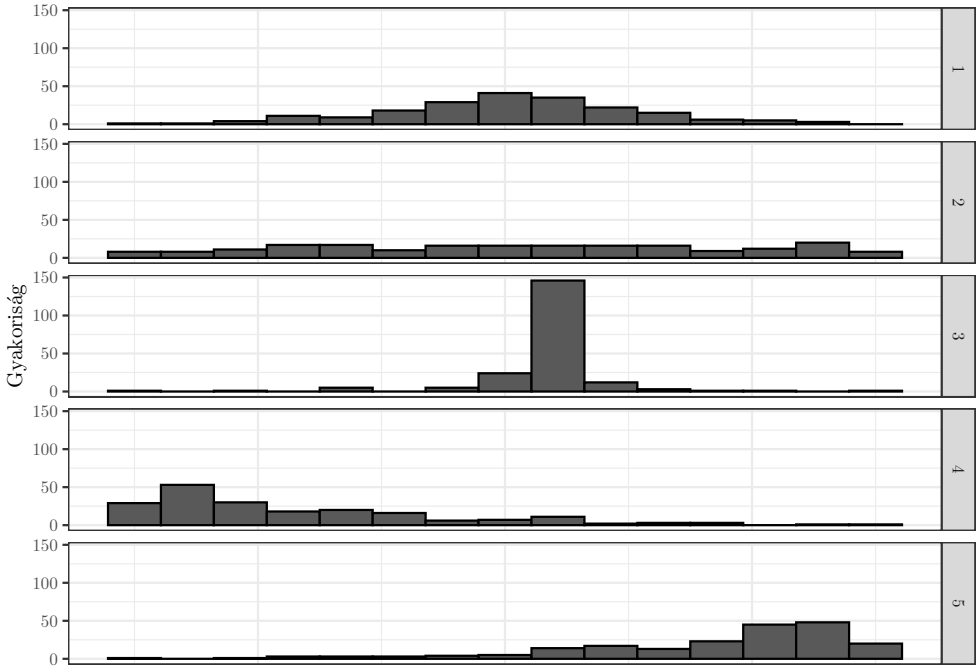
Az osztályközök számának, valamint az osztályközök határainak meghatározására nincs egyetlen jó megoldás, néhány fontos szempont az alábbiakban foglalható össze:

- tipikusan ajánlott egyenlő hosszúságú osztályközök alkalmazása,
- áttekinthető, egyértelmű, teljes (nyitott alsó és felső közök) legyen az osztályközös gyakorisági sor, minden megfigyelés pontosan egy osztályba tartozzon, azaz tisztázni kell az osztályköz határra eső megfigyelések helyét,
- lehetőleg ne legyenek üres (0 gyakoriságú) osztályközök, ami néhány extrém kiugró értéknél jelenthet gondot, erre szintén megoldás a nyitott osztályköz,
- „szép” osztályköz határok (az Excel jellemzően ezt a kritériumot nem teljesíti, más szoftverek, pl. az R igen), ami alatt azt értjük, hogy lehetőség szerint kerek, vagy kevés tizedessel rendelkező, jól azonosítható törtek legyenek a határok,
- több különböző szabály/képlet is van az osztályközök számára vonatkozóan, fő szabályként a minél több adat, megfigyelés egyre részletesebb (több osztályközt tartalmazó) gyakorisági sort engedélyez.

Az osztályközös gyakorisági sorhoz tartozó statisztikai ábra neve hisztogram, mely a vízszintes tengelyen az osztályközöket tartalmazza, míg a függőleges tengelyen a gyakoriságokat.

A 3.3 ábra alapján jól látható, hogy nagyon különböző sokasági eloszlások képzelhetőek el, itt is ugyanazokat a sokaságokat ábrázoltuk hisztogram segítségével, mint a boxplotok segítségével a 3.1. ábrán:

1. szimmetrikus, nem túl lapos, nem túl csúcsos
2. szimmetrikus, lapult eloszlás
3. szimmetrikus, csúcsos eloszlás
4. jobboldali aszimmetria, csúcsos eloszlás
5. baloldali aszimmetria, csúcsos eloszlás



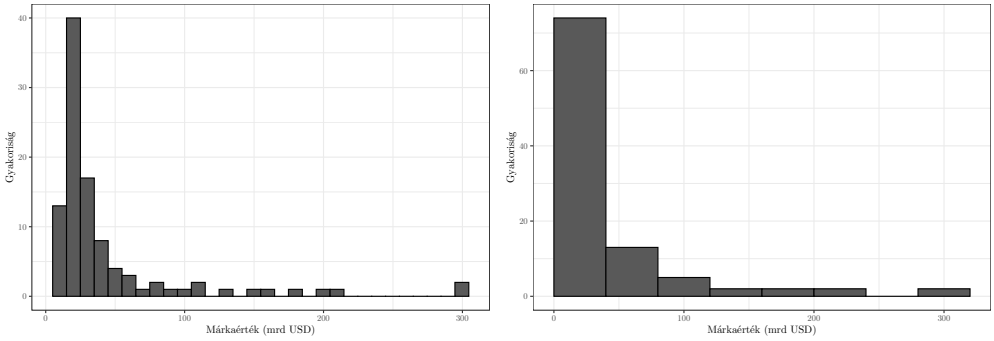
3.3. ábra. Különböző eloszlással rendelkező sokaságok alapján készített hisztogramok

Készíthetünk gyakorisági sort a TOP100 vállalat márkaértékei alapján. Legyenek az osztályközök 40 milliárd dollár hosszúak, és az első osztályköz kezdődjön 0-tól! (Vegyük észre, hogy más értékeket is választhattunk volna, ráadásul a példánk nem is teljesíti az üres osztályközökkel kapcsolatos szempontot!) Ekkor a gyakorisági sor:

3.2. táblázat: Osztályközös gyakorisági sor a TOP100 adatok alapján

osztályköz	gyakoriság
-40]	74
(40-80]	13
(80-120]	5
(120-160]	2
(160-200]	2
(200-240]	2
(240-280]	0
(280-	2
összesen	100

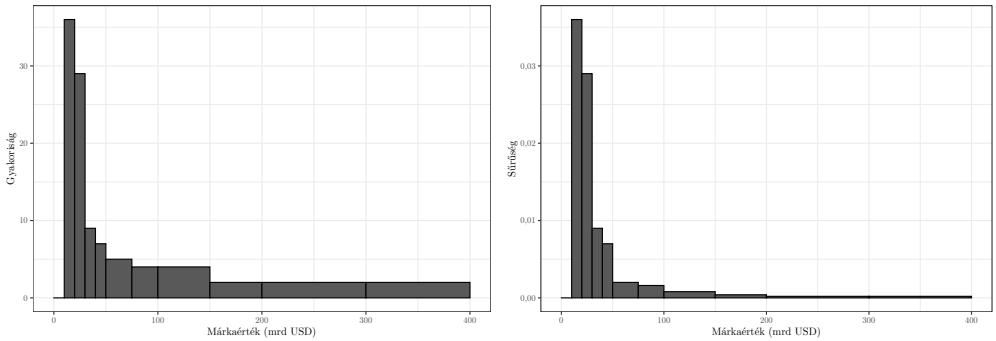
Az osztályközök kialakításakor a határokat úgy állapítottuk meg, hogy a 40 milliárdos érték még az első osztályközbe esik, a 40 milliárd feletti, 80 milliárdig bezárólag a másodikba, stb. A határra eső értékek hovatartozását érzékeltettük a] és (jelek alkalmazásával. A TOP100 vállalat márkaértékeire vonatkozó hisztogramok a 3.4. ábrán láthatók, két különböző osztályköz hossz esetén, azonos adatok alapján. Mindkét hisztogram alapján elmondható, hogy erős jobboldali aszimmetria figyelhető meg. Míg a baloldali ábra talán túl sok, rövid osztályköz alapján, addig a jobboldali a fenti táblázatban szereplő adatokból készült.



3.4. ábra. A márkaértékek hisztogramja

A nem egyenlő hosszúságú osztályközök esete meghaladja tananyagunk kereteit, azonban a példában is látott erős aszimmetria esetén alkalmazása szükséges lehet. Ebben az esetben arra kell figyelni az ábra elkészítésekor, hogy az eltérő hosszúságú osztályközökhöz kiszámított nyers gyakoriságok torzítanak abban az értelemben, hogy az osztályköz hosszát nem veszik figyelembe, amit korrigálni kell.

Egy, a TOP100 márkaérték adatállománya alapján készített, nem egyenlő hosszúságú osztályközöket ábrázoló hisztogramokat mutat be a 3.4. ábra. Az osztályközök kezdetben 10, 25, 50, majd 100 hosszúságúak. A baloldali ábra az egyes osztályközökhöz tartozó gyakoriságokat ábrázolja, ugyanakkor nem jogos összehasonlítani egy 10 hosszúságú és egy 100 hosszúságú osztályköz gyakoriságát. A statisztikai szoftverek a szükséges korrekciót automatikusan elvégzik (jobboldali ábra).



3.5. ábra. A márkaértékek hisztogramja – eltérő osztályköz hosszok esetén

Az osztályközös gyakorisági sor további elemzésekre is lehetőséget nyújt a hisztogram elkészítésén túl. A gyakoriságok alapján gyakran számítunk relatív gyakoriságokat (G_j), azaz megoszlási viszonyszámokat (1.5) alapján. A másik gyakran alkalmazott művelet a kumulálás, vagy felösszegzés, amit a gyakoriságokra és a relatív gyakoriságokra is elvégezhetünk.

$$F'_j = \sum_{k=1}^{j} F_k \quad G'_j = \sum_{k=1}^{j} G_k \tag{3.2}$$

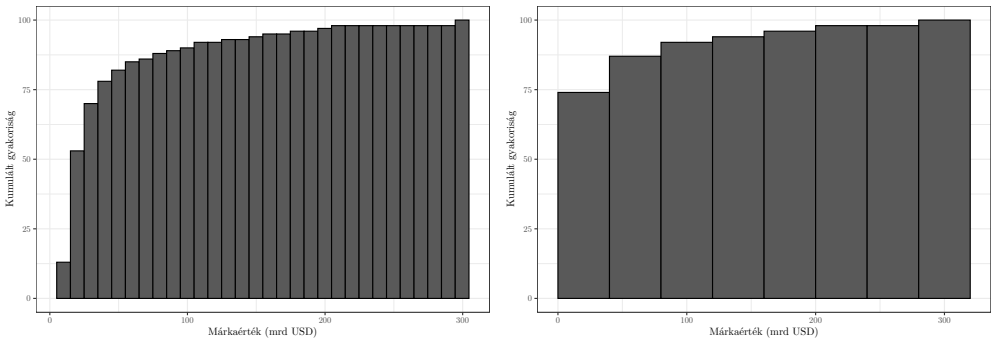
Azaz a j. osztályköz kumulált gyakorisága az első j osztályköz gyakoriságainak az összege. A kumulált relatív gyakoriság hasonlóan képezhető. A kumulált értékek ábrázolása kumulált hisztogramot eredményez.

A gyakorisági táblázatot kiegészíthetjük a relatív gyakoriságokkal (ami 100 elemű sokaság esetén triviális), illetve kumulált értékeket is feltüntethetünk.

3.3. táblázat: Relatív gyakoriságok és kumulált értékek a TOP100 adatok alapján

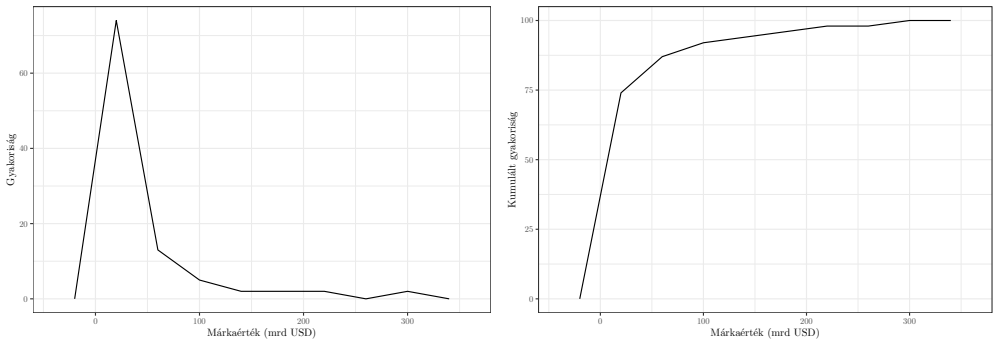
osztályköz	F_j	G_j	F'_j	G'_j
-40]	74	0,74	74	0,74
(40-80]	13	0,13	87	0,87
(80-120]	5	0,05	92	0,92
(120-160]	2	0,02	94	0,94
(160-200]	2	0,02	96	0,96
(200-240]	2	0,02	98	0,98
(240-280]	0	0	98	0,98
(280-	2	0,02	100	1,00
összesen	100	1	-	-

A második osztályköz kumulált gyakorisága azt jelenti, hogy 87 vállalatnak van 80 milliárd dolláros, vagy az alatti márkaértéke, ami jelen példában 87%-nak felel meg. A 3.4. ábrán látható két hisztogram párjaként, kumulált gyakoriságokból készült a 3.6. ábra két kumulált hisztogramja (a jobboldali adatai láthatók a fenti táblázatban).



3.6. ábra. A márkaértékek kumulált hisztogramja

A hisztogram az egyik legnépszerűbb eszköz egy sokaság elhelyezkedésének, eloszlásának bemutatására, ugyanakkor hátránya, hogy összehasonlításra kevésbé alkalmas, hiszen a két sokaságból készített hisztogramok eltakarnák egymást egy közös ábrára helyezve. Ennek orvoslására gyakran készítenek (kumulált) gyakoriségi poligonokat, ami gyakorlatilag a (kumulált) hisztogram oszlopainak középpontját összekötő egyenesekből áll. A 3.7. ábra ezeket a grafikonokat mutatja be a TOP100 adatokra.



3.7. ábra. Poligon és kumulált poligon a márkaértékre

3.3. Alakmutatók

Az eloszlás alakjának jellemzésére egy harmadik megközelítést is bemutatunk, méghozzá számított mutatószámok segítségével.

3.3.1. Ferdeség

Az aszimmetria mérésére több alkalmas mutatószám is kidolgozásra került. Ahogy azt a 2.2.6. fejezetben is említettük, a számtani átlag és a medián egymáshoz viszonyított helyzete árulkodik az eloszlás alakjáról, hiszen míg a medián nem érzékeny a kiugró értékekre, addig a számtani átlag igen. Így, ha a két középérték között jelentős különbség van, valószínűleg aszimmetrikus a vizsgált eloszlás alakja. A két középérték különbségén alapuló mutatószámok is léteznek, de a leggyakrabban mégis az alábbi – később megismerendő fogalommal élve harmadrendű momentumokon alapuló – mutatót alkalmazzuk a ferdeség jellemzésére.

$$\gamma_1 = \frac{\frac{1}{N}\sum (X_i - \mu)^3}{\sigma^3} \tag{3.3}$$

Amennyiben a mutató értéke

- 0 körüli, úgy szimmetrikus eloszlásról
- negatív, akkor baloldali aszimmetriáról
- pozitív, akkor jobboldali aszimmetriáról

beszélünk. A gazdasági jelenségek esetén nagyon gyakran találkozunk a pozitív ferdeséggel, azaz a jobboldali aszimmetriával.

3.4. táblázat. Ferdeség és csúcsosság értékek

Sokaság	Ferdeség	Csúcsosság
1. szimmetrikus, nem túl lapos, nem túl csúcsos	-0,12	0,18
2. szimmetrikus, lapult eloszlás	-0,01	-1,13
3. szimmetrikus, csúcsos eloszlás	-1,76	14,70
4. jobboldali aszimmetria, csúcsos eloszlás	1,32	1,32
5. baloldali aszimmetria, csúcsos eloszlás	-1,40	1,76

A TOP100 vállat márkaértéke esetén pozitív ferdeségi mutatóértéket várunk. A mutató értéke 3,06, ami elég erősnek számít, azaz igazolja a mutató értéke is a boxplot és a hisztogram alapján tapasztaltakat. A mutatónak nincs elvi határa, tehát nem pl. -1 és 1 közötti, ahogy azt majd sok esetben látjuk a későbbiekben, ebben a példában is meghaladta az 1-es értéket.

Léteznek egyéb mutatók a szakirodalomban, melyek az átlag és a medián, esetleg az átlag és a módusz egymáshoz viszonyított helyzete alapján kerülnek kiszámításra. A nemzetközi szakirodalom azonban a leggyakrabban a (3.3) formula alapján vizsgálja az aszimmetriát.

3.3.2. Csúcsosság

A csúcsosság mutatója azt méri, hogy a módusz (vagy modális, leggyakrabban előforduló osztályköz) környékén mennyire sűrűsödnek a megfigyelések – egy negyedrendű momentumokon alapuló – mérőszám segítségével.

$$\gamma_2 = \frac{\frac{1}{N}\sum (X_i - \mu)^4}{\sigma^4} - 3 \tag{3.4}$$

Amennyiben a mutató értéke

- 0 körüli, úgy átlagos csúcsosságú, „normális” eloszlásról
- negatív, akkor lapult eloszlásról
- pozitív, akkor csúcsos eloszlásról

beszélünk. A tananyag későbbi, 6. fejezetében részletesen foglalkozunk a most csak „normálisnak” hívott elméleti eloszlással.

A 3.1. és 3.2. fejezetekben boxplotokon, illetve hisztogramokon bemutatott öt sokaság esetére kiszámítottuk a ferdeség és csúcsosság mutatókat, amiket a 3.4. táblázat mutat be.

A TOP100 vállat márkaértéke esetén pozitív mutatóértéket várunk. A mutató értéke 10 feletti, ami extrém csúcsos eloszlásnak számít, azaz igazolja a mutató értéke is a boxplot és a hisztogram alapján látottakat. Ennek a mutatónak sincs elvi határa.

3.4. Koncentráció

A koncentráció szó ismert köznapi jelentéssel bír, statisztikában azonban a koncentrációs vizsgálat a sokaság egyedeinek részesülését elemzi valamilyen változó alapján. A koncentráció elemezhető az egyedi megfigyelések alapján is, de jóval elterjedtebb annak vizsgálata a gyakorisági táblázat további elemzése kapcsán. Az elemzéshez szükségünk lesz a kumulált relatív gyakoriság mellett a kumulált relatív értékösszeg fogalmának bevezetésére.

3.4.1. Koncentrációs táblázat

Egy osztályközhöz tartozó értékösszeg alatt egyszerűen az abba az osztályba eső megfigyelések adott változó szerinti összegét értjük. Amennyiben az alapadatokat ismerjük, ez pontosan számítható egy egyszerű összegként, de gyakran élünk az alábbi közelítő meghatározással:

$$S_j = F_j X_j^* \tag{3.5}$$

ahol F_j a j. csoport gyakorisága, X_j^* az ún. osztályközép, az az érték, ami az adott osztályközt leginkább jellemzi, félúton az alsó és a felső határ között. Innen a relatív értékösszeg kiszámítása logikusan:

$$Z_j = \frac{S_j}{\sum_{k=1}^{J} S_k} \tag{3.6}$$

ahol J az osztályközök száma. A kumulálás fogalmát már megismertük, a művelet mind az értékösszegre, mind a relatív értékösszegre elvégezhető, eredményül a kumulált értékösszeget (S'_j) és a kumulált relatív értékösszeget (Z'_j) kapjuk.

A TOP100 vállat márkaértékére vonatkozó további elemzéseket az alábbi táblázat tartalmazza:

3.5. táblázat: Értékösszegek a TOP100 adatok alapján

osztályköz	X_j^*	F_j	S_j	S'_j	Z_j	Z'_j
-40]	20	74	1480	1480	0,333	0,333

(40-80]	60	13	780	2260	0,176	0,509
(80-120]	100	5	500	2760	0,113	0,622
(120-160]	140	2	280	3040	0,063	0,685
(160-200]	180	2	360	3400	0,081	0,766
(200-240]	220	2	440	3840	0,099	0,865
(240-280]	260	0	0	3840	0,000	0,865
(280-	300	2	600	4440	0,135	1,000
összesen	-	100	4440	-	1,000	-

Az S_j közelítő értékösszeg szerint az első osztályközbe tartozó (40 milliárd dollár érték alatti) cégek összesen 1480 milliárd dolláros márkaértékkel rendelkeznek (ne feledjük, hogy ez egy közelítő érték, a tényleges összérték 1591,3 milliárd dollár, amit a nyers adatok ismeretében ebben a példában ki tudunk számítani), míg a második osztályközbe tartozók 780 milliárdos összértékkel. Ez a 4440 milliárdra becsült összérték 17,6%-át teszi ki. A Z'_j oszlop harmadik sora például azt jelenti, hogy a 120 milliárd dollár alatti márkaértékek a teljes márkaérték mintegy 62,2%-át teszik ki.

A koncentráció elemzéséhez a kumulált relatív gyakoriság (G'_j), valamint a kumulált relatív értékösszeg (Z'_j) mutatóira van szükségünk, tulajdonképpen ezt a két számsort hasonlítjuk össze. A koncentrációs táblázat sematikus felépítése a 3.6. táblázatban látható. Az utolsó sorban természetesen mindkét kumulált sor esetén 1 érték szerepel, azaz a sokaság egésze felöleli az értékösszeg egészét.

3.6. táblázat: Koncentrációs táblázat

osztályköz	G'_j	Z'_j
$(X_1^{\text{alsó}})$ - $X_1^{\text{felső}}$	G'_1	Z'_1
$X_2^{\text{alsó}}$ - $X_2^{\text{felső}}$	G'_2	Z'_2
...	...	...
$X_J^{\text{alsó}}$ - $(X_J^{\text{felső}})$	1,000	1,000

A TOP100 márkaértékre vonatkozó koncentrációs táblázat az alábbi.

3.7. táblázat: Koncentrációs táblázat a TOP100 adatok alapján

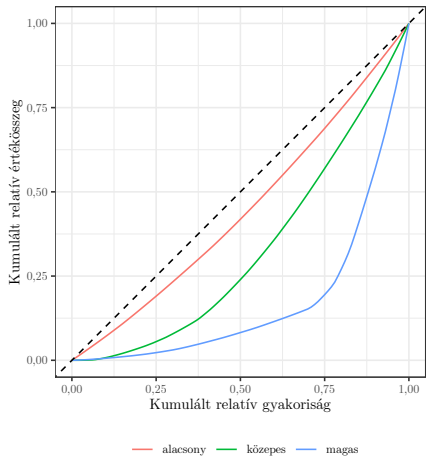
osztályköz	G'_j	Z'_j
-40]	0,740	0,333
(40-80]	0,870	0,509
(80-120]	0,920	0,622
(120-160]	0,940	0,685
(160-200]	0,960	0,766

(200-240]	0,980	0,865
(240-280]	0,980	0,865
(280-	1,000	1,000

A 3.7. táblázat alapján például azt olvashatjuk le, hogy a legkisebb márkaértékkel rendelkező vállalatok 87%-a a teljes márkaérték 50,9%-ával rendelkezik, vagy a vállalatok 98%-a az összérték 86,5%-ával, azaz a két legnagyobb márkaértékű vállalat rendelkezik a maradék 13,5%-kal. Vegyük észre, hogy két osztályköz esetén is azonosak a G'_j és természetesen a Z'_j értékek, ami azért következett be, mert üres osztályköz figyelhető meg az osztályközös gyakorisági sorban.

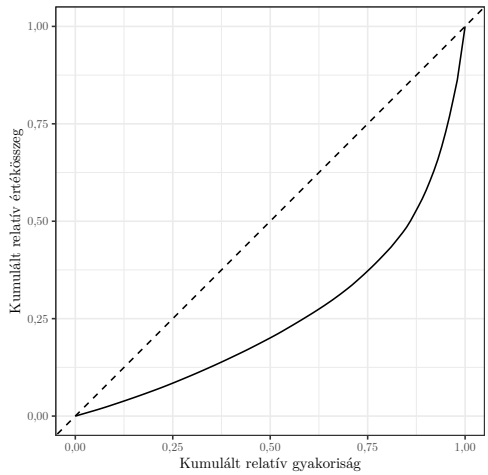
3.4.2. Lorenz-görbe

A Lorenz-görbe tulajdonképp a koncentrációs táblázat vizuális megjelenítését szolgálja egy pontdiagram segítségével. A vízszintes tengelyre a kumulált relatív gyakoriság, míg a függőlegesre a kumulált relatív értékösszeg kerül, az egyes pontok a gyakorisági táblából kerülnek ábrázolásra, majd összekötésre. Amennyiben nem lenne koncentráció, valamennyi pont a főátlón helyezkedne el, ezért a főátlót mintegy referenciaként ábrázolni szokás. A 3.8. ábra három különböző sokaságra vonatkozó görbét mutat be, rendre alacsony, közepes és magas koncentrációval.



3.8. ábra. Három különböző Lorenz-görbe

A TOP100 márkaértékre vonatkozó (egyedi megfigyelések alapján készített) Lorenz-görbét mutatja be a 3.9. ábra. A koncentráció az ábra alapján is magasnak mondható.



3.9. ábra. Lorenz-görbe a TOP100 márkaérték alapján

A Lorenz-görbe esetén tehát minél messzebb fekszik a görbe az átlótól, annál erősebb a koncentráció. A Gini-együttható a görbe és az átló által bezárt terület nagyságát méri, a maximális 0,5-höz viszonyítva. A Gini-együttható értéke 0 és 1 közé normalizált, 0 értéke a koncentráció hiányát, 1 értéke a maximális koncentrációt mutatja.

A TOP100 márkaértékre vonatkozó Gini-index meghatározása meghaladja tananyagunk kereteit, értéke megközelíti a 0,5-öt, azaz elég erős koncentrációról beszélhetünk, még akkor is, ha csak a TOP100 márkaérték esetét vizsgáljuk. Ha a világ valamennyi márkáját tudnánk vizsgálni, a koncentráció még nagyobb lenne.

A koncentráció az egyedi adatok alapján is vizsgálható, ebben az esetben a relatív gyakoriságokat $\frac{1}{N}$ adja minden megfigyelésre, míg a kumulált értékösszeget egyszerűen a növekvő rendbe állított megfigyelések kumulálásával kapjuk. Az egyedi adatokból készített Lorenz-görbe megrajzolását az érdeklődő Olvasóra bízzuk.

3.5. Excel tippek

Hasznos Excel függvények:

- kvantilisek:
  - KVARTILIS.KIZÁR, KVARTILIS.TARTALMAZ
  - PERCENTILIS.KIZÁR, PERCENTILIS.TARTALMAZ
- osztályközös gyakorisági sor: GYAKORISÁG

A GYAKORISÁG egy ún. tömbfüggvény. Alkalmazásához előre ki kell jelölnünk azokat a cellákat ahova a gyakoriságokat meg szeretnénk kapni. A Csoport-tömbnek a kívánt osztályközök felső határait kell megadnunk, majd a függvényt az ENTER helyett a CTRL+SHIFT+ENTER billentyűkkel kell meghívni.

- alakmutatók: FERDESÉG.P, CSÚCSOSSÁG

Az Excelben szerepel a FERDESÉG és a FERDESÉG.P függvény, amelyből a második az, ami a (3.3) formulával adott mérőszámot kiszámítja. A FERDESÉG függvény is hasonló eredményt ad, az a függvény a minta ferdeségéből enged következtetést levonni, ami a későbbi fejezetek tananyaga. Sajnos a csúcsosság esetén nem került implementálásra a sokasági adatokra alkalmazható függvény (nincs CSÚCSOSSÁG.P függvény), így a némileg eltérő, de implementált függvényt fogjuk használni, vállalva a kis pontatlanságot.

Hasznos Excel funkciók:

- Hisztogram és Doboz-diagram (boxplot) beszúrása 
- Adatelemzés menü Hisztogram eszköze 

Az Excelbe beépített szabály általában túl sok osztályközt hoz létre, de lehetőség van az osztályközök számának módosítására. A 2016-os Excel verzióban már közvetlenül elérhető a hisztogram és a boxplot ábrája, és a hisztogram viszonylag rugalmasan testreszabható. Az Adatelemzés menü használatakor a Rekesztartomány opciónál megadhatjuk a kívánt osztályközök felső határait. A leginkább rugalmas megoldás (pl. eltérő hosszúságú osztályközök készítéséhez) természetesen a GYAKORISÁG tömbfüggvény alkalmazása.

4. fejezet

Bevezetés a valószínűségszámításba

A Valószínűségszámítás és statisztika tárgy keretében az első három fejezetben sokaságokat leíró módszereket ismertünk meg. A statisztika másik (jóval nagyobb) nagy területe a következtetéses statisztika mintából következtet a sokaságra. A mintavétel véletlen jellege miatt a valószínűségszámítás alapjait ismerjük meg a 4. fejezetben, majd a következő két fejezetben a legfontosabb diszkrét és folytonos eloszlásokat taglaljuk. A tananyagban a teljes matematikai pontosságot esetenként feláldozzuk az érthetőség oltárán, de úgy gondoljuk, hogy ez vállalható kompromisszum ebben az esetben.

4.1. Valószínűségszámítás alapfogalmai

A társadalmi, gazdasági, természeti jelenségek véletlen jellegük alapján az alábbi csoportokba sorolhatók.

4.1.1. Jelenségek csoportosítása

- determinisztikus
- sztochasztikus (véletlen jelenség)
  - egyszeri véletlen jelenség
  - véletlen tömegjelenség

A determinisztikus jelenségről beszélünk, ha adott körülmények mellett minden esetben ugyanaz az esemény következik be. Amennyiben adott körülmények között

nem mindig azonos esemény következik be, a jelenséget sztochasztikusnak nevezzük. A két nagy csoport között nem annyira éles a határvonal, mint azt elsőre gondolnánk: azt, hogy egy jelenséget melyik csoportba sorolunk, nagyban befolyásolja, hogy mik azok az említett „adott körülmények".

A determinisztikus jelenségek leginkább a természettudományokat jellemzik. Amennyiben a megfelelő körülmények fennállnak (légnyomás, hőmérséklet, stb.), a jég elolvad. Ezzel szemben, amikor egy pénzérmét feldobunk, az hol a fej, hol az írás oldalával felfelé fog leesni. Ilyen szempontból a pénzérme feldobása sztochasztikus jellegű. Amennyiben lenne egy pénzfeldobó gépünk, ami figyelembe veszi az érme fizikai jellemzőit, a légmozgást, stb, úgy a jelenség többé már nem lenne sztochasztikus. Egy jelenség jellege tehát nagyban függ attól, hogy ésszerű kereteken belül mennyire tudjuk a körülményeket kontrollálni.

A sztochasztikus jelenségek még két csoportra bonthatók, az egyszeri véletlen jelenségek jellemzője, hogy az őket befolyásoló körülmények nem ismételhetők meg, míg a véletlen tömegjelenségek olyan jelenségek csoportját alkotják, amelyek – legalábbis elvileg – változatlan körülmények között akár végtelen sokszor megismételhetők. A valószínűségszámítás a fenti két csoport közül az egyszeri véletlen jelenségekkel nem foglalkozik, további vizsgálatainkban a véletlen tömegjelenségek vizsgálatára fókuszálunk.

Egyszeri véletlen jelenségek közé soroljuk például a sporteseményeket, amik jellemzően olyanok, hogy azonos körülmények között nem ismételhető meg, valószínűleg pontosan ez adja a sportfogadás népszerűségét. A véletlen tömegjelenségek közé tartozik például az egyszerű kockadobástól a lottóhúzáson át a statisztikai mintavétel is.

A statisztikai alkalmazásokon kívül a véletlen jelenségek és azok kezelése a gazdasági élet rengeteg területén felmerül. A pénzügyi élet, a biztosítótársaságok mind alkalmaznak valószínűségszámításon alapuló modelleket és módszereket. A következő három fejezetben olvasható ismeretek mindegyik témakör szempontjából hasznos ismeretek.

4.1.2. Elemi esemény, esemény, eseménytér

A véletlen jelenségek (más szóval véletlen kísérletek) során megfigyelünk egy kimenetelt, melyeket elemi eseményeknek hívunk, az elemi eseményeket ω jelöli. Az összes elemi eseményt tartalmazó halmazt eseménytérnek nevezzük és Ω-val jelöljük. A gyakorlatban jellemzően közvetlenül nem az elemi eseményekkel foglalkozunk, hanem eseményeket szeretnénk vizsgálni. Az események az elemi események egy részhalmazaként definiáljuk, és jellemzően az ABC elejéről származó nagybetűkkel jelöljük őket: $A, B, C, ...$, vagy amennyiben sok eseményről beszélünk, gyakran indexekkel különböztetjük meg őket: $A_1, A_2, A_3, ...$ Egy esemény tehát állhat 0, 1, vagy akár több elemi eseményből is. Azt az eseményt, ami 0 elemi eseményből áll (üres halmaz, jele $\emptyset$), lehetetlen eseménynek nevezzük. Az Ω halmaz is egy esemény,

ezt az eseményt biztos eseménynek nevezzük. Végül az adott kísérlethez tartozó események halmazát $\mathcal{A}$ jelöli.

Már itt bevezetjük a valószínűség jelölését, a tananyagban $\mathbf{P}()$ fogja jelölni egy adott esemény valószínűségét. A valószínűség tehát minden eseményhez egy számot rendel.

Tekintsük át a fenti fogalmakat és jelöléseket a kockadobás példáján!

- elemi esemény: a dobható számok, $\omega_1 = 1$, $\omega_2 = 2$, ..., $\omega_6 = 6$
- eseménytér: $\Omega = \{\omega_i : i = 1, 2, 3, 4, 5, 6\} = \{1, 2, 3, 4, 5, 6\}$
- esemény:
  - A: páros dobás, $A = \{\omega_2, \omega_4, \omega_6\} = \{2, 4, 6\}$
  - B: páratlan dobás, $B = \{\omega_1, \omega_3, \omega_5\} = \{1, 3, 5\}$
- események halmaza:
  $\mathcal{A} = \{\{\emptyset\}, \{1\}, \{2\}, \ldots, \{1, 2\}, \{1, 3\}, \ldots, \{1, 2, 3\}, \ldots, \{\Omega\}\}$
- valószínűség: $\mathbf{P}(A) = \mathbf{P}(B) = 0{,}5$

Természetesen a páros és páratlan dobás esetén jelen tananyag szerint még nem tudjuk, hogy a valószínűség 0,5 értéket vesz fel, de korábbi ismereteinkből ezt már felírhatjuk, feltételezve az érme szabályosságát.

4.1.3. Események közti műveletek

Az események között definiálhatóak műveletek, illetve bemutatunk olyan jelöléseket, melyek eseményekhez kapcsolódnak. Legyen $A, B \in \mathcal{A}$, azaz tartozzon mindkét esemény egy adott eseménytérhez.

- Azt az eseményt, amikor A és B esemény is bekövetkezik, a két esemény szorzatának, vagy a két esemény metszetének hívjuk és $A \cap B$ módon jelöljük.
- Azt az eseményt, amikor az A és B események közül legalább az egyik bekövetkezik, a két esemény összegének, vagy a két esemény uniójának hívjuk és $A \cup B$ módon jelöljük.
- Azt az eseményt, amikor nem az A esemény következik be, az A esemény Ω-ra vett komplementer eseményének hívjuk és $\overline{A}$ módon jelöljük.
- Ha A és B esemény olyan, hogy valahányszor A bekövetkezik, akkor B is, akkor A-ból következik B és $A \subset B$ (vagy $A \subseteq B$) módon jelöljük. Amennyiben a két esemény egyezőségét is megengedjük, a $A \subseteq B$ jelölést használjuk.
- Azt az eseményt, amikor A bekövetkezik, de B nem, a két esemény különbségének hívjuk és $A - B$ (vagy $A \cap \overline{B}$) módon jelöljük.

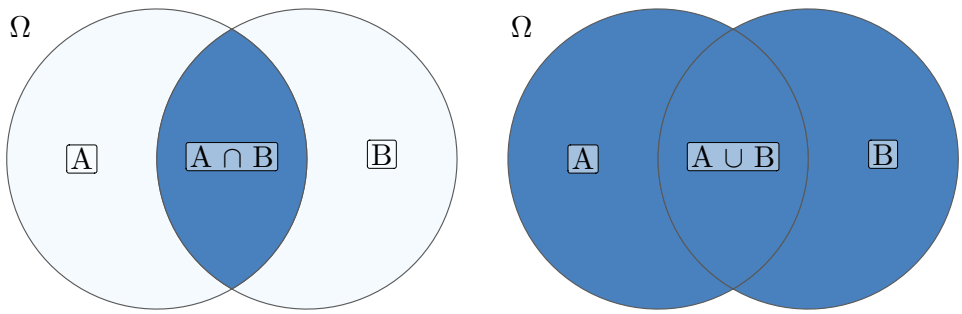
Legyenek adottak az alábbi események:

- $A = \{\text{páros dobás}\}$
- $B = \{\text{5-nél kisebb dobás}\}$
- $C = \{\text{3-nál kisebb dobás}\}$

Ekkor

- $A \cap B = \{2, 4\}$
- $A \cup B = \{1, 2, 3, 4, 6\}$
- $\overline{A} = \{1, 3, 5\}$
- $C \subseteq B$ teljesül
- $A \cap \overline{B} = \{6\}$

Az eseményeket gyakran ábrázoljuk Venn diagram segítségével. Az összegzés és a szorzás műveletét mutatja be a 4.1. ábra.



4.1. ábra. Az összegzés és a szorzás művelete

A fenti műveletekkel és jelölésekkel kapcsolatosan néhány egyszerű szabály, összefüggés írható fel. Mind az események összege, mind pedig az események szorzata rendelkezik az idempotencia, kommutativitás, asszociativitás és disztributivitás tulajdonságával:

4.1. táblázat: Az összegzés és szorzás tulajdonságai

tulajdonság	összeg	szorzat
idempotencia	$A \cup A = A$	$A \cap A = A$
kommutativitás	$A \cup B = B \cup A$	$A \cap B = B \cap A$

tulajdonság	összeg	szorzat
asszociativitás	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
disztributivitás	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Az A esemény és annak $\overline{A}$ komplementer eseményére vonatkozóan belátható, hogy

$$A \cap \overline{A} = \emptyset \qquad A \cup \overline{A} = \Omega$$

azaz tetszőleges A esemény és komplementere egyszerre soha nem következnek be, illetve az A esemény és komplementere együttesen az eseményteret adják.

Szintén az események közötti elemi műveletekkel kapcsolatosak a DeMorgan azonosságok, melyek szerint tetszőleges A és B eseményekre igaz:

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \qquad \overline{A \cap B} = \overline{A} \cup \overline{B}$$

A fenti két azonosság Venn-diagram segítségével egyszerűen ellenőrizhető.

Az események egy $A_1, A_2, \ldots, A_n$ rendszerét teljes eseményrendszernek hívjuk, amennyiben az alábbi két feltétel teljesül:

$$A_1 \cup A_2 \cup \ldots A_n = \Omega \tag{4.1}$$

azaz az események összege a teljes eseményrendszer, valamint

$$A_i \cap A_j = \emptyset \text{ minden } i \neq j, i, j = 1, 2, \ldots, n\text{-re} \tag{4.2}$$

4.2. Klasszikus valószínűségi mező

Az események és a köztük elvégezhető műveletek, jelölések bemutatása után vezessük be a valószínűség fogalmát. Ahogy azt már említettük, bármely A esemény bekövetkezésének valószínűségét $\mathbf{P}(A)$ módon jelöljük. A valószínűségszámítás az alábbi három, Kolmogorov-féle axiómára támaszkodik:

$$\begin{gathered} 0 \leq \mathbf{P}(A) \leq 1 \\ \mathbf{P}(\Omega) = 1 \\ \text{ha } A \cap B = \emptyset, \text{ akkor } \mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) \end{gathered} \tag{4.3}$$

Azaz a valószínűség egy olyan függvény, amely minden eseményhez egy számot rendel hozzá, méghozzá a 0-1 zárt intervallumból.

A valószínűség meghatározásának és interpretációjának három alapvető megközelítését különböztethetjük meg:

1. szubjektív: az eseménnyel kapcsolatos egyéni várakozások kifejeződése, amely ennek megfelelően egyénenként változhat. Az ún. bayesi statisztika erősen támaszkodik a szubjektív valószínűségekre, ennek tárgyalása azonban meghaladja tananyagunk kereteit.

2. objektív: az esemény valószínűségének meghatározásához sok kísérletet végzünk, és feljegyezzük, hogy az esemény bekövetkezett-e, vagy sem. Az objektív valószínűség szerint a relatív gyakoriságok a valószínűséghez fognak közelíteni, ha a kísérletek számát minden határon túl növeljük.

3. logikai: az esemény valószínűségét elméleti módon, a kísérlet elméleti jellemzői alapján vezeti le, tananyagunkban leginkább ezt a megközelítést alkalmazzuk.

A valószínűség fenti három megközelítése természetesen nem zárják ki egymást, alkalmazásuk inkább a vizsgálandó jelenség összetettségétől függ. Viszonylag egyszerű problémák esetén (érmedobás, kockadobás, mintavétel, stb.) a logikai út vezet eredményre. Amennyiben a kísérlet, vizsgálni kívánt jelenség bonyolultabb, logikai úton nem határozható meg a valószínűsége, de nincs elvi akadálya a kísérlet sokszori elvégzésének, akkor az objektív megközelítés adhat eredményt (számítógépes szimulációk, véletlenszám generálás, stb.). Szubjektív valószínűségeket gyakran használunk a mindennapi életben is, amikor nem képzelhető el a kísérlet sokszori megismétlése, azaz nem véletlen tömegjelenséggel, hanem egyszeri véletlen jelenséggel van dolgunk. A szubjektív valószínűség azonban a tudományos megismerésben is egyre nagyobb szerepet tölt be.

A (4.3) axiómákból levezethető néhány fontos, eseményekre vonatkozó tulajdonság:

- $\mathbf{P}(\emptyset) = 0$, azaz a lehetetlen esemény valszínűsége 0
- $\mathbf{P}(\overline{A}) = 1 - \mathbf{P}(A)$, azaz a komplementer esemény valószínűsége egyből való kivonással kapható meg
- $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$
- ha $A \subseteq B$, akkor $\mathbf{P}(A) \leq \mathbf{P}(B)$

Klasszikus valószínűségi mezőről beszélünk, ha

- az eseménytérben található elemi események száma véges (jelölje a számukat $|\Omega| = n \in \mathbb{N}$),
- az elemi események valószínűsége pozitív és egyenlő.

Amennyiben a két fenti feltétel teljesül, könnyen belátható, hogy minden elemi esemény valószínűsége

$$\mathbf{P}(\omega_1) = \cdots = \mathbf{P}(\omega_n) = 1/n \tag{4.4}$$

hiszen

$$1 = \mathbf{P}(\Omega) = \mathbf{P}\left(\bigcup_{i=1}^{n} \omega_i\right) = \sum_{i=1}^{n} \mathbf{P}(\omega_i) = n\mathbf{P}(\omega_1)$$

A gyakorlatban jellemzően nem egyetlen elemi esemény, hanem egy komplexebb jelenség, azaz egy A esemény valószínűségét szeretnénk meghatározni. Tegyük fel, hogy az $A \in \mathcal{A}$ esemény k darab elemi – kedvező – eseményből áll, azaz $A = \{\omega_1, \ldots, \omega_k\}$, ekkor

$$\mathbf{P}(A) = \mathbf{P}(\omega_1 \cup \cdots \cup \omega_k) = \mathbf{P}(\omega_1) + \cdots + \mathbf{P}(\omega_k) = k \cdot \frac{1}{n} = \frac{k}{n} \tag{4.5}$$

azaz az esemény valószínűségét a klasszikus valószínűségi mező feltételrendszerében a kedvező események és az összes esemény számának hányadosaként számítjuk.

A pénzérme, vagy a kocka egyszeri feldobásának modellje könnyen belátható illeszkedik a klasszikus valószínűségi mező feltételrendszeréhez. A lehetséges események száma véges ($n = 2$, illetve $n = 6$), és azt feltételezzük, hogy bekövetkezési valószínűségeik megegyeznek (fair érme, illetve dobókocka). Ennek megfelelően egy elemi esemény valószínűsége $1/2$, illetve $1/6$. A pénzérme esetén nehéz komplex eseményt definiálni az elemi események kis száma miatt, de a kockadobás esetére (4.5) már alkalmazható. Legyen az A esemény az, hogy a dobás páros, ekkor a kedvező esetek ($A = \{2, 4, 6\}$) könnyen megszámlálhatók, azaz $k = 3$, a keresett valószínűség pedig $\mathbf{P}(A) = \frac{1}{2}$.

Más esetekben a klasszikus valószínűségi mező alkalmazhatósága nem ilyen egyszerűen belátható. A lottóhúzás esetén például tudjuk, hogy az öttalálatos szelvény valószínűsége jóval kisebb, mint például az egytalálatosé, így első ránézésre azt gondolhatjuk, hogy itt nem alkalmazható a klasszikus valószínűségi mező feltételrendszere és képletei. Ne felejtsük el azonban, hogy a lehetséges számötösök száma véges, valószínűségük pedig egyenlő, így ebben a szituációban is alkalmazható (4.5), ahogy azt a 5.3.4. fejezetben látni fogjuk.

Nem csak a klasszikus valószínűségi mező létezik, jelen tananyagunkban azonban csak ezt az egyet említjük meg. Az ún. geometriai valószínűségi mező is fontos szerepet játszik, alkalmazásakor jellemzően idomok területeinek, vagy térfogatainak arányaként határozunk meg valószínűségeket.

A (4.5) egyenlet tehát azt mondja röviden, hogy a klasszikus valószínűségi mező feltételrendszerében a valószínűséget a „kedvező per összes” módon számítjuk. Sok egyszerű esetben ez nem okoz különösebb problémát, azonban összetettebb szituációkban a kedvező, és/vagy összes esetek megszámlálása nem mindig egyszerű.

Az esetek megszámlálásában segít a kombinatorika, jelen tananyagban három, elemi események megszámlálását segítő technikát említünk meg.[1]

- permutáció: arra a kérdésre ad választ, hogy n objektumot hányféleképpen lehet sorba rakni. Mivel az első helyre n objektum közül választhatunk, a másodikra $n-1$, stb, az összes lehetséges sorrend:

$$P_n = n! = n \cdot (n-1) \cdots 2 \cdot 1 \tag{4.6}$$

- variáció: arra a kérdésre ad választ, hogy n objektumot hányféleképpen lehet k helyre sorba rakni, ha ismétlődhetnek az elemek. Mivel minden helyre minden objektumot elhelyezhetjük, ezért a lehetséges esetek száma:

$$V_n^k = n^k \tag{4.7}$$

- kombináció: arra a kérdésre ad választ, hogy n objektum közül hányféleképpen lehet kiemelni k darabot úgy, hogy a sorrend nem számít:

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{4.8}$$

4.3. Feltételes valószínűség

Nagyon gyakran arra vagyunk kíváncsiak, hogy egy esemény bekövetkezése befolyásolja-e, és ha igen, milyen mértékben egy másik esemény valószínűségét. Az ehhez hasonló kérdések vizsgálatában a feltételes valószínűség fogalma lesz segítségünkre. Ebben az alfejezetben elsőként definiáljuk a feltételes valószínűséget, majd a definíció segítségével több hasznos tételt, állítást is megfogalmazunk. Az itt megismert fogalmakra épít majd a Statisztikai modellezés tárgy több témaköre is.

Tegyük fel, hogy egy A esemény $\mathbf{P}(A)$ valószínűségét vizsgáljuk, majd tudomásunkra jut, hogy egy B esemény bekövetkezett ($\mathbf{P}(B) \neq 0$). Ennek tükrében az A esemény valószínűsége megváltozhat, ezt a feltételes valószínűséget $\mathbf{P}(A \mid B)$ jelöli, a feltételes eseményt A feltéve B, vagy A vonás B eseménynek nevezzük.

Egy új termék bevezetése előtt a vezérigazgató egy adott valószínűséget rendel a termék sikeréhez (A). A bevezetés előtti fogyasztói tesztek eredményei (B) megváltoztathatják az erről alkotott véleményét mind pozitív, mind negatív irányban.

[1] A három technika tárgyalásakor k és n betűket használunk, ami azért nem szerencsés, mert a kedvező esetek és az összes eset jelölésére is ezeket a betűket használtuk. Az alábbi pontokban n és k csupán tetőszleges számokat jelölnek, amik segítségével számláljuk meg az elemi események számát.

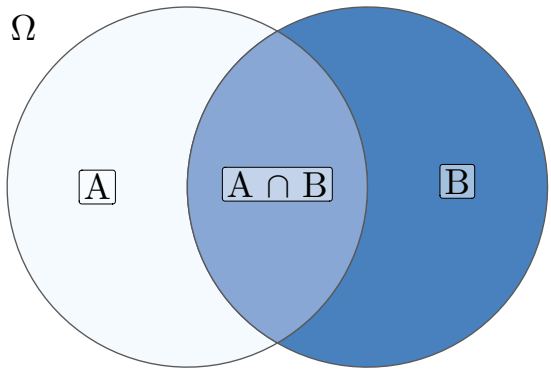
A sztárbefektető a tőzsdeindex következő negyedévben történő 10%-os emelkedéséhez (A) egy adott valószínűséget rendel. A jegybanki alapkamat megváltozása (B) ezt a valószínűséget megváltoztathatja, különösen, ha meglepetésként hat.

A társasjáték vége felé közeledve akkor tudok nyerni, ha egy hatoldalú kockával kétszer dobva több mint 10-et dobok (A). Az első dobás értéke (B) nyilvánvalóan befolyásolja a nyerési esélyeimet.

Ha tehát már tudjuk, hogy B esemény bekövetkezett, akkor a Ω eseménytér egy – a B-n kívüli – része már nem következhet be, innentől B válik az eseménytérré. Ez azt jelenti, hogy az eseményteret le kell szűkítenünk. Mivel eddig minden eseményt az eseménytérben található összes elemi esemény számához hasonlítottunk, az új viszonyítási alap B esemény elemi eseményei lesznek. Azt tehát már tudjuk, hogy egy B-re feltételes valószínűség nevezőjében $\mathbf{P}(B)$ szerepel. A számlálóban pedig azok az események szerepelnek, ahol B mellett A is bekövetkezik, azaz $\mathbf{P}(A \cap B)$. A gondolatmenet alapján levezetett képlet ténylegesen a feltételes valószínűség definíciója:

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \tag{4.9}$$

azaz az A feltéve B esemény valószínűségét a két esemény együttes bekövetkezési valószínűségének és B valószínűségének hányadosaként definiáljuk, feltéve hogy $\mathbf{P}(B) > 0$. A feltételes valószínűség kiszámításának logikáját mutatja be a 4.2. ábra.



4.2. ábra. A feltételes valószínűség illusztrációja

A (4.9) definícióból egyben a fordított feltételes valószínűség is adódik (feltételezve, hogy $\mathbf{P}(A) \neq 0$), azaz

$$\mathbf{P}(B \mid A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} \tag{4.10}$$

hiszen az együttes bekövetkezési valószínűség felcserélhető, $\mathbf{P}(A \cap B) = \mathbf{P}(B \cap A)$.

Folytassuk a fentiek közül a legegyszerűbb példával, legyen tehát

- $A = \{\text{a dobott számok összege több mint } 10\}$,
- $B = \{\text{az első dobás 6-os}\}$.

Ebben az esetben a (4.9) és (4.10) formulák valamennyi elemét egyszerű módszerekkel ki tudjuk számítani, így mintegy „ellenőrizve" a képletek helyességét. Ekkor

- $\mathbf{P}(A) = \frac{3}{36}$, hiszen a $6^2 = 36$ lehetőségből (lásd (4.7)) mindössze 3 sorozat kedvező: $(6,5); (6,6); (5,6)$
- $\mathbf{P}(B) = \frac{1}{6}$, egyszerűen annak a valószínűsége, hogy az első dobás hatos.
- $\mathbf{P}(A \cap B) = \frac{2}{36}$, tehát annak a valószínűségét keressük, hogy nyerünk ÉS elsőre hatost dobunk. Az összes eset száma szintén 36, a kedvező események száma pedig 2: a $(6,5)$ és $(6,6)$ dobás sorozatok.
- $\mathbf{P}(A \mid B) = \frac{2}{6}$, hiszen ha már tudjuk, hogy elsőre 6-ost dobtunk, akkor az összes esemény száma 6, ebből pedig két kedvező esemény van, ha 5-öst, vagy ha újra 6-ost dobunk.
- $\mathbf{P}(B \mid A) = \frac{2}{3}$, hiszen ha már tudjuk, hogy tíznél többet dobtunk, akkor az összes esemény száma 3, ebből pedig két kedvező esemény van.

Behelyettesítve tehát

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{2}{6} = \frac{\frac{2}{36}}{\frac{1}{6}} \quad \mathbf{P}(B \mid A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{2}{3} = \frac{\frac{2}{36}}{\frac{3}{36}}$$

Az esetek nagy részében természetesen nem az a cél, hogy a (4.9) és (4.10) formulák valamennyi elemét kiszámítsuk, ahogy ebben az egyszerű példában történt, hanem pontosan az, hogy általában két tag könnyebben meghatározható, mint a harmadik, mely meghatározásában segítenek a formulák.

Sok esetben épp a (4.9) és (4.10) formulákban szereplő együttes bekövetkezési valószínűségek kiszámítása okoz problémát, azonban ismert valamelyik feltételes valószínűség. Elemi matematikai ismeretek segítségével belátható a valószínűségek szorzási szabálya

$$\mathbf{P}(A \cap B) = \mathbf{P}(A \mid B)\mathbf{P}(B) = \mathbf{P}(B \mid A)\mathbf{P}(A) \tag{4.11}$$

A szorzási szabály kettőnél több esemény esetére is értelmezhető, ez azonban meghaladja tananyagunk kereteit.

Szintén a feltételes valószínűség definícióján alapul két esemény sztochasztikus függetlensége, az A eseményt a B eseménytől függetlennek nevezzük, ha

$$\mathbf{P}(A \mid B) = \mathbf{P}(A) \tag{4.12}$$

A definíció logikus, azt mondja ki, hogy ha B bekövetkezése nem változtatja meg A bekövetkezési valószínűségéről alkotott véleményünket, akkor A esemény független B-től.

Abban az esetben, ha A és B események nem nulla valószínűségűek, belátható, hogy A függetlensége B-től egyben B függetlenségét is jelenti A-tól, azaz a tulajdonság szimmetrikus.

Felhasználva (4.11) és (4.12) definíciókat a függetlenség egy másik, gyakran használt definícióját kapjuk. Az A és B eseményeket tehát függetlennek nevezzük, ha

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \tag{4.13}$$

A (4.13) formulát kimondva, kimondatlanul nagyon gyakran fogjuk használni mind a következő két fejezetben, mind később, a Statisztikai modellezés kurzuson. A függetlenség kettőnél több eseményre is definiálható, jelen tananyagban azonban nem térünk ki erre az esetre.

A teljes valószínűség tétele szintén feltételes valószínűségekre vonatkozó összefüggés. A korábbiakban ((4.1) és (4.2)) már definiáltuk a teljes eseményrendszer fogalmát. A teljes valószínűség tétele azt mondja, hogy ha $A_1, A_2, \ldots, A_n$ események teljes eseményrendszert alkotnak, akkor az eseménytér bármely B eseményének valószínűsége meghatározható a

$$\mathbf{P}(B) = \sum_{i=1}^{n} \mathbf{P}(B \mid A_i)\mathbf{P}(A_i) \tag{4.14}$$

módon.

A feltételes valószínűségek egy újabb fontos alkalmazási területe az ún. Bayes-tétel. Tekintsük (4.11) formulát, amiből egyszerű osztással (feltéve, hogy $\mathbf{P}(B) > 0$)

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B)} \tag{4.15}$$

adódik. A tételben tehát két feltételes valószínűség szerepel. A $\mathbf{P}(A)$ valószínűséget gyakran „a priori”, vagy röviden prior, míg a $\mathbf{P}(A \mid B)$ valószínűséget „a posteriori”, vagy röviden poszterior valószínűségnek nevezzük, utalva arra, hogy az A esemény eredeti, valamint a B esemény bekövetkezése utáni valószínűségeiről van szó. Gyakran

van szükségünk az utóbbi két tétel, azaz (4.14) és (4.15) együttes alkalmazására, hiszen a Bayes-tétel nevezőjében szereplő $\mathbf{P}(B)$ valószínűség gyakran nem ismert közvetlenül. Tegyük fel, hogy a teljes eseményrendszert alkotó események közül A_k poszterior valószínűségére vagyunk kíváncsiak, ekkor a Bayes-tétel egy alternatív formája

$$\mathbf{P}(A_k \mid B) = \frac{\mathbf{P}(B \mid A_k)\mathbf{P}(A_k)}{\sum_{i=1}^{n} \mathbf{P}(B \mid A_i)\mathbf{P}(A_i)} \tag{4.16}$$

4.4. Excel tippek

Hasznos Excel függvények:

- permutáció: FAKT
- variáció: VARIÁCIÓK, VARIÁCIÓK.ISM
- kombináció: KOMBINÁCIÓK, KOMBINÁCIÓK.ISM

5. fejezet

Diszkrét valószínűségi változó

A 4. fejezetben események valószínűségének kiszámításával foglalkoztunk. A gyakorlatban a véletlen kísérletek kimeneteleit számokkal jellemezzük, ezt a minden kimenetelhez egy számot rendelő függvényt valószínűségi változónak hívjuk, és jellemzően X-szel jelöljük. A statisztikai változó fogalomhoz hasonlóan pl. Y-nal jelölünk egy további valószínűségi változót, vagy ha sok változóval dolgozunk a későbbiekben, akkor gyakori az $X_1, X_2, \dots, X_k$ jelölés is. Fontos azonban megjegyezni, hogy – bár nagyon hasonlók – nem keverendő össze a 2. fejezetben megismert sokaságot leíró változó és a valószínűségi változó. Utóbbi egy absztraktabb fogalom, bár leírására hasonló fogalmakat is használunk majd, mint egy statisztikai változó esetén.

Egy adott eseménytérhez (végtelen) sok valószínűségi változó definiálható. Amennyiben a véletlen kísérlet a jövő heti (ötös)lottóhúzáson való részvétel egy szelvénnyel, akkor X lehet például az elért találatok száma, egy másik Y valószínűségi változó lehet az elért nyeremény nagysága. Egy másik példában a véletlen kísérletek lehetnek egy call-centerbe beérkező hívások. Ekkor X lehet az egységnyi időtartam (pl. egy óra) alatt beérkező hívások száma, Y pedig a hívások hossza, stb.

A statisztikai mintavétel is egy véletlen kísérletnek fogható fel és ne feledjük, hogy a valószínűségszámítás elsősorban a következtetéses statisztika megalapozásához ad számunkra segítséget. Valószínűségi változó lehet a mintába került első megfigyelés testmagassága, vagy egy másik esetben a mintába került vállalatok átlagos árbevétele is.

Attól függően, hogy a valószínűségi változó értékkészlete milyen számosságú, megkülönböztetünk:

- diszkrét
  - véges számosságú

  - megszámlálhatóan végtelen számosságú

- folytonos

valószínűségi változókat.

A diszkrét valószínűségi változók jellemzően számlálás útján keletkeznek, így a lehetséges értékek a természetes számok körében keresendők. A folytonos valószínűségi változók jellemzően valamilyen mérési folyamat eredményeképp állnak elő, így a lehetséges értékek általában a valós számok valamely részhalmazán találhatók, azaz kontinuum számosságú potenciális kimenettel rendelkeznek.

Az előző példák közül az ötöslottón elért találatok száma egyértelműen véges számosságú, diszkrét valószínűségi változó, hiszen a lehetséges találatok száma mindössze hat: $0, 1, \ldots, 5$. Az egy óra alatt beérkező telefonhívások száma megszámlálhatóan végtelenül sok különböző értéket vehet fel, azaz nincs egy olyan felső határ, mint a lottó találatok esetén, a beérkező hívások száma lehet $0, 1, 2, \ldots$. A hívások hossza, vagy a mintába került vállalatok átlagos árbevétele folytonos változó, hiszen bizonyos korlátok között bármilyen, nemnegatív valós értéket felvehetnek, a lehetséges értékek halmazának felsorolása lehetetlen lenne.

5.1. Súly- és eloszlásfüggvény

A fejezetben a diszkrét valószínűségi változók jellemzésére szolgáló két eszközt, a súlyfüggvényt és az eloszlásfüggvényt mutatjuk be.

Egy X diszkrét valószínűségi változó súlyfüggvénye alatt a

$$\mathbf{P}(X = x_k) = p_k \tag{5.1}$$

függvényt értjük, ami egy tetszőleges valós számhoz az érték bekövetkezési valószínűségét rendeli. A hozzárendelést megadhatjuk az összes lehetséges érték, és a valószínűségek felsorolásával, vagy akár képletek segítségével is, ahogy azt a 5.3. pontban látni fogjuk.

Diszkrét valószínűségi változók esetén az x_k lehetséges értékeket gyakran k-val jelöljük, jelezve, hogy csak természetes számok esetén vesz fel a függvény 0-tól különböző értéket. Emiatt a fejezetben gyakran fogunk élni a $\mathbf{P}(X = k) = p_k$ egyszerűsített jelöléssel.

A p_k valószínűségeknek két feltételt kell teljesíteniük ahhoz, hogy egy diszkrét eloszlást írjanak le:

- $p_k \geq 0$, azaz a valószínűségek nemnegatívak,
- $\sum_k p_k = 1$, mivel a lehetséges értékek teljes eseményrendszert alkotnak.

Ha a szóban forgó kísérletet nagyon sokszor elvégezzük, akkor a megfigyelt relatív gyakoriságok a valószínűségekhez fognak tartani. A súlyfüggvény ábrázolására jellemzően oszlop-, vagy pálcikadiagramot használhatunk.

Az egyik legegyszerűbb diszkrét valószínűségi változó a kockadobás példájához tartozik, ebben az esetben a súlyfüggvény

$$p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$$

ami azt jelenti, hogy ha nagyon sok kockadobást végeznénk el, akkor azt várjuk, hogy minden dobott szám megközelítőleg $\frac{1}{6}$ relatív gyakorisággal fog előfordulni.

Egy X diszkrét valószínűségi változó eloszlásfüggvénye alatt a

$$F(x) = \mathbf{P}\left(X \leq x\right) \tag{5.2}$$

függvényt értjük, ami egy tetszőleges valós számhoz az $X \leq x$ esemény bekövetkezési valószínűségét rendeli. Az eloszlásfüggvény a diszkrét esetben nem annyira fontos, hiszen a súlyfüggvény segítségével könnyedén meghatározhatók a szükséges valószínűségek, de a folytonos eloszlásoknál definiált eloszlásfüggvény kiemelten fontos lesz a következő fejezetben, ezért már itt is megemlítjük. Diszkrét esetben az eloszlásfüggvény értéke az a helyen $(F\left(a\right))$ egyszerűen azon súlyfüggvény-értékek összege, melyekre az $x_k \leq a$ feltétel teljesül.

Az egyik legegyszerűbb diszkrét valószínűségi változó a kockadobás példájához tartozik, ebben az esetben az eloszlásfüggvény néhány pontja például:

- $F(2) = \mathbf{P}\left(X \leq 2\right) = \mathbf{P}\left(X = 1\right) + \mathbf{P}\left(X = 2\right) = p_1 + p_2 = \frac{1}{3}$
- $F(4{,}5) = \mathbf{P}\left(X \leq 4{,}5\right) = p_1 + p_2 + p_3 + p_4 = \frac{2}{3}$

A második példa azt is mutatja, hogy az eloszlásfüggvény minden x-re, nem csak egész számokra értelmezett.

Az eloszlásfüggvény monoton növekvő, hiszen x növelésével a $\mathbf{P}\left(X \leq x\right)$ valószínűség nem csökkenhet. A diszkrét értékek között a függvény szakaszonként konstans, hiszen például az $F(4{,}5)$ és az $F(4{,}8)$ valószínűségek megegyeznek. A függvény határértéke $-\infty$-ben 0, ∞-ben pedig egy, azaz minél nagyobb x-re a valószínűség minden határon túl megközelíti (vagy eléri) az egyet. Ahogy a súlyfüggvény statisztikai megfelelője a relatív gyakoriság, úgy az eloszlásfüggvény esetében ezt a szerepet a kumulált relatív gyakoriság tölti be.

Könnyen belátható, hogy az eloszlásfüggvény akkor igazán hasznos, ha a kiszámítandó valószínűség nem $\mathbf{P}\left(X = x\right)$ típusú, hanem egy nyílt, vagy zárt intervallumba esés

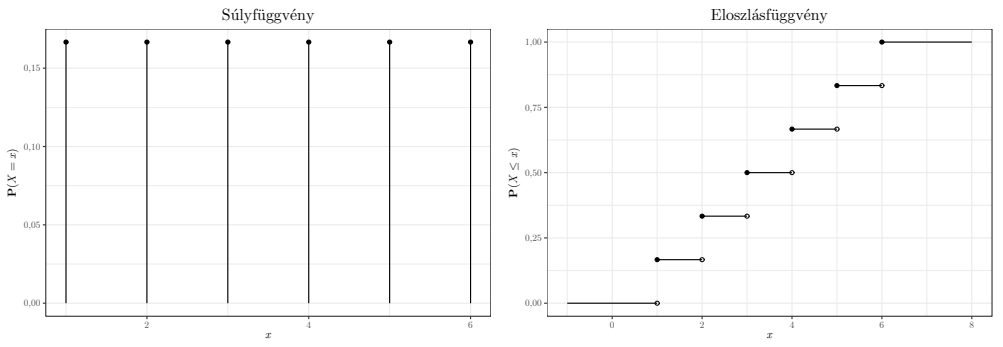
valószínűségét szeretnénk kiszámítani. Az alábbiakban bizonyítás nélkül mutatjuk be azokat az eseteket egy táblázatban, melyek esetén az eloszlásfüggvény is alkalmazható. Természetesen minden valószínűség megadható a megfelelő súlyfüggvényértékek összeadásával is.

5.1. táblázat: Eloszlásfüggvény alkalmazási lehetőségei diszkrét esetben (a és b valós számok)

keresett valószínűség	alkalmazható formula
$\mathbf{P}(X \leq a)$	$F(a)$
$\mathbf{P}(X < a)$	$F(a-1)$
$\mathbf{P}(X \geq a)$	$1 - F(a-1)$
$\mathbf{P}(X > a)$	$1 - F(a)$
$\mathbf{P}(a < X \leq b)$	$F(b) - F(a)$
$\mathbf{P}(a \leq X \leq b)$	$F(b) - F(a-1)$
$\mathbf{P}(a < X < b)$	$F(b-1) - F(a)$
$\mathbf{P}(a \leq X < b)$	$F(b-1) - F(a-1)$

A súlyfüggvény és az eloszlásfüggvény szoros kapcsolatban állnak egymással, ahogy láttuk, az eloszlásfüggvény előállítható a súlyfüggvény megfelelő értékeinek összeadásával.

A kockadobás példájához tartozó súlyfüggvény és az eloszlásfüggvény az 5.1. ábrán látható. A súlyfüggvény azt mutatja meg, hogy a valószínűségek hol helyezkednek el, míg az eloszlásfüggvény egy lépcsős, szakaszonként konstans függvény képét mutatja. A két függvény egymással szorosan összefügg, az egyikből kiszámítható a másik, összeadás, illetve ennek inverz művelete, kivonás segítségével: az eloszlásfüggvényben ott van ugrás, ahol lehetséges érték van, ennek a lehetséges értéknek a valószínűsége pedig az ugrás nagyságával azonos.



5.1. ábra. A kockadobás valószínűségi változójának súly- és eloszlásfüggvénye

5.2. Momentumok

A valószínűségi változókat – csakúgy mint a statisztikai sokaságokat – gyakran jellemezzük azok momentumaival. Végtelen sok momentum definiálható, ebben a tankönyvben csupán a legfontosabb kettővel fogunk megismerkedni: a várható értékkel és a varianciával (illetve a szórással).

Egy X diszkrét valószínűségi változó várható értékén a

$$\mathbf{E}(X) = \sum_k x_k \cdot p_k \tag{5.3}$$

kifejezést értjük, ahol k a valószínűségi változó összes lehetséges értékét befutja. A várható érték a súlyfüggvény „tömegközéppontjaként" értelmezhető, a statisztikai megfelelője a számtani átlag. Ez azt jelenti, hogy a kísérletet sokszor ismételve a kapott értékek átlaga a várható értékhez fog tartani.

Visszatérve a kockadobás példájára a várható érték az alábbi módon határozható meg:

$$\mathbf{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3{,}5$$

Ez tehát egyrészt azt jelenti, hogy a 3,5 pontban van a súlyfüggvény tömegközéppontja, illetve nagyon sokszor elvégezve a kockadobás kísérletét, a kapott értékek átlaga egyre közelebb fog kerülni a 3,5 értékhez.

Egy X diszkrét valószínűségi változó varianciáján a

$$\mathbf{D}^2(X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \sum_k (x_k - \mathbf{E}(X))^2 \cdot p_k \tag{5.4}$$

kifejezést értjük, ahol k a valószínűségi változó összes lehetséges értékét befutja. A variancia gyökét a valószínűségi változó szórásának nevezzük és $\mathbf{D}(X)$ módon jelöljük. A szórás azt mutatja meg, hogy a kísérletet sokszor ismételve a kapott értékek átlagosan milyen messze fognak elhelyezkedni a várható értéktől (és egyben a saját átlaguktól).

A kockadobás példája esetén a variancia az alábbi módon határozható meg:

$$\mathbf{D}^2(X) = (1-3{,}5)^2\frac{1}{6} + (2-3{,}5)^2\frac{1}{6} + \cdots + (6-3{,}5)^2\frac{1}{6} = \frac{35}{12}$$

Ez tehát azt jelenti, hogy a kockadobást sokszor elvégezve a kapott számok (statisztikai) szórása egyre inkább megközelíti a $\sqrt{\frac{35}{12}}$ értéket.

5.3. Nevezetes diszkrét eloszlások

Sok esetben olyan jelenségeket, kísérleteket vizsgálunk, melyek egy adott sémát követnek, így a kiszámításukhoz szükséges – egyszer már kitalált – képleteket újra tudjuk alkalmazni kissé eltérő módon. Amennyiben például kigondoltuk, hogy az ötöslottón (ahol 90 számból 5-öt kell eltalálni) elért találatok számát milyen módon kell kiszámítani, a hatoslottó esetén már nem kell újra kigondolnunk a számítás menetét, csupán a megfelelő értékeket (későbbi szóhasználattal paramétereket) kell módosítanunk. Az ilyen, adott formula segítségével meghatározható valószínűségekkel rendelkező valószínűségi változókat nevezetes eloszlásoknak nevezzük. A nevezetes eloszlások további előnye, hogy a valószínűségek kiszámításán túl a momentumok meghatározása is jóval egyszerűbb, mint az általános (5.3), vagy (5.4) képletek segítségével.

Az alábbiakban a legegyszerűbb, leggyakrabban használt öt diszkrét nevezetes valószínűségi változót mutatjuk be, de ezen kívül is sok fontos eloszlás létezik, amelyre tananyagunk nem terjed ki. Az alábbi eloszlások logikájának megértése azonban megkönnyíti más eloszlások alkalmazását is. Azt, hogy egy X valószínűségi változó adott eloszlást követ, azt $\sim$ módon jelöljük, amit zárójelben a már említett paraméter(ek) követ(nek).

5.3.1. Bernoulli-eloszlás

A Bernoulli-eloszlás (más néven indikátor eloszlás) a legegyszerűbb diszkrét eloszlás, mindössze két lehetséges értéke van: 0 és 1. Amennyiben egy adott, számunkra fontos, p valószínűségű esemény bekövetkezik, akkor $X = 1$ értéket vesz fel, ellenkező esetben pedig $X = 0$-t. Vegyük észre, hogy egyetlen paraméter, a p értéke teljesen meghatározza az egyes lehetséges értékek valószínűségét. Az $X \sim \mathrm{Bern}(p)$ jelölést alkalmazva tehát a súlyfüggvény:

$$\mathbf{P}(X = 1) = p \text{ és } \mathbf{P}(X = 0) = 1 - p = q \tag{5.5}$$

Vegyük észre, hogy a két egyszerű formula helyett a valószínűségeket a

$$\mathbf{P}(X = k) = p^k \, (1 - p)^{(1-k)}, \quad k \in \{0, 1\} \tag{5.6}$$

módon is felírhatjuk, amely képlet már sokkal inkább hasonlít a később megismerendő eloszlásoknál látottakra. Könnyen ellenőrizhető, hogy $k = 0$ esetre a valószínűség $1-p$,

míg a $k = 1$ esetre a valószínűség p. A (5.6) formulához nagyon hasonló képlettel a későbbiekben még találkozni fogunk.

Ahogy említettük, a nevezetes eloszlások előnye, hogy a momentumok meghatározásához nem kell az általános képleteket használni, levezethetők speciális, csak az adott eloszlásra alkalmazható momentum-formulák is. Mivel ezt a Bernoulli-eloszlás esetén könnyű megtenni, ezért ezeket a levezetéseket meg is tesszük.

Meg kívánjuk tehát határozni az $X \sim \text{Bern}(p)$ valószínűségi változó várható értékét, amihez a (5.3) általános definíciót használjuk fel, mely szerint az összes lehetséges érték és a hozzájuk tartozó valószínűségek szorzatösszege adja meg a keresett értéket:

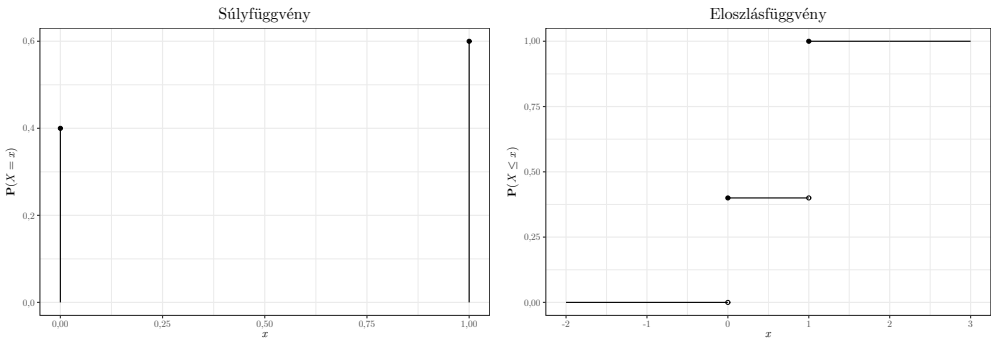
$$\mathbf{E}(X) = \sum_k x_k \cdot p_k = 0 \cdot (1-p) + 1 \cdot p = p \tag{5.7}$$

Hasonlóan egyszerű levezetni az $X \sim \text{Bern}(p)$ valószínűségi változó varianciáját a (5.4) általános definíció felhasználásával:

$$\mathbf{D}^2(X) = \sum_k (x_k - \mathbf{E}(X))^2 \cdot p_k = (0-p)^2(1-p) + (1-p)^2 p = p(1-p) \tag{5.8}$$

Azt kaptuk tehát, hogy a p paraméter ismeretében mind a várható érték, mind a variancia könnyedén meghatározható. A Bernoulli-eloszlás klasszikus példája a pénzérme feldobás, ahol a p valószínűségű esemény lehet az, hogy fejet dobunk. Amennyiben $p = 0{,}5$, azt mondjuk, hogy a pénzérme fair.

A 5.2. ábra a Bernoulli-eloszlás súly- és eloszlásfüggvényét mutatja be.



5.2. ábra. A Bernoulli-eloszlás súly- és eloszlásfüggvénye, $p = 0{,}6$

5.3.2. Diszkrét egyenletes eloszlás

Amennyiben egy diszkrét valószínűségi változó n darab, egymást követő egész számmal jellemezhető lehetséges értékkel rendelkezik, melyek azonos valószínűséggel

következnek be, egyenletes eloszlásról beszélünk. Az eloszlás pontos leírásához két paraméterre van szükség, amiket a, a lehetséges értékek alsó határa és b, a lehetséges értékek felső határa jelöl. Az $X \sim \mathcal{U}(a, b)$ jelölést alkalmazva tehát a súlyfüggvény:

$$\mathbf{P}(X = k) = \frac{1}{n} \tag{5.9}$$

ahol $n = b - a + 1$. A diszkrét egyenletes eloszlás momentumai a paraméterek segítségével könnyen megadhatók. A várható érték és a variancia a

$$\mathbf{E}(X) = \frac{a+b}{2}, \qquad \mathbf{D}^2(X) = \frac{n^2-1}{12} \tag{5.10}$$

formulák segítségével számítható. A legkézenfekvőbb példa a már többször említett kockadobás, $a = 1, b = 6$ paraméterekkel. A (5.10) képletek lehetőséget adnak a 5.2. fejezet alapján ellenőrizni a kapott várható érték és szórás eredményeket. Könnyen belátható, hogy ugyanazokat az eredményeket kapjuk ($\mathbf{E}(X) = \frac{7}{2}, \mathbf{D}^2(X) = \frac{6^2-1}{12} = \frac{35}{12}$), mint az általános képletek segítségével, de jóval egyszerűbben kaptuk meg az eredményeket, kihasználva hogy tudjuk, a kockadobás egyenletes eloszlást követ. A szóban forgó eloszlás ábráját már bemutattuk a 5.1. ábrán, így itt azt nem ismételjük meg, más paraméterű esetekre is hasonló ábrát kapnánk.

5.3.3. Binomiális eloszlás

Az egyik leggyakrabban előforduló eset és emiatt az egyik legfontosabb valószínűségi változó a mintavételhez kötődik. A binomiális eloszlás alkalmazásának feltételei az alábbiak:

- előre adott, n számú megfigyelést végzünk,
- valamennyi megfigyelés pontosan két kategóriába sorolható (sikeres, vagy sikertelen)
- minden egyes megfigyelés esetén a siker valószínűsége állandó, amit π jelöl,
- a megfigyelések egymástól függetlenek (pl. az első mintaelem kiválasztásának eredménye nem befolyásolja a második és azt követő mintaelemek eloszlását, vagy más szavakkal, nem változtatja meg a sokaság összetételét),
- a vizsgálatunk tárgya az összes sikeres megfigyelés száma a mintában (k).

Ezek a feltételek két gyakran előforduló gyakorlati mintavételi esetben is érvényesek:

- végtelen sokaságból vett, illetve

- véges sokaságból visszatevéssel vett

minta esetén. Általánosságban a fenti két eset az alábbiakkal írható le.

- Végtelen alapsokaságból egymástól függetlenül vett n elemű mintába az elemenként π valószínűségű adott tulajdonságú objektumból k darab kerül be.
- Véges, N elemű alapsokaságban K darab adott tulajdonsággal rendelkező elem van. Ebből a sokaságból visszatevéssel vett n elemű mintába k darab adott tulajdonságú elem kerül be. Ekkor a $\pi = \frac{K}{N}$ jelöléssel élünk.

Jelölje a X valószínűségi változó a mintába kerülő, adott tulajdonsággal rendelkező objektumok számát (vagy röviden a sikerek számát). Mindkét fenti esetben azt mondjuk, hogy $X \sim \mathrm{Bin}(n, \pi)$, azaz X binomiális eloszlást követ, n és π paraméterekkel. A súlyfüggvényt az alábbi

$$\mathbf{P}(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} \tag{5.11}$$

formula írja le. A képlet emlékeztet a (5.6) egyenletben látott kifejezésre. A kapcsolat a két eloszlás között igen szoros, hiszen míg a Bernoulli-eloszlás egy kísérletet írt le, addig a binomiális eloszlás n egymás követő kísérletet. Formálisan is könnyű belátni, hogy n darab p paraméterű Bernoulli-eloszlású véletlen változó összege n, p paraméterpárral rendelkező binomiális eloszlást ad. Vagy a másik oldalról, a Bernoulli-eloszlás a binomiális eloszlás speciális esete, arra az esetre, amikor $n = 1$.

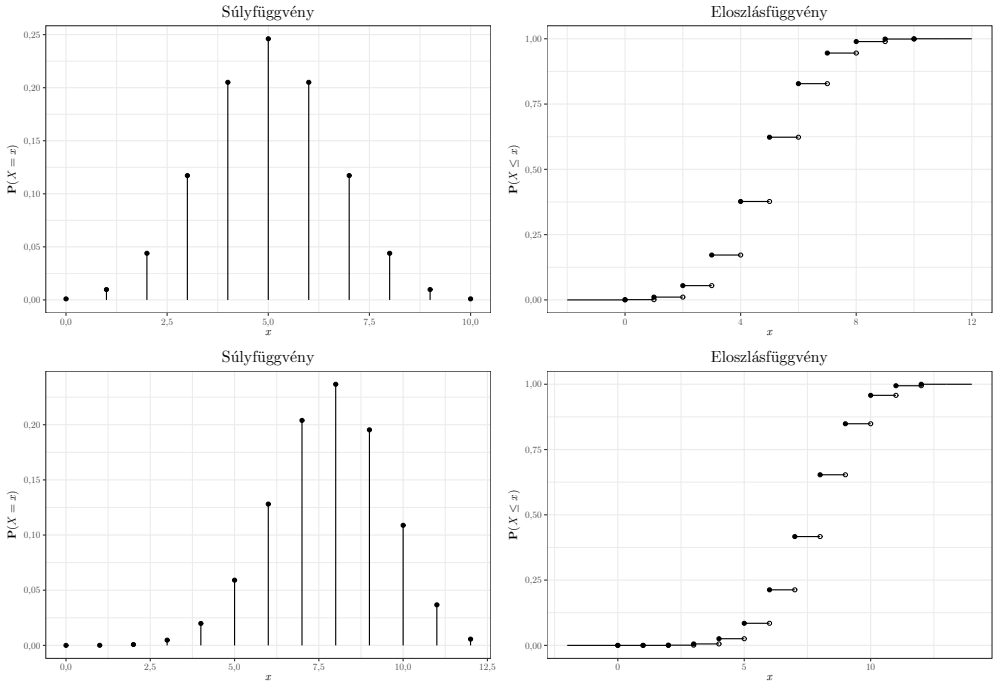
A binomiális eloszlás paramétereinek ismeretében a megismert momentumok könnyedén számíthatók:

$$\mathbf{E}(X) = n\pi, \qquad \mathbf{D}^2(X) = n\pi(1 - \pi) \tag{5.12}$$

A binomiális eloszlás súly- és eloszlásfüggvénye változatos alakot ölt a paraméterek függvényében. Az 5.3. ábra két paraméterpár esetén mutatja be a függvényeket. Az első esetben a $\pi = 0{,}5$ választás miatt a sűrűségfüggvény szimmetrikus, illetve az eloszlásfüggvény is középpontosan szimmetrikus, míg a 0,5 feletti paraméter baloldali (hiszen a magas k értékekhez fog magas valószínűség tartozni), a 0,5 alatti paraméter jobboldali aszimmetriát (hiszen az alacsony k értékekhez fog magas valószínűség tartozni) okoz a súlyfüggvényben.

A binomiális eloszlásra a klasszikus példa a pénzérme feldobása n alkalommal, X pedig jelölje a dobott fejek számát (azaz a fejeket tekintjük sikeres eseménynek). Ennek a modellezésére fair érme esetén az $n, \pi = 0{,}5$ paraméterpárral rendelkező binomiális eloszlás felel meg.

Itt jegyezzük meg, hogy a gyakorlatban a nagyon nagy alapsokaságot is végtelenként kezeljük. Gyakorlati szempontból egy Magyarország népességéből vett viszonylag



5.3. ábra. A binomiális eloszlás súly- és eloszlásfüggvénye, $n = 10, \pi = 0{,}5$, illetve $n = 12, \pi = 0{,}65$

kis elemszámú visszatevés nélküli minta is függetlennek tekinthető, gyakorlati szempontból nem változik meg a sokaság összetétele jelentősen attól, hogy az első elemeket már kihúztuk, és nem tettük őket vissza. Abban az esetben viszont, ha a minta a sokaság méretéhez képest nem elhanyagolható méretű, és a kiválasztás visszatevés nélkül történik, a binomiális eloszlás nem alkalmazható.

5.3.4. Hipergeometriai eloszlás

Az előző pontban bemutatott binomiális eloszlás a mintában előforduló sikeres esetek számához rendel valószínűségeket több különböző esetre vonatkozóan is, de a véges alapsokaságból visszatevés nélkül vett minták esetén más eloszlással kell megismerkednünk. Amennyiben annak a valószínűségére vagyunk kíváncsiak, hogy N elemű véges alapsokaságból visszatevés nélkül vett n elemű mintába, az összesen K tulajdonságú objektumból pontosan k db kerül be, akkor a hipergeometriai, vagy hipergeometrikus eloszlásról beszélünk. A $X \sim \mathrm{Hip}(n, K, N)$ eloszlás súlyfüggvénye a

$$\mathbf{P}(X = k) = \frac{\binom{N-K}{n-k}\binom{K}{k}}{\binom{N}{n}} \tag{5.13}$$

formulával írható le. A számítás logikája tulajdonképp egyszerű, az összes lehetséges $\binom{N}{n}$ mintából azoknak a számát szeretnénk meghatározni, amiben pontosan k adott tulajdonságú szerepel. Ennek meghatározására egyrészt a mintánkba pontosan $\binom{K}{k}$ módon választhatunk az adott tulajdonsággal rendelkezők közül, majd – hogy teljes legyen a minta – ki kell választanunk az adott tulajdonsággal nem rendelkező mintaelemeket, amit $\binom{N-K}{n-k}$ módon tehetünk meg. Mivel minden sikeres csoporthoz minden nem sikeres csoportot választhatjuk, ezért a lehetőségek számát a számlálóban összeszorozzuk.

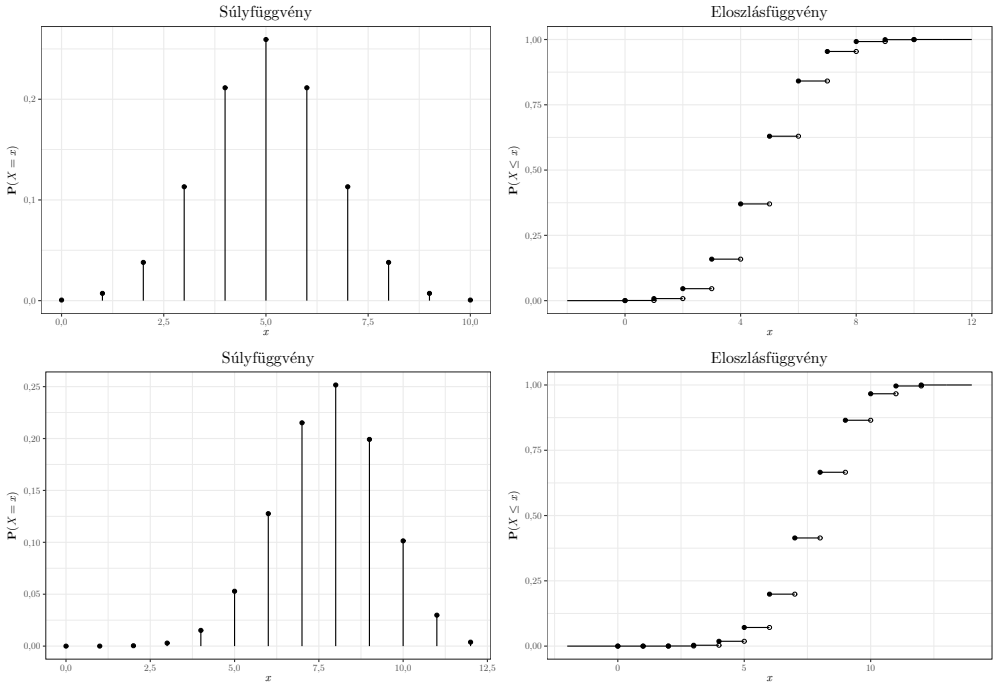
A $X \sim \mathrm{Hip}(n, K, N)$ eloszlás momentumai a paraméterek ismeretében könnyen számíthatóak.

$$\mathbf{E}(X) = n\frac{K}{N}, \qquad \mathbf{D}^2(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1} \tag{5.14}$$

Amennyiben ezt a formulát összehasonlítjuk a binomiális eloszlás várható értékével és varianciájával a (5.12) egyenletben, és a hipergeometriai eloszlásnál is bevezetjük a $\pi = \frac{K}{N}$ jelölést, akkor azt látjuk, hogy a két várható érték megegyezik, a varianciák pedig abban különböznek, hogy a hipergeometriai eloszlás esetén szerepel egy $\frac{N-n}{N-1}$ tényező. Mivel ezt a különbséget a véges sokaság miatt figyelhetjük meg, a tényező neve véges szorzó. A későbbi fejezetekben még találkozni fogunk a kifejezéssel.

A hipergeometriai eloszlás klasszikus példája a lottón elért találatok valószínűsége, ami olyan szempontból speciális példa, hogy a sokaságban a kedvező esetek száma is $K = 5$ (hiszen öt nyerőszám van) és a mintaelemszám is $n = 5$ (hiszen öt számra kell tippelni). Ez a hipergeometriai eloszlás egyéb felhasználásai esetén nem feltétlenül van így.

A hipergeometriai eloszlás súlyfüggvénye és eloszlásfüggvénye két paraméterezés mellett az 5.4. ábrán látható.



5.4. ábra. A hipergeometriai eloszlás súly- és eloszlásfüggvénye, $n = 10, K = 50, N = 100$, illetve $n = 12, K = 65, N = 100$

5.3.5. Poisson-eloszlás

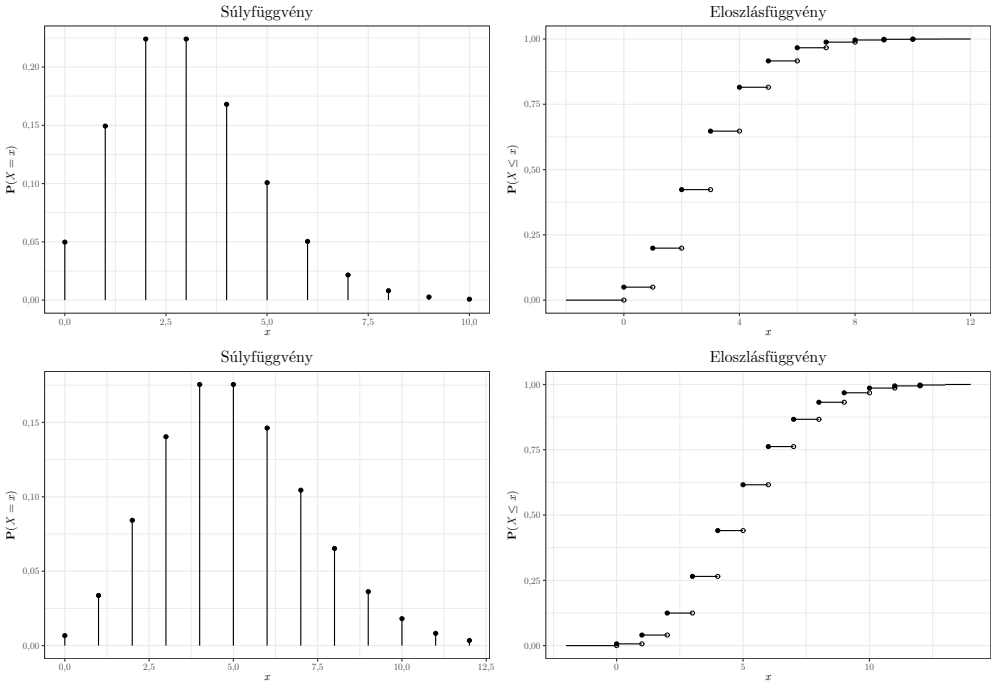
A Poisson-eloszlás nem közvetlenül a mintavételhez kapcsolódik, mégis hasznos, gyakran alkalmazott eloszlás. Az eloszlás egy rögzített időtartam alatt, vagy felszínen megfigyelhető események számának valószínűségét írja le, ha az események egymástól függetlenül következnek be, állandó ráta mellett. A Poisson eloszlás egyetlen paraméterrel írható le, a $X \sim \mathrm{Poi}(\lambda)$ eloszlás súlyfüggvénye a

$$\mathbf{P}(X = k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} \tag{5.15}$$

formulával írható le. A Poisson-eloszlás különleges abból a szempontból, hogy a várható értéke és varianciája megegyezik.

$$\mathbf{E}(X) = \lambda, \qquad \mathbf{D}^2(X) = \lambda \tag{5.16}$$

A klasszikus példa a Poisson-eloszlásra az óránként egy call-centerbe érkező hívások száma. Az első közismert alkalmazása a porosz hadseregben évente halálra rúgott katonák számának modellezése volt. A 5.5. ábra két Poisson-eloszlást mutat be



5.5. ábra. A Poisson-eloszlás súly- és eloszlásfüggvénye, $\lambda = 3, \lambda = 5$

5.4. Excel tippek

Hasznos Excel függvények:

- SZORZATÖSSZEG (a várható érték meghatározásához súlyfüggvény alapján)
- BINOM.ELOSZL, BINOM.INVERZ
- HIPGEOM.ELOSZLÁS
- POISSON.ELOSZLÁS

Az Excel ELOSZLÁS függvényei egyaránt alkalmasak a súly-, és eloszlásfüggvények értékeinek (valószínűségek) kiszámítására, az első paraméterben megadott k értékének függvényében. Odafigyelést igényel a függvények paramétereinek helyes

megadása, különösen ha a függvénybe nem számokat, hanem abszolút és relatív hivatkozásokat írunk!

Hasznos Excel funkciók:

- A súly- és eloszlásfüggvény ábrázolása oszlopdiagramon 

6. fejezet

Folytonos valószínűségi változó

A 4. fejezetben események valószínűségének kiszámításával foglalkoztunk, majd a 5. fejezetben diszkrét valószínűségi változókkal. Amint azt láttuk, a gyakorlatban a véletlen kísérletek kimeneteleit számokkal jellemezzük, ezt a minden kimenetelhez egy számot rendelő függvényt valószínűségi változónak hívjuk és jellemzően X-szel jelöljük. A statisztikai változó fogalomhoz hasonlóan pl. Y-nal jelölünk egy további valószínűségi változót, vagy ha sok változóval dolgozunk a későbbiekben, akkor gyakori az $X_1, X_2, \dots, X_k$ jelölés is. Fontos azonban megjegyezni, hogy – bár nagyon hasonlók – nem keverendő össze a 2. fejezetben megismert sokaságot leíró változó és a valószínűségi változó. Utóbbi egy absztraktabb fogalom, bár leírására hasonló fogalmakat is használunk majd, mint egy statisztikai változó esetén.

Attól függően, hogy a valószínűségi változó értékkészlete milyen számosságú, megkülönböztetünk a tananyagunkban:

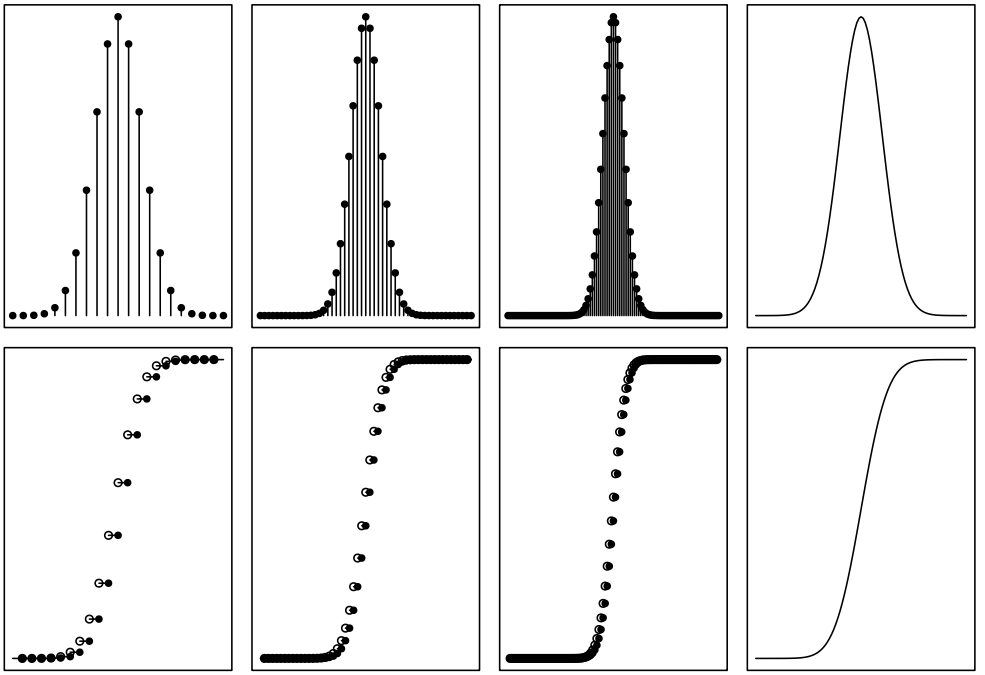
- diszkrét
  - véges számosságú
  - megszámlálhatóan végtelen számosságú
- folytonos

valószínűségi változókat. A fenti besorolás nem teljes, vannak például vegyes eloszlások is, de ezeket nem tárgyaljuk.

A diszkrét valószínűségi változókkal ellentétben a folytonos valószínűségi változók jellemzően valamilyen mérési folyamat eredményeképp állnak elő, így a lehetséges értékek általában a nemnegatív valós számok valamely véges, vagy végtelen részhalmazán találhatók. Ennek megfelelően a folytonos valószínűségi változók

kontinuum számosságú potenciális kimenettel rendelkeznek. Pontosan amiatt, mert a lehetséges értékek számossága nem megszámlálhatóan végtelen, az egyes elemi események valószínűségének szükségszerűen nullának kell lennie, azaz minden folytonos eloszlásra $\mathbf{P}(X = x) = 0$ bármely x értékre.

A 6.1. ábra azt mutatja be, hogy ha végtelenül sűrűen helyezkednek el a lehetséges értékek egy intervallumon, és teljesül a $\mathbf{P}(X = x) = 0$ feltétel, akkor a súlyfüggvény és az eloszlásfüggvény is folytonossá válik.



6.1. ábra. Átmenet a diszkrét és folytonos valószínűségi változók között

A súlyfüggvény, amely definíció szerint a $\mathbf{P}(X = x)$ valószínűséget adja meg diszkrét esetben, folytonos esetben értelmét veszti, a fentiek miatt. A súlyfüggvény helyett a sűrűségfüggvény fogalmát fogjuk bevezetni a következő alpontban, az eloszlásfüggvény szerepe azonban nem változik. Mivel tetszőleges valószínűség meghatározásában a sűrűségfüggvény közvetlenül nem, csak az eloszlásfüggvény segít folytonos esetben, ezért gyakran elsőként az eloszlásfüggvényt definiáljuk.

6.1. Eloszlás- és sűrűségfüggvény

A fejezetben a folytonos valószínűségi változók jellemzésére szolgáló két eszközt, az eloszlásfüggvényt és a sűrűségfüggvényt mutatjuk be.

Egy X folytonos valószínűségi változó eloszlásfüggvénye alatt a

$$F(x) = \mathbf{P}\left(X \leq x\right) \tag{6.1}$$

függvényt értjük, ami egy tetszőleges valós számhoz az $X \leq x$ esemény bekövetkezési valószínűségét rendeli. Ahogy azt már említettük, a folytonos valószínűségi változók esetében a valószínűségek kiszámításához minden esetben az eloszlásfüggvényt fogjuk használni, ezért érdemes a lehetséges változatokat felsorolni:

- $\mathbf{P}\left(X \leq a\right) = \mathbf{P}\left(X < a\right) = F(a)$
- $\mathbf{P}\left(X \geq a\right) = \mathbf{P}\left(X > a\right) = 1 - F(a)$
- $\mathbf{P}\left(a \leq X \leq b\right) = \mathbf{P}\left(a < X \leq b\right) = \mathbf{P}\left(a \leq X < b\right) = \mathbf{P}\left(a < X < x\right) = F(b) - F(a)$

Amint az a fentiekből is látható, folytonos esetben az események esetén az egyenlőségek nem relevánsak, hiszen a definíció szerint az elemi események valószínűsége 0.

Az eloszlásfüggvény definíciójából adódóan az 0 és 1 közötti értéket vehet fel, monoton növekvő, folytonos függvény. Határértéke $\lim_{x\to-\infty} F(x) = 0$, illetve $\lim_{x\to\infty} F(x) = 1$.

Egy X folytonos valószínűségi változó sűrűségfüggvénye alatt az f függvényt értjük, ha

$$F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u \tag{6.2}$$

amiből szokásos feltételek mellett következik, hogy

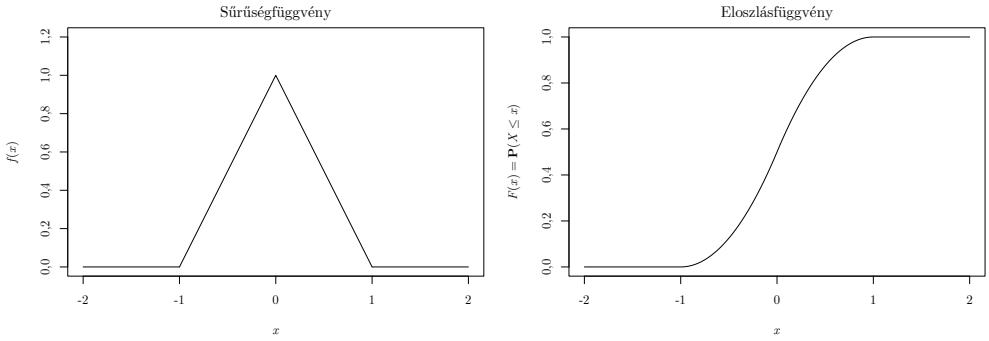
$$f(x) = F'(x) \tag{6.3}$$

A sűrűségfüggvényre igaz, hogy $f(x) \geq 0$, illetve $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$. A teljes görbe alatti terület tehát egységnyi, ami a valószínűséget reprezentálja. Fontos azonban hangsúlyozni, hogy a sűrűségfüggvény $f(x)$ értéke nem valószínűségként értelmezendő, hanem csupán a valószínűség sűrűsödéséről. Matematikailag ez azt jelenti, hogy a sűrűségfüggvény x kis sugarú környezetében a valószínűség sűrűsödését írja le: $\mathbf{P}(x < X < x + \Delta x) \approx f(x)\Delta x$.

A görbe alatti terület tehát egységnyi, a (zárt, vagy nyílt) intervallumba esés valószínűségét a sűrűségfüggvény segítségével integrál segítségével határozhatjuk meg, azaz $\mathbf{P}(a < X < b) = \int_a^b f(x)\mathrm{d}x$.

A sűrűség- és eloszlásfüggvény – hasonlóan a diszkrét eloszlásokhoz – egyértelmű kapcsolatban van egymással. Ahogy diszkrét esetben összegzéssel, illetve kivonással meghatározható egyik függvényből a másik, úgy folytonos esetben az integrálás, illetve a deriválás művelete segít a meghatározásban.

A 6.2. ábrán egy egyszerű folytonos eloszlás sűrűségfüggvénye, illetve a hozzá tartozó eloszlásfüggvény látható. A valószínűségi változó -1 és 1 közötti értékeket vehet fel, a legerősebb sűrűsödés az $x = 0$ környékén látható, $f(0) = 1$, de ez értelemszerűen nem jelenti azt, hogy a 0 értéket 1 valószínűséggel veszi fel a változó. A görbe alatti terület összesen egységnyi, ami egyszerű matematikai módszerekkel két háromszög területeként számítható ki. Az $x = 0$ ponttól balra eső görbe alatti terület a sűrűségfüggvény esetén 0,5, azaz $\mathbf{P}(X \leq 0) = \int_{-\infty}^{0} f(x)\mathrm{d}x = 0{,}5$. Ugyanez a valószínűség olvasható le az eloszlásfüggvényről (jóval egyszerűbben), hiszen $\mathbf{P}(X \leq 0) = F(0) = 0{,}5$.



6.2. ábra. Folytonos valószínűségi változó sűrűség- és eloszlásfüggvénye

Vegyük észre, hogy az eloszlásfüggvény folytonossága miatt a függvényérték bejárja a 0-1 intervallumot, azaz tetszőleges p valószínűséghez meghatározható az az x érték, melyre igaz, hogy $F(x) = \mathbf{P}(X \leq x) = p$, ezt a kifejezést az eloszlásfüggvény inverzének hívjuk és $x = F^{-1}(p)$ módon jelöljük. Az eloszlásfüggvény inverzének segítségével tehát adott p valószínűséghez tudjuk meghatározni azt az x értéket, melyhez pontosan a p eloszlásfüggvény-érték tartozik.

6.2. Momentumok

A 5.2. fejezetben megismerkedtünk a diszkrét valószínűségi változók két legfontosabb momentumával, a várható értékkel és a varianciával. A folytonos esetekben a momentumok jelentése, tartalma azonos, de az alkalmazandó képlet természetesen módosul, az összeadás folytonos megfelelőjére, az integrálásra támaszkodunk.

Egy X folytonos valószínűségi változó várható értékén a

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x \tag{6.4}$$

kifejezést értjük. A várható érték ebben az esetben is a sűrűségfüggvény tömegközéppontjaként értelmezhető, a statisztikai megfelelője a számtani átlag.

Ez azt jelenti, hogy a kísérletet sokszor ismételve a kapott értékek átlaga a várható értékhez fog tartani.

Egy X folytonos valószínűségi változó varianciáján a

$$\mathbf{D}^2(X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \int_{-\infty}^{\infty} (x - \mathbf{E}(X))^2 f(x) \mathrm{d}x \tag{6.5}$$

kifejezést értjük. A variancia gyökét a valószínűségi változó szórásának nevezzük és $\mathbf{D}(X)$ módon jelöljük. A szórás azt mutatja meg, hogy a kísérletet sokszor ismételve a kapott értékek átlagosan milyen messze fognak elhelyezkedni a várható értéktől (és egyben a saját átlaguktól).

Ahogy azt már említettük, a momentumok számítása során a diszkrét összegzés helyett folytonos esetben integrálni szükséges, illetve a súlyfüggvény ($\mathrm{P}(X = x)$) szerepét a sűrűségfüggvény ($f(x)$) veszi át mind (6.4), mind (6.5) esetében.

Eddig nem ejtettünk szót a valószínűségi változók transzformációjáról, de ezen a ponton egy speciális műveletet meg kell említenünk, amely azonban elveiben már ismerős a 2.4. fejezetből.

Legyen X (nem feltétlenül folytonos) valószínűségi változó, ekkor a

$$Z = \frac{X - \mathbf{E}(\mathbf{X})}{\mathbf{D}(X)} \tag{6.6}$$

transzformált valószínűségi változót standardizált valószínűségi változónak nevezzük. Könnyen belátható, hogy

- $\mathbf{E}(Z) = \mathbf{E}\left(\dfrac{X - \mathbf{E}(X)}{\mathbf{D}(X)}\right) = \dfrac{\mathbf{E}(X) - \mathbf{E}(X)}{\mathbf{D}(X)} = 0$
- $\mathbf{D}^2(Z) = \mathbf{D}^2\left(\dfrac{X - \mathbf{E}(X)}{\mathbf{D}(X)}\right) = \dfrac{1}{\mathbf{D}^2(X)}\mathbf{D}^2(X) = 1$

azaz a standardizált valószínűségi változó várható értéke 0, varianciája (és szórása) pedig 1.

6.3. Nevezetes folytonos eloszlások

A diszkrét eloszlásokhoz hasonlóan a folytonos eloszlások közül is vannak olyanok, melyek valamilyen tulajdonságuk miatt gyakrabban alkalmazhatóak. Ezen eloszlások esetén jellemzően a sűrűség- és az eloszlásfüggvény is ismert, a szükséges paraméterek megadásával az eloszlás jól kezelhetővé válik. A sűrűségfüggvény alapján megtörténhet az eloszlás ábrázolása, ami segít elképzelni a valószínűség sűrűsödését, azonban a sűrűségfüggvény csak korlátozottan alkalmas valószínűségek számítására, a görbe

alatti terület gyakran nehezen számítható. Az eloszlásfüggvény ebben segít, gyakorlatilag tetszőleges x értékhez megadva az adott értékhez tartozó baloldali görbe alatti területet.

A nevezetes eloszlások alkalmazása a valószínűségek könnyebb számíthatósága mellett azért is előnyös, mert a momentumok is a paraméterektől függő zárt képlettel rendelkeznek, azaz nem kell a (6.4) és (6.5) általános képleteket és integrálást alkalmazni, hanem egyszerűbb speciálisan az adott eloszlásra jellemző képletek állnak rendelkezésünkre.

Az alábbiakban a legegyszerűbb folytonos eloszlásokat mutatjuk be röviden, a további tanulmányok során ezeket még több folytonos eloszlás fogja követni. Az első két eloszlás jelentőségét az egyszerűségük adja, ennek ellenére fontos, gyakorlatban alkalmazható eloszlások, a normális eloszlás pedig a statisztika számára az egyik legfontosabb eloszlás.

6.3.1. Folytonos egyenletes eloszlás

Az egyenletes eloszlás az egyik legegyszerűbb folytonos eloszlás. Azt az esetet írja le, ha egy intervallumon valamennyi azonos hosszúságú részintervallum ugyanolyan valószínűséggel következik be, vagy más szavakkal, a valószínűség sűrűsödése egyenletes, azaz a sűrűségfüggvény konstans. A folytonos egyenletes eloszlásnak tehát két paramétere van, melyeket jelöljünk a diszkrét esethez hasonlóan a-val és b-vel, ahol a legkisebb, b pedig a legnagyobb lehetséges értéke X-nek.

Az $X \sim \mathcal{U}(a,b)$ jelölést alkalmazva tehát a sűrűség- és eloszlásfüggvény:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{ha } a \le x \le b, \\ 0 & \text{egyébként} \end{cases} \qquad F(x) = \begin{cases} 0 & \text{ha } x < a \\ \frac{x-a}{b-a} & \text{ha } a \le x < b \\ 1 & \text{ha } x \ge b \end{cases} \tag{6.7}$$

A folytonos egyenletes eloszlás momentumai a paraméterek segítségével könnyen megadhatók, azok hasonlítanak a (5.9) egyenletekben látottakhoz. A várható érték és a variancia a

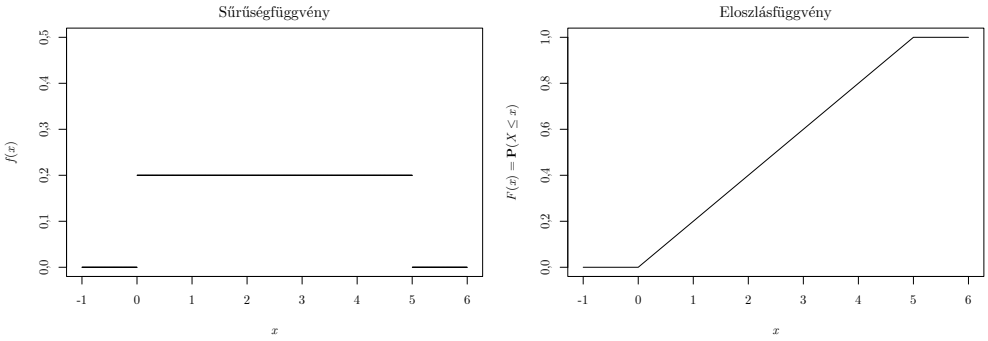
$$\mathbf{E}(X) = \frac{a+b}{2}, \qquad \mathbf{D}^2(X) = \frac{(b-a)^2}{12} \tag{6.8}$$

formulák segítségével számítható. Ha tudjuk, hogy a $0-t$ időintervallumban pontosan 1 telefonhívás érkezik be, akkor joggal feltételezhetjük, hogy a telefonhívás ideje leírható az $X \sim \mathcal{U}(0,t)$ eloszlással. Egyenletes eloszlással közelíthető egy szakaszra eső véletlenszerű pont helyzetének eloszlása, stb.

A fenti elméleti fejtegetéseket az alábbi példával szeretnénk érthetőbbé tenni. Tegyük fel, hogy egy telefonhívás érkezik be a következő 6 percen belül, jelölje X az eltelt időt és tételezzük fel, hogy $X \sim \mathcal{U}(0,6)$. Ekkor

$$f(x) = \begin{cases} \frac{1}{6} & \text{ha } 0 \le x \le 6, \\ 0 & \text{egyébként} \end{cases} \qquad F(x) = \begin{cases} 0 & \text{ha } x < 0 \\ \frac{x}{6} & \text{ha } 0 \le x < 6 \\ 1 & \text{ha } x \ge 6 \end{cases}$$

Mivel X folytonos, ezért tudjuk, hogy $\mathrm{P}(X = x) = 0$ minden x-re. A sűrűségfüggvény a 0-6 intervallumon konstans, könnyen ellenőrizhetjük azt is, hogy $f(x)$ sűrűségfüggvény, hiszen $f(x) \ge 0$ minden x-re, illetve a görbe alatti terület 1, hiszen egy olyan téglalapról van szó, melynek egyik oldala 6, a másik $\frac{1}{6}$ hosszúságú.



6.3. ábra. Egyenletes eloszlású változó sűrűség- és eloszlásfüggvénye

Könnyen ellenőrizhető, hogy a sűrűségfüggvény $-\infty$-től x-ig vett integrálja épp az eloszlásfüggvény, illetve az eloszlásfüggvény deriváltja épp $\dfrac{1}{6}$.

Példaként tekintsük annak az eseménynek a valószínűségét, hogy 3 percet, vagy annál kevesebbet kell várni a hívásra. A $\mathbf{P}(X \le 3)$ valószínűséget két módon is meghatározhatjuk, egyrészt a sűrűségfüggvényen az $x = 3$ értéktől balra eső görbe alatti terület meghatározásával. A görbe alatti terület a $\int_{-\infty}^{3} \frac{1}{6} \mathrm{d}u$ kifejezés segítségével határozható meg, ami jelen esetben felesleges, hiszen egy téglalapról van szó, melynek egyik oldala 3, a másik $\dfrac{1}{6}$ hosszúságú. A keresett valószínűség tehát a sűrűségfüggvény segítségével 0,5. Az eloszlásfüggvény segítségével sokkal könnyebb a keresett valószínűség meghatározása, hiszen csupán be kell helyettesítenünk az $F(x) = \dfrac{x}{6}$ képletbe, amiből közvetlenül adódik, hogy a keresett valószínűség 0,5. Az eloszlásfüggvény ábrájáról ez a pont könnyedén leolvasható.

A fenti, baloldali valószínűség mellet a jobboldali valószínűségek, vagy az intervallumba esés valószínűsége is analóg módon számítható, akár a

sűrűségfüggvény megfelelő integrálásával, vagy az eloszlásfüggvény alkalmazásával a fejezet elején felírt azonosságok alapján.

A momentumok segítségével a várható érték és a variancia könnyedén meghatározható, $\mathbf{E}(X) = \frac{0+6}{2} = 3$, illetve $\mathbf{D}^2(X) = \frac{36}{12} = 3$. Az általános (6.4) és (6.5) formulák alkalmazásával szintén megkaphattuk volna a momentumokat.

Kérdésként merülhet fel, hogy mi az az időtartam, ami esetén annak a valószínűsége, hogy maximum annyit kell várnunk épp 0,8. Most tehát egy valószínűséget ismerünk, méghozzá azt az x értéket keressük, amire teljesül, hogy $\mathbf{P}(X \leq x) = 0{,}8$. Erre a kérdésre az eloszlásfüggvény inverze ad választ. Azt az x-et keressük tehát, amire $F(x) = \frac{x}{6} = 0{,}8$, amiből egyszerűen kapjuk az $x = 4{,}8$ megoldást. Tehát 0,8 valószínűséggel legfeljebb 4,8 percet kell várni a hívásra.

6.3.2. Exponenciális eloszlás

Az exponenciális eloszlást gyakran alkalmazzák véletlenszerű időtartamok leírására. Értelmezési tartománya a nemnegatív valós számok halmaza, egyetlen, pozitív paraméterrel rendelkezik, melyet λ jelöl. Az $X \sim \mathrm{Exp}(\lambda)$ jelölést alkalmazva tehát a sűrűség- és eloszlásfüggvény:

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & \text{ha } x \geq 0 \\ 0 & \text{egyébként} \end{cases} \qquad F(x) = \begin{cases} 1 - \mathrm{e}^{-\lambda x}, & \text{ha } x \geq 0 \\ 0, & \text{egyébként} \end{cases} \tag{6.9}$$

A várható érték és a variancia a

$$\mathbf{E}(X) = \frac{1}{\lambda}, \qquad \mathbf{D}^2(X) = \frac{1}{\lambda^2} \tag{6.10}$$

formulák segítségével számítható, tehát az exponenciális eloszlás esetén a várható érték megegyezik a szórással. Az exponenciális eloszlás egy fontos tulajdonsága, hogy „nincs memóriája”, azaz belátható, hogy $\mathbf{P}(X > x + s | X > s) = \mathbf{P}(X > x)$ minden nemnegatív x és s esetén. Azaz ha például egy meghibásodásig tartó időtartamot vizsgálunk, akkor ha a meghibásodás nem következett be az első x másodperc alatt, akkor annak a valószínűsége, hogy a meghibásodás nem következik be a következő s másodpercben pontosan ugyanannyi, mintha a 0 időponttól kezdve mérnénk az s másodpercet. Ez a tulajdonság nem mindig reális, ezért az exponenciális eloszlás egyéb változatait is gyakran alkalmazzák például meghibásodások vizsgálata esetén. A folytonos eloszlások közül egyedül az exponenciális eloszlás rendelkezik ezzel a tulajdonsággal.

Vizsgáljuk egy call-centerbe érkező hívások hosszát. Ha tudjuk, hogy a hívások átlagos hossza két perc, akkor gyakran alkalmazzuk a $\lambda = 0{,}5$ exponenciális eloszlást (hiszen $\mathbf{E}(X) = \frac{1}{\lambda} = 2$).

Mi a valószínűsége, hogy a következő hívás 1 és 3 perc közötti időtartamot fog felölelni? Tudjuk, hogy

$$\mathbf{P}\left(1 < X < 3\right) = F(3) - F(1) = 1 - \mathrm{e}^{-0,5\cdot 3} - (1 - \mathrm{e}^{-0,5}) = \mathrm{e}^{-0,5} - \mathrm{e}^{-1,5} = 0,3834$$

A valószínűséget a sűrűségfüggvény 1-3 intervallumon vett integrálásával is megkaphattuk volna, ennek belátását az olvasóra bízzuk.

Mi az a híváshossz, amelynél a hívások 95%-a rövidebb?

$$\mathbf{P}\left(X < x\right) = F(x) = 1 - \mathrm{e}^{-0,5\cdot x} = 0,95$$

A fenti egyenletet rendezve azt kapjuk, hogy

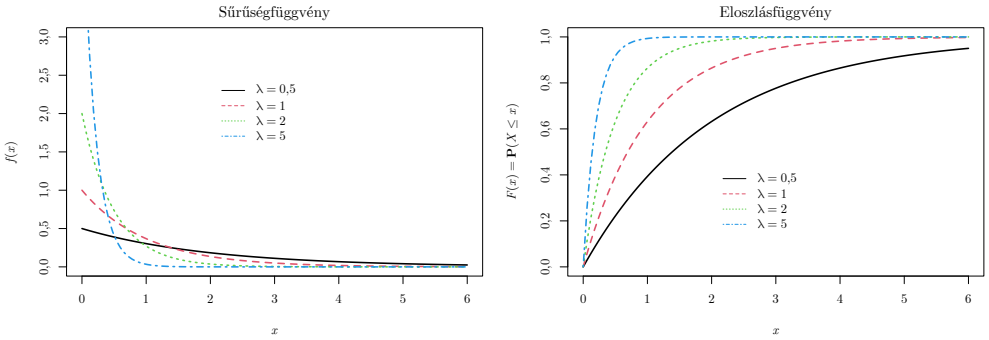
$$\mathrm{e}^{-0,5\cdot x} = 0,05$$

amiből

$$x = -\frac{\ln(0,05)}{0,5} = 5,991$$

percnél rövidebb a hívások 95%-a.

Az exponenciális eloszlás sűrűség- és eloszlásfüggvényét mutatja be a 6.4. ábra különböző λ paraméterek mellett.



6.4. ábra. Exponenciális eloszlású változó sűrűség- és eloszlásfüggvénye

6.3.3. Normális eloszlás

A normális eloszlás az egyik legfontosabb, leggyakrabban alkalmazott eloszlás. Ennek egyik oka, hogy a természetben is sok jelenség jó közelítéssel leírható ezzel az eloszlással, a másik ok, hogy a következtetéses statisztikában központi szerepet tölt be ez az eloszlás. A normális eloszlás két paraméterrel rendelkezik, melyeket μ és σ jelöl (néhol μ és σ^2 paramétereket alkalmaznak), amely jelöléseket már alkalmaztunk a leíró statisztikával foglalkozó 2. fejezetben. A tananyag írásakor

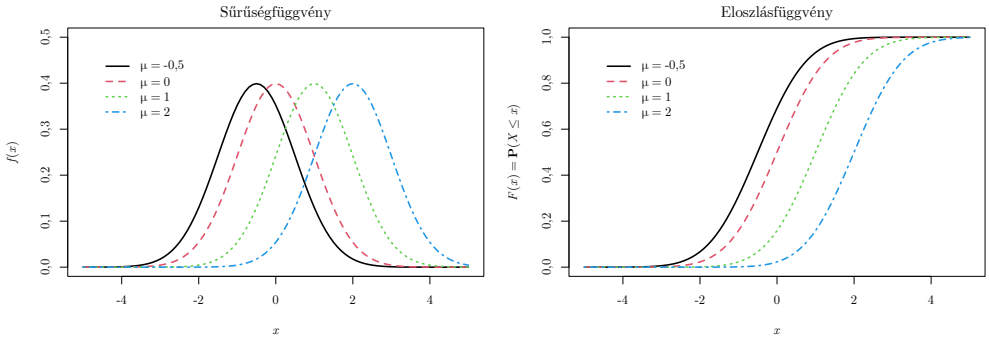
kínosan ügyelünk arra, hogy különböző fogalmakat ne jelöljünk ugyanazzal a betűvel, itt azonban mégis ezt tesszük, de a két paramétert klasszikusan ezekkel a betűkkel szokás jelölni, a későbbiekben kiderül, hogy miért. Legyen $X \sim \mathcal{N}(\mu, \sigma)$, ekkor az eloszlás sűrűségfüggvénye:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{6.11}$$

ahol $x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0$, $\pi \approx 3{,}1416$ a Ludolph-féle szám, $\mathrm{e} \approx 2{,}7182$ pedig az Euler-féle szám. A függvény a valós számok teljes körén értelmezett, így az X valószínűségi változó $-\infty$ és ∞ között elvileg bármilyen értéket felvehet. A μ paraméter tetszőleges, a σ paraméter viszont nemnegatív. Az eddig megismert folytonos valószínűségi változókkal ellentétben az eloszlásfüggvény nem adható meg zárt alakban. A momentumok azonban a lehető legegyszerűbben számíthatóak, ugyanis

$$\mathbf{E}(X) = \mu \qquad \mathbf{D}^2(X) = \sigma^2 \tag{6.12}$$

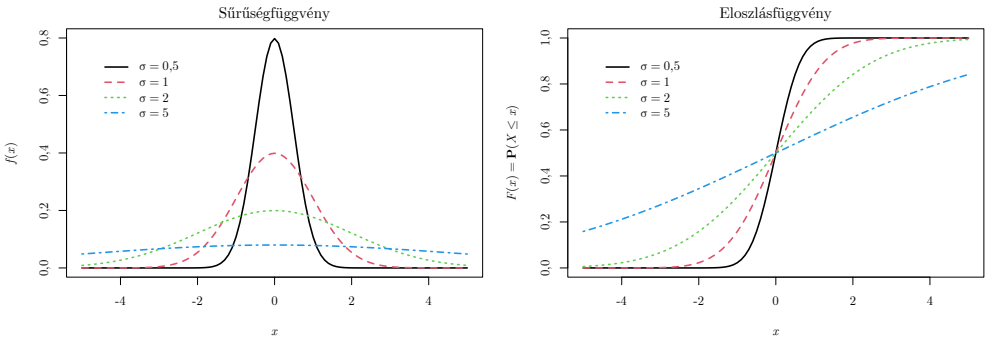
azaz azt tapasztaljuk, hogy a két paraméter értéke épp a várható értéket és a varianciát „állítja be”, egymástól függetlenül. Éppen ez az a tulajdonság, ami miatt a jelölésük megegyezik a sokasági átlag és a sokasági variancia/szórás általunk is alkalmazott jelölésével.



6.5. ábra. Különböző várható értékű, 1 szórású normális eloszlások

Az ábrák alapján látható, hogy a normális eloszlás a μ paraméterre szimmetrikus, illetve a módusza (a sűrűségfüggvény legmagasabb pontja) és a mediánja (a görbe alatti területet felező pont) is itt található. A függvények csupán eltolásban különböznek. Valamennyi esetben az értelmezési tartomány a teljes számegyenes, de a valószínűségek döntő része a $-5, 5$ intervallumba esik ezekben a példákban.

Az azonos μ paraméternek köszönhetően a sűrűségfüggvények azonos szimmetria tengellyel rendelkeznek, illetve valamennyi eloszlásfüggvény átmegy a (0; 0,5) ponton.



6.6. ábra. Különböző szórású, 0 várható értékű normális eloszlások

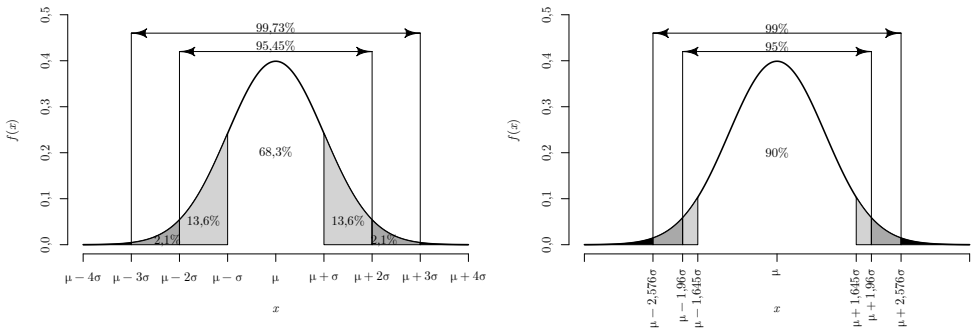
Minél kisebb a σ paraméter értéke, jellemzően annál közelebb esnek az értékek a várható értékhez. A $\sigma = 5$ paraméter mellett az eloszlás nagy része már nem fér rá a diagramra.

Legyen $X \sim \mathcal{N}(100, 15)$, ami jó közelítéssel egy populáció IQ hányadosát leíró valószínűségi változó. Vegyük észre, hogy a feladat az IQ-t folytonos változóként kezeli, annak ellenére, hogy az IQ tesztek eredménye tipikusan egész szám. A modell feltételezése szerint az IQ nagyon sok lehetséges értéket vesz fel, az a mérőeszközünk jellemzője, hogy csak egész értékekre mér. Ebből azt tudjuk, hogy a legtöbb embernek 100 körül van az IQ-ja, illetve az emberek felének ennél alacsonyabb, a másik felének ennél az értéknél magasabb.

A normális eloszlás egyik fontos tulajdonsága, hogy a paraméterek értékétől függetlenül a várható érték adott szórásnyi környezetében minden esetben azonos terület található, ami a valószínűségek meghatározásában a későbbiekben segítségünkre lesz. Ahogy láttuk, az eloszlásfüggvénynek nincs zárt alakja, így ezeket a valószínűségeket numerikus úton lehetséges meghatározni (közelítés, szimuláció, stb.), ez azonban nem képezi tananyagunk részét. A valószínűségeket táblázat, vagy szoftver segítségével fogjuk meghatározni.

A 6.7. ábráról tehát azt olvashatjuk le, hogy – függetlenül a paraméterek konkrét értékétől – a várható érték egy szórásnyi környezetében a valószínűség mintegy 68,3%-a található, két σ esetén a valószínűség 95,45%, három σ esetén pedig már 99,73%. Gyakorlati szempontból nagyobb a jelentősége azoknak az eseteknek, ahol a valószínűség egy kerek, 1-hez közeli érték, a jobb oldali ábrán néhány ilyen eset látható. Ahhoz, hogy a terület 90%-át lefedjük a várható érték körül, 1,645 szórásnyit kell eltávolodnunk μ-től mindkét irányban. 95%-os valószínűség eléréséhez 1,96, 99%-hoz pedig 2,576 szórásnyi távolság szükséges.

A normális eloszlás fenti jellemzői alapján tehát azt tudjuk, hogy a népesség mintegy 68,3%-ának 85 és 115 között van az IQ-ja, mintegy 95,5%-nak pedig 70



6.7. ábra. Görbe alatti területek a várható érték körül

és 130 között. A szimmetria miatt az is meghatározható, hogy a fennmaradó 4,5% fele 70 alatti, fele pedig 130 feletti IQ-val rendelkezik. Amennyiben az emberek középső 90%-ának IQ-jára vagyunk kíváncsiak, úgy $1{,}645 \cdot 15 = 24{,}675$ pontot kell a 100 várható értékből levonnunk, illetve hozzáadnunk.

A normális eloszlások közül kiemelt szerepe van a $\mu = 0$ várható értékű, $\sigma = 1$ szórású normális eloszlásnak, amit standard normális eloszlásnak nevezünk. Amint azt a normális eloszlás tulajdonságainál láttuk, a várható érték adott szórásnyi környezetében minden normális eloszlás esetén azonos valószínűség található. Ez lehetőséget ad arra, hogy egy kitüntetett normális eloszláson keresztül határozzuk meg a keresett valószínűségeket, méghozzá egy transzformáció – a standardizálás – segítségével. Mivel ezzel az eljárással minden normális eloszlás standard normálissá alakítható, ezért elegendő a standard normális eloszlás esetén egyszer meghatároznunk a különböző értékekhez tartozó valószínűségeket, majd ezeket egy táblázatba foglaljuk. Legyen tehát $Z \sim \mathcal{N}(0,1)$, ekkor az eloszlás sűrűségfüggvénye:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(z)^2} \tag{6.13}$$

ahol az általános (6.11) sűrűségfüggvénybe a $\mu = 0$ és $\sigma = 1$ helyettesítéseket végeztük el, illetve – jelezve, hogy standardizált változóról van szó – a változó elnevezésére X helyett Z-t használtunk. A sűrűségfüggvények általános $f(x)$ jelölése helyett gyakran a $\phi(z)$ jelölést alkalmazzuk, ami önmagában arra utal, hogy standard normális eloszlással van dolgunk. A standard normális eloszlás eloszlásfüggvényét ennek megfelelően a $\Phi(z)$ módon jelöljük $F(x)$ analógiájára. A $\Phi(z)$ eloszlásfüggvénynek sincs ugyan zárt alakja, de léteznek táblázatok, melyek z bizonyos (pl. -3-tól 3-ig századonként) értékeire megadják az eloszlásfüggvény értékét, azaz a $\mathbf{P}(Z \leq z)$ valószínűségeket. Néhány táblázat (például a tankönyvhöz tartozó képletgyűjteményben található) nem az eloszlásfüggvény értékét, hanem annak komplementerét tartalmazza, így feltétlenül szükséges a használt táblázat alapos ismerete. A táblázatok tipikusan a pozitív z

értékeket tartalmazzák, az eloszlás szimmetriája miatt azonban a negatív z értékekre is alkalmazhatóak. Belátható, hogy minden z-re

$$\Phi(z) = 1 - \Phi(-z) \tag{6.14}$$

ami az eloszlás szimmetriája alapján grafikusan is könnyen ábrázolható.

Határozzuk meg annak a valószínűségét, hogy egy véletlenszerűen kiválasztott megfigyelés esetén az IQ 90-nél alacsonyabb! Tudjuk, hogy $Z = \dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ és a normális eloszlás tulajdonságai alapján

$$\mathbf{P}(X < 90) = \mathbf{P}\left(Z < \frac{90-100}{15}\right) = \mathbf{P}\left(Z < -\frac{2}{3}\right) = \Phi\left(-\frac{2}{3}\right)$$

A (6.14) alapján tehát

$$\mathbf{P}(X < 90) = 1 - \Phi\left(\frac{2}{3}\right)$$

azaz az eloszlásfüggvény értékeit tartalmazó táblázatból a 0,67 helyen található valószínűség komplementerét keressük. Az eloszlásfüggvény komplementerét tartalmazó táblázatból azonban közvetlenül kiolvasható a keresett valószínűség, ami 0,2514. Az Excel, illetve a statisztikai programok a keresett valószínűséget közvetlenül (az eredeti eloszlásból) is képesek szolgáltatni, nem csupán két tizedes figyelembevételével, de fontosnak tartjuk a táblázat ismeretét is.

Sok esetben nem egy x érték alapján keresünk valószínűséget, hanem fordítva, azt az x értéket keressük, amelyhez adott valószínáség tartozik. Ezt az eloszlásfüggvény inverzének nevezzük. Az eloszlásfüggvény zárt alakjának hiánya miatt alapvetően itt is a standard normális eloszláshoz tartozó táblázatra támaszkodhatunk, épp fordítva gondolkodva. A táblázat belsejében keressük meg a leginkább megfelelő valószínűséget (nem feltétlenül találjuk meg egészen pontosan a keresett értéket), majd ehhez keressük meg z-t. A standardizálás műveletét is épp fordítva végezzük el, amennyiben $Z \sim \mathcal{N}(0,1)$, úgy $\mu + Z\sigma = X \sim \mathcal{N}(\mu, \sigma)$.

Határozzuk meg azt az x értéket, melynél az emberiség csupán 1%-ának magasabb az IQ-ja! Ehhez elsőként azt a z értéket keressük, melyre igaz az ekvivalens állítás, miszerint az értékek 99%-a alacsonyabb nála.

$$\Phi(z) = \mathbf{P}(Z \leq z) = 0{,}99$$

ekkor a

$$z = \Phi^{-1}(0{,}99)$$

megoldást (az eloszlásfüggvény inverzét) keressük. A táblázat alapján a keresett z érték 2,32 és 2,33 között található (0,0102 és 0,0099 valószínűség tartozik a két

z értékhez). A 2,33 értékkel számolva a keresett IQ

$$x = \mu + \sigma z = 100 + 15 \cdot 2{,}33 = 134{,}95$$

Ez az az IQ érték tehát, aminél az emberiség csupán 1%-ának magasabb az intelligenciahányadosa.

6.4. Excel tippek

Hasznos Excel függvények:

- EXP.ELOSZL
- NORM.S.ELOSZLÁS, NORM.S.INVERZ
- NORM.ELOSZLÁS, NORM.INVERZ

Az Excel INVERZ függvényei az adott nevezetes eloszlásfüggvény alapján adják meg azt az x értéket, amelynél az eloszlásfüggvény éppen a megadott valószínűséget adja vissza.

- egyéb eloszlások és kapcsolataik: http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

7. fejezet

Valószínűségi vektorváltozó

A 4. fejezetben események valószínűségének kiszámításával foglalkoztunk, majd az 5. fejezetben diszkrét, a 6. fejezetben folytonos valószínűségi változókkal. Amint azt láttuk, a gyakorlatban a véletlen kísérletek kimeneteleit számokkal jellemezzük. Sok esetben azonban egy kísérletnek nem csupán egy, hanem több jellemzőjére is kíváncsiak vagyunk. Ezeket a kimeneteleket egy vektorba rendezhetjük, ennek megfelelően ebben a többváltozós esetben valószínűségi változó helyett valószínűségi vektorváltozóról beszélhetünk.

Egy dobozból történő húzások során nem csak a kék, hanem a piros és zöld golyók számára is kíváncsiak lehetünk. A holnapi időjárás jellemzésére a középhőmérséklet mellett a csapadék mennyiségét és a napos órák számát is feljegyezhetjük. Általánosságban egyedek több tulajdonságát is vizsgálhatjuk egyszerre. A fejezetben egyszerű példaként egy kis benzinkúton az egy óra alatt eladott kávék és ásványvizek együttes eloszlását fogjuk vizsgálni.

A megfigyelt véletlen értékek száma alapján beszélhetünk a valószínűségi vektorváltozó dimenziójáról. Jelen anyagban csak a kétdimenziós vektorváltozók esetére térünk ki részletesebben, de a legtöbb megismerendő koncepció hasonlóan általánosítható többváltozós esetekre is.

Az egyváltozós esethez hasonlóan megkülönböztetünk:

- diszkrét és
- folytonos

valószínűségi vektorváltozókat.

7.1. Kétváltozós diszkrét eloszlások

A fejezetben a kétváltozós diszkrét eloszlásokhoz kapcsolódó fogalmakat és elemzési eszközöket mutatunk be. A fogalmak egy része az 5. fejezetben megismertek általánosítása, a több valószínűségi változó együttes vizsgálata azonban lehetőséget ad azok kapcsolatának vizsgálatára is, amire a statisztika is nagy mértékben támaszkodik.

Jelölje (X, Y) a kétváltozós, mindkét változójában diszkrét vektorváltozót (többváltozós esetben célszerűbb indexeket alkalmazni a változók jelölésére $(X_1, X_2, \dots, X_k)$). Jelölje továbbá rendre x_i és y_j a lehetséges értékeket.

7.1.1. Súly- és eloszlásfüggvény

A valószínűségi vektorváltozókat az egyváltozós esethez hasonlóan jellemezhetjük a lehetséges értékek, valamint azok valószínűségének felsorolásával. Jelölje az együttes bekövetkezési valószínűségeket

$$\mathbf{P}\left(X = x_i, Y = y_j\right) = p_{ij} \tag{7.1}$$

amivel tulajdonképp a kétváltozós diszkrét eloszlás súlyfüggvényét definiáltuk. A két változó együttes eloszlását (azaz a lehetséges értékeket és valószínűségeiket) érdemes egy táblázat segítségével összefoglalni:

7.1. táblázat: A lehetséges értékek és azok együttes bekövetkezési valószínűségei, valamint a peremvalószínűségek

X, Y	y_1	y_2	...	y_m	$\mathbf{P}(X = x_i)$
x_1	p_{11}	p_{12}	...	p_{1m}	$p_{1.}$
x_2	p_{21}	p_{22}	...	p_{2m}	$p_{2.}$
...	...	...	...	...	...
x_n	p_{n1}	p_{n2}	...	p_{nm}	$p_{n.}$
$\mathbf{P}(Y = y_j)$	$p_{.1}$	$p_{.2}$	...	$p_{.m}$	1

Természetesen a fenti valószínűségekre minden i és j esetén $0 \leq p_{ij} \leq 1$, valamint $\sum_i \sum_j p_{ij} = 1$ feltételeknek teljesülniük kell, hogy valóban kétváltozós diszkrét súlyfüggvényről beszélhessünk. A táblázat peremén – az utolsó sorban és oszlopban – feltüntettük az ún. peremvalószínűségeket, melyek a két változó szerinti súlyfüggvényeket adják meg.

A peremvalószínűségek egyszerűen a megfelelő sorban, illetve oszlopban található együttes bekövetkezési valószínűségek összegei, azaz $p_{i.} = \sum_j p_{ij} = \mathbf{P}(X = x_i)$, illetve $p_{.j} = \sum_i p_{ij} = \mathbf{P}(Y = y_j)$.

Az egyváltozós esethez hasonlóan az eloszlásfüggvény egyszerű felösszegzéssel képezhető, mind a peremeloszlásokra, mind az együttes eloszlásra vonatkozóan. Az X változó szerinti perem eloszlásfüggvény tehát

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{x_i \leq x} \sum_j p_{ij} = \sum_{x_i \leq x} p_{i\cdot} \tag{7.2}$$

módon számítható, azaz az együttes bekövetkezési valószínűségek és a peremvalószínűségek összegzésének segítségével is. Az együttes eloszlásfüggvény az egyváltozós eset általánosítása, azt mutatja meg, hogy a valószínűségi vektorváltozó milyen valószínűséggel teljesíti a $X \leq x$ és $Y \leq y$ feltételeket:

$$F_{XY}(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{ij} \tag{7.3}$$

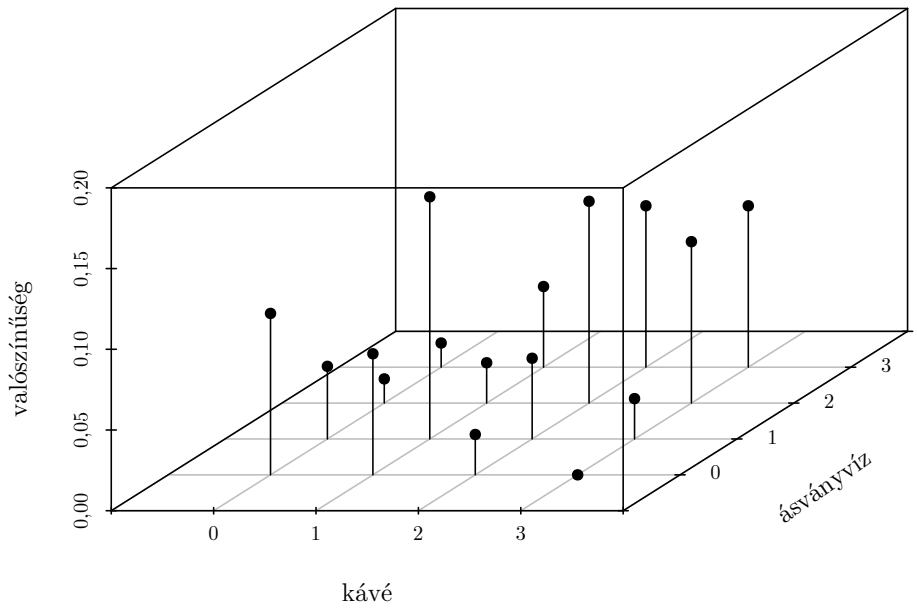
Legyen tehát X az adott órában eladott kávék, míg Y az ásványvizek száma. Tudjuk, hogy az együttes bekövetkezési valószínűségeket az alábbi táblázat írja le.

7.2. táblázat: Az együttes bekövetkezési valószínűségek és peremvalószínűségek

kávé, víz	0	1	2	3	összesen
0	0,1	0,045	0,015	0,015	0,175
1	0,075	0,15	0,025	0,05	0,3
2	0,025	0,05	0,125	0,1	0,3
3	0	0,025	0,1	0,1	0,225
összesen	0,2	0,27	0,265	0,265	1

Az együttes bekövetkezési valószínűségek az egyváltozós esethez hasonlóan ábrázolhatók, az egy tengely helyett azonban már kettőre van szükségünk, a lehetséges értékek egy síkon helyezkednek el, a bekövetkezési valószínűségeket pedig a harmadik dimenzióban ábrázoljuk. Három, vagy több valószínűségi változóra vonatkozó súlyfüggvény ábrázolása már nehezen megoldható, mert legalább négy dimenzióra lenne szükség. A 7.1. ábrán a példára vonatkozó kétváltozós súlyfüggvény látható.

Az ábra alapján megállapíthatjuk, hogy a legnagyobb valószínűségek a főátló mentén láthatók, azaz ahol az értékesített ásványvizek és kávék mennyisége megegyezik. Ez arra utal, hogy a két termék kiegészítő termékként viselkedik a mikroökonómia szóhasználatával élve.



7.1. ábra. Kétváltozós súlyfüggvény

7.1.2. Feltételes eloszlások

Az, hogy két valószínűségi változót vizsgálunk lehetőséget ad az ún. feltételes eloszlások vizsgálatára is. A feltételes eloszlás alatt azt értjük, hogy az egyik változó fix értéke mellett a másik változó eloszlása hogyan alakul. A feltételes eloszlás vizsgálatához a már korábban tárgyalt, eseményekre vonatkozó feltételes valószínűség fogalmához nyúlunk vissza, amit a (4.9) formulában definiáltunk. A kétváltozós diszkrét eloszlások jelölésrendszerét alkalmazva

$$\mathbf{P}(X = x_i \mid Y = y_j) = \frac{\mathbf{P}(X = x_i, Y = y_j)}{\mathbf{P}(Y = y_j)} = \frac{p_{ij}}{p_{.j}} \tag{7.4}$$

amennyiben az Y változó értéke y_j. Az X változó értékét rögzítve a feltételes valószínűségek analóg módon határozhatók meg:

$$\mathbf{P}(Y = y_j \mid X = x_i) = \frac{\mathbf{P}(X = x_i, Y = y_j)}{\mathbf{P}(X = x_i)} = \frac{p_{ij}}{p_{i.}} \tag{7.5}$$

azaz a feltételes valószínűség kiszámítható az együttes bekövetkezés valószínűségének és a megfelelő peremvalószínűségnek a hányadosaként. Adott feltételhez tartozó valamennyi feltételes valószínűség a feltételes eloszlást adja meg. Mivel az egyik valószínűségi változót adott értéken fixáltuk, így a feltételes eloszlás ebben az esetben egy egyváltozós valószínűségi változó, így az 5. fejezetben megismert elemzési eszközök is alkalmazhatók, így például a feltételes eloszlás várható értéke is kiszámítható, ami a későbbiekben is fontos szerepet kap. Rögzített y_j esetén a formula

$$\mathbf{E}(X \mid Y = y_j) = \sum_i x_i \mathbf{P}(X = x_i \mid Y = y_j) = \sum_i x_i \frac{p_{ij}}{p_{.j}} \tag{7.6}$$

míg rögzített x_i esetén

$$\mathbf{E}(Y \mid X = x_i) = \sum_j y_j \mathbf{P}(Y = y_j \mid X = x_i) = \sum_j y_j \frac{p_{ij}}{p_{i.}} \tag{7.7}$$

Tegyük fel, hogy tudjuk, az elmúlt órában $X = 2$ kávét értékesítettek a benzinkúton, ekkor kíváncsiak lehetünk az egyes ásványvíz értékesítések valószínűségére, valamint a feltételes várható értékre is.

$$\mathbf{P}(Y = 0 \mid X = 2) = \frac{0{,}025}{0{,}3} = \frac{1}{12} \quad \mathbf{P}(Y = 1 \mid X = 2) = \frac{0{,}05}{0{,}3} = \frac{1}{6}$$

$$\mathbf{P}(Y = 2 \mid X = 2) = \frac{0{,}125}{0{,}3} = \frac{5}{12} \quad \mathbf{P}(Y = 3 \mid X = 2) = \frac{0{,}1}{0{,}3} = \frac{1}{3}$$

azaz ha tudjuk, hogy két kávét értékesítettek, akkor annak a valószínűsége, hogy nem értékesítettek ásványvizet mindössze $\frac{1}{12}$, míg a 3 ásványvíz valószínűsége $\frac{1}{3}$. Az ezekből számítható feltételes várható érték Y-ra vonatkozóan pedig a

$$\mathbf{E}(Y \mid X = 2) = \sum_j y_j \frac{p_{ij}}{p_{i.}} = 0 \cdot \frac{1}{12} + 1 \cdot \frac{1}{6} + 2 \cdot \frac{5}{12} + 3 \cdot \frac{1}{3} = 2$$

Azaz ebben a példában, amennyiben tudjuk, hogy egy adott órában két kávét értékesítünk, akkor az eladott ásványvizek várható darabszáma is épp kettő, bár természetesen előfordulhat 0, 1, 2 és 3 is eladott mennyiségként, ahogy azt a feltételes valószínűségek mutatják. A várható érték (hasonlóan az egyváltozós esethez) nem feltétlenül olyan érték, amely lehetséges, így például

$$\mathbf{E}(Y \mid X = 3) = 2{,}333$$

Adott mennyiségű ásványvízhez tartozó kávéra vonatkozó valószínűségek és várható értékek analóg módon számíthatók.

A feltételes várható érték mellett a feltételes eloszlás varianciája, vagy bármilyen egyéb momentuma is meghatározható, az egyváltozós esettel analóg módon, ez azonban meghaladja tananyagunk kereteit.

7.1.3. Függetlenség

Két esemény függetlenségét a (4.12) formulával, a 4.3. fejezetben definiáltuk, két valószínűségi változó függetlensége támaszkodik ezekre az ismeretekre. Azt mondjuk, hogy X és Y valószínűségi változó független, ha

$$\mathbf{P}(X = x_i \mid Y = y_j) = \mathbf{P}(X = x_i) \tag{7.8}$$

minden i és j esetén teljesül, azaz bármilyen, Y-ra vonatkozó információ sem változtatja meg az x_i események valószínűségét. A definíció fordítottja is kimondható, azaz függetlenség esetén

$$\mathbf{P}(Y = y_j \mid X = x_i) = \mathbf{P}(Y = y_j) \tag{7.9}$$

Ez pontosan akkor teljesül, ha

$$\mathbf{P}(X = x_i, Y = y_j) = \mathbf{P}(X = x_i)\mathbf{P}(Y = y_j) \tag{7.10}$$

minden i és j esetén teljesül, azaz az együttes bekövetkezési valószínűségek a teljes táblázatban a hozzájuk tartozó peremvalószínűségek szorzataiként előállíthatók.

A példánkban szereplő X és Y valószínűségi változók nem függetlenek, hiszen már korábban láttuk, hogy pl.

$$\mathbf{P}(Y = 0 \mid X = 2) = \frac{1}{12} \neq \mathbf{P}(Y = 0) = 0{,}2$$

Mivel az azonosságnak minden feltételes valószínűségre teljesülni kell függetlenség esetén, ezért már egy ellenpélda esetén is kimondhatjuk, hogy a valószínűségi változók nem függetlenek. Egyezőség esetén azonban tovább kell vizsgálódnunk, hogy minden esetben fennáll-e az.

A függetlenség ellenőrizhető a (7.10) egyenlet alapján is, amihez – és később megismerendő vizsgálatokhoz – érdemes elkészíteni a függetlenség esetén érvényes valószínűségeket a peremvalószínűségek segítségével. A példánk esetén ez három tizedesre kerekítve az alábbi táblázatban látható.

7.3. táblázat: Függetlenség esetén érvényes valószínűségek

kávé, víz	0	1	2	3	összesen
0	0,035	0,047	0,046	0,046	0,175
1	0,060	0,081	0,080	0,080	0,3
2	0,060	0,081	0,080	0,080	0,3
3	0,045	0,061	0,060	0,060	0,225
összesen	0,2	0,27	0,265	0,265	1

7.1.4. Vektorváltozó momentumai

Az egyváltozós eloszlások legfontosabb momentumai (5.2. és 6.2. fejezetek) a várható érték és a variancia. Vektorváltozó esetén a várható érték szerepét skalár helyett egy vektor, míg a variancia szerepét egy mátrix veszi át.

Kétváltozós diszkrét eloszlás esetében a várható érték egy kétdimenziós vektor

$$\mathbf{E}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \sum_i x_i p_{i\cdot} \\ \sum_j y_j p_{\cdot j} \end{bmatrix} \tag{7.11}$$

azaz a várható érték vektor egyszerűen a komponensek várható értékeinek felsorolásából áll, amiből következik, hogy az egyváltozós esetnél megfigyelt tulajdonságok öröklődnek.

Az egyváltozós esetben megismert variancia helyét a $\mathbf{C}$ variancia-kovariancia mátrix veszi át, amely kétváltozós esetben 2×2-es:

$$\mathbf{C} = \begin{bmatrix} \mathbf{D}^2(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathbf{D}^2(Y) \end{bmatrix} \tag{7.12}$$

amiben a két változó varianciáját mérő $\mathbf{D}^2(X)$ és $\mathbf{D}^2(Y)$ mellett a két változó együttmozgását mérő kovariancia is megjelenik, amely szimmetrikus, azaz $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$. A kovariancia az alábbi módon számítható

$$\mathrm{Cov}(X,Y) = \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))) = \mathbf{E}(XY) - \mathbf{E}(X) \cdot \mathbf{E}(Y) \tag{7.13}$$

ami diszkért esetben a

$$\mathrm{Cov}(X,Y) = \sum_i \sum_j x_i y_j p_{ij} - \left(\sum_i x_i p_{i\cdot} \right) \left(\sum_j y_j p_{\cdot j} \right) \tag{7.14}$$

formában írható fel. A szimmetria mellett a kovariancia alábbi fő tulajdonságait említjük meg:

- bármely valószínűségi változó konstanssal vett kovarianciája 0, azaz $\mathrm{Cov}(X,a) = 0$
- lineáris transzformációk esetén $\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X,Y)$
- a variancia tulajdonképp egy valószínűségi változó önmagával vett kovarianciája $\mathrm{Cov}(X,X) = \mathbf{D}^2(X)$

A példánkban a várható érték vektor a két (feltétel nélküli) várható értéket tartalmazza

$$\mathbf{E}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 0 \cdot 0{,}175 + 1 \cdot 0{,}3 + 2 \cdot 0{,}3 + 3 \cdot 0{,}225 \\ 0 \cdot 0{,}2 + 1 \cdot 0{,}27 + 2 \cdot 0{,}265 + 3 \cdot 0{,}265 \end{bmatrix} = \begin{bmatrix} 1{,}575 \\ 1{,}595 \end{bmatrix}$$

A variancia-kovariancia mátrix elemei közül a kovariancia kiszámítását szemléltetjük, elsőként számítsuk ki a $\mathbf{E}(XY)$ várható értéket.

$$\mathbf{C} = \begin{bmatrix} \mathbf{D}^2(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathbf{D}^2(Y) \end{bmatrix} = \begin{bmatrix} 1{,}044375 & 0{,}612875 \\ 0{,}612875 & 1{,}170975 \end{bmatrix}$$

$$\begin{aligned} \mathbf{E}(XY) &= \sum_i \sum_j x_i y_j p_{ij} = \\ &= 0 \cdot 0 \cdot 0{,}1 + \cdots + 1 \cdot 0 \cdot 0{,}075 + 1 \cdot 1 \cdot 0{,}15 + \cdots + 3 \cdot 3 \cdot 0{,}1 = 3{,}125 \end{aligned}$$

amiből a kovariancia egyszerűen adódik:

$$\mathrm{Cov}(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X) \cdot \mathbf{E}(Y) = 3{,}125 - 1{,}595 \cdot 1{,}575 = 0{,}612875$$

A kovariancia tehát pozitív, ami azt jelenti, hogy az egyik változó nagyobb értékei a másik változó nagyobb értékeivel fordulnak elő együtt gyakran, illetve a kis értékek is gyakran járnak együtt. Ezt fogalmaztuk meg már korábban, a súlyfüggvény vizsgálatakor is úgy, hogy a kávé és az ásványvíz ezen a benzinkúton kiegészítő termékek. Helyettesítő termékek esetén a kovariancia negatív értéket venne fel.

Az egyváltozós diszkrét valószínűségi változókkal analóg módon kiszámított varianciák és a kovariancia alapján felírható tehát a variancia-kovariancia mátrix:

$$\mathbf{C} = \begin{bmatrix} \mathbf{D}^2(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathbf{D}^2(Y) \end{bmatrix} = \begin{bmatrix} 1{,}044375 & 0{,}612875 \\ 0{,}612875 & 1{,}170975 \end{bmatrix}$$

7.1.5. Korreláció

A kovariancia tehát két valószínűségi változó együttmozgásának mértékét méri, hátránya azonban, hogy függ a változók nagyságrendjétől. Ahogyan azt láttuk, ha az X változót a-szorosára változtatjuk, a kovariancia is a-szorosára változik, annak ellenére, hogy ez nem jelenti azt, hogy az Y változóval való együttmozgása megváltozott. Ha például X költséget, vagy profitot jelöl, és forint helyett 1000 forintban mérjük, a kovariancia 0,001-szeresére változna. Ezt a tulajdonságot hivatott kiküszöbölni egy nagyon gyakran alkalmazott mérőszám, a korreláció.

Két valószínűségi változó közötti kapcsolat szorosságát és irányát a

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbf{D}^2(X) \cdot \mathbf{D}^2(Y)}} \tag{7.15}$$

ún. lineáris korrelációs együtthatóval mérjük. Legfontosabb tulajdonságai:

- szimmetrikus, azaz $\rho(X,Y) = \rho(Y,X)$,
- $-1 \leq \rho(X,Y) \leq 1$,
- az előjel a kapcsolat irányát jelöli,
- $\rho(X,Y) = 0$ neve korrelálatlanság,
- $|\rho(X,Y)| = 1$ neve függvényszerű lineáris kapcsolat, tökéletes negatív, vagy pozitív lineáris korreláltság.

Általánosságban minél közelebb van a korrelációs együttható abszolút értéke 1-hez, annál erősebb korrelációs kapcsolatról beszélünk a két valószínűségi változó között.

Amint azt láttuk, a kovariancia értéke pozitív a példánkban, előjele igen, nagysága nem értelmezhető önmagában. A kovarianciából számított korrelációs együttható

azonban biztosan -1 és 1 közötti értéket vesz fel.

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbf{D}^2(X)\cdot\mathbf{D}^2(Y)}} == \frac{0{,}612875}{\sqrt{1{,}044375\cdot 1{,}170975}} = 0{,}5542$$

Azaz a korrelációs együttható – a kovarianciával szükségszerűen azonos – pozitív előjelű, azaz pozitív, de nem tökéletes lineáris korrelációt tapasztalunk a két valószínűségi változó között.

Gyakran keveredik össze a 7.1.3. fejezetben tárgyalt függetlenség, valamint a korrelálatlanság fogalma, ezért ezen a helyen néhány szót ejtünk a két fogalom közötti kapcsolatról

- bizonyítható, hogy ha X és Y függetlenek, akkor korrelálatlanok, azaz $\rho(X,Y) = 0$
- a korrelálatlanság azonban nem jelenti egyben azt, hogy a valószínűségi változók függetlenek!

Ilyen értelemben a két valószínűségi változó közötti függetlenség erősebb állítás, mint a korrelálatlanság. A 0 korrelációs együttható csak annyit jelent, hogy lineáris összefüggés nem figyelhető meg a váltózók között, de más, pl. parabolaszerű kapcsolat elképzelhető.

Legyen X és Y együttes eloszlása az alábbi

X, Y	-1	0	1
0	0	0,5	0
1	0,25	0	0,25

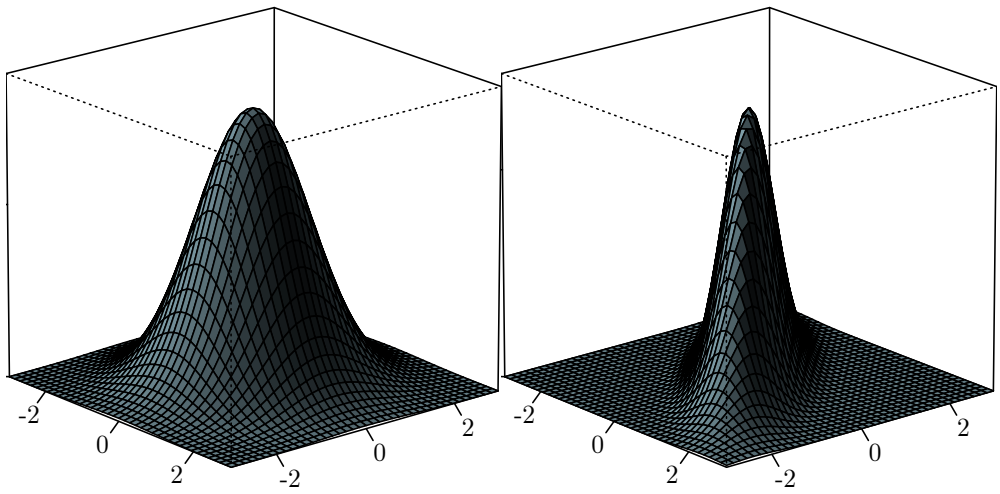
Ebben az esetben könnyen belátható, hogy a kovariancia, és ezzel együtt a korreláció is 0, azaz a két változó korrelálatlan, de a változók nem függetlenek. A számítások elvégzését az olvasóra hagyjuk.

7.2. Kétváltozós folytonos eloszlások

Kétváltozós folytonos eloszlásokról jelen tananyagban részletesen nem lesz szó, azonban a később tanult módszerek miatt elengedhetetlen egy rövid betekintés. Kétváltozós esetben a sűrűségfüggvényt egy görbe helyett egy felület írja le. Ahogy az egyváltozós esetben a görbe alatti terület, úgy kétváltozós esetben a felület alatti térfogat írja le az adott eseményhez tartozó valószínűséget. Az egyváltozós eset analógiájára a sűrűségfüggvény alatti térfogat egységnyi.

Az eloszlásfüggvény szintén egy monoton növekvő felületként képzelhető el, határértéke dél-nyugati irányban (ahol a két változó kis értékei szerepelnek) 0, észak-keleti irányban 1.

Az egyváltozós esethez hasonlóan a többváltozós valószínűségi változók esetén is beszélhetünk nevezetes eloszlásokról, és a többváltozós normális eloszlás a leggyakrabban alkalmazott eloszlások közé tartozik. A 7.2. ábra két különböző, kétváltozós normális eloszlás sűrűségfüggvényét mutatja be. Mindkét eloszlás várható érték vektora $[0, 0]$, a varianciák pedig minden esetben egységnyiek. A különbség a kétváltozós normális eloszlás fontos paramétere, a korrelációs együttható. A baloldali ábrán 0, míg a jobboldali sűrűségfüggvény esetén $-0{,}8$ az értéke.



7.2. ábra. Kétváltozós sűrűségfüggvény

7.3. Excel tippek

- Az 1. fejezetben megismert abszolút és relatív hivatkozások, valamint
- az 5. fejezetben megismert SZORZATÖSSZEG függvény alkalmazása javasolt.

8. fejezet

Mintavétel, mintavételi eloszlás

Az előző fejezetekben a valószínűségszámítás alapjaival, illetve néhány nevezetes diszkrét és folytonos eloszlással ismerkedtünk meg. Ebben a fejezetben a következtetéses statisztika alapjait rakjuk le. A következtetéses statisztika mintából következtet a sokaság egészére, így elsőként néhány fontos mintavételi módszert tekintünk át. Továbbá azt vizsgáljuk ebben a fejezetben, hogy adott sokaságból kiválasztható összes mintát figyelembe véve hogyan alakulnak a minta alapján számított jellemzők (pl. átlag, arány, szórás). Míg ezeket a jellemzőket sokaság esetén összefoglalóan paramétereknek nevezzük, mintabeli megfelelőik összefoglaló neve mintabeli statisztika. A különböző mintákból számított mintabeli statisztikák véletlen változóként viselkednek, eloszlásukat mintavételi eloszlásnak nevezzük.

A mintavétel során az alapsokaságra

- jellemző paraméter közelítő értékére, vagy
- megfogalmazott állítás igazságtartalmára, vagy
- vonatkozó modell összefüggéseire

vagyunk kíváncsiak megfelelő módon kiválasztott minta alapján, mert

- nincs elegendő idő, pénz, egyéb erőforrás, vagy
- nem lehetséges

a teljes sokaság felmérése.

Tekintsük a magyar felsőoktatásban tanulók sokaságát. Egy kutatás során kíváncsiak lehetünk arra, hogy átlagosan hány forintot költenek albérletre, vagy arra, hogy körükben milyen egy adott politikai párt támogatottsága (jellemző paraméter közelítő értéke). Egy másik esetben az állíthatjuk, hogy körükben a férfiak és a nők átlagosan azonos időt töltenek tanulással, vagy a már említett párt támogatói azonos arányban vannak a férfiak és nők körében (megfogalmazott állítás), és ezt egy minta alapján próbáljuk ellenőrizni. Még nehezebb a dolgunk, ha azt szeretnénk körükben vizsgálni, hogy a család jövedelme, vagy a hallgató lakóhelye hogyan befolyásolja az albérletre költött összeget (modell összefüggés). Valamennyi esetre elmondható, hogy valamennyi felsőoktatási hallgató megkérdezése rengeteg időbe és pénzbe kerülne, ezért a fenti vizsgálatokat leginkább egy minta alapján reális elvégezni. Más esetekben a teljes sokaság megismerése egyáltalán nem lehetséges, például töréstesztek esetén.

A minta alapján levont következtetés eredménye mindig bizonytalansággal terhelt, egyrészt azért, mert nem a teljes sokaságot ismerjük meg, másrészt egyéb hibaforrások is jelen vannak: nemválaszolás, rossz adat rögzítése, félreértett kérdés, rossz mértékegység használata, stb. Vegyük észre, hogy ezek a potenciális hibák akkor is jelen vannak, ha a teljes sokaságot próbáljuk vizsgálni. A mintavételi hiba és az egyéb hibák közötti jelentős különbség az, hogy a mintavételi hiba bizonytalansága számszerűsíthető. Ebben a fejezetben azzal kezdünk foglalkozni, hogy ez a bizonytalanság hogyan kvantifikálható.

A minta alapján szerzett információ tehát bizonytalan, azonban az üzleti döntéshozatalhoz elengedhetetlen információkat szerezhetünk segítségével. A tananyag hátralévő részében a sokasági paraméter közelítő értékének meghatározásával foglalkozunk, amit szakszóval becslésnek nevezünk. A megfogalmazott állítás igazságtartalma, a hipotézis vizsgálat, és a modellezés témaköre a Statisztikai modellezés tárgy anyagához tartozik.

8.1. Mintavételi módszerek

A mintavételi módszerek széles tárháza áll a kutatást végzők rendelkezésére, valamennyi módszer bemutatása meghaladja a tananyagunk kereteit. A mintavételi módszereket két nagy csoportba soroljuk, majd a minta reprezentativitásáról ejtünk néhány szót.

8.1.1. Véletlen mintavételi módszerek

A véletlen mintavételi módszerek közös jellemzője, hogy a sokaság elemei előre meghatározható (nem feltétlenül azonos) valószínűséggel kerülnek a mintába. A véletlen mintavételi módszereknek ez a tulajdonsága lehetővé teszi, hogy a

mintavételhez kapcsolódó bizonytalanságot számszerűsítsük, így elsősorban az ilyen mintavételek eredményeiből levonható következtetést tárgyaljuk.

Véletlen mintavételi módszerek közül az alábbiakat említjük meg:

- független azonos eloszlású (FAE): a FAE mintavétel feltétele, hogy rendelkezésünkre álljon egy lista a sokaság elemeiről, majd pl. véletlenszámok generálásával kiválasztjuk a megvizsgálandó sokasági elem sorszámát. A mintavételt visszatevéssel végezzük, azaz egy sokasági elem akár többször is a mintába kerülhet. A FAE mintavétel nem elsősorban gyakorlati szempontból fontos, hanem matematikai kezelhetősége miatt. A gyakorlatban gyakran a végtelen nagy sokaságból kiválasztott mintákat is FAE mintaként kezeljük, természetesen ilyenkor a véletlen kiválasztást valami más módszerrel kell biztosítanunk.
- egyszerű véletlen: az egyszerű véletlen mintavétel nagyon hasonló a FAE mintához, azonban a kiválasztás visszatevés nélkül történik. Ennek a módszernek nagyobb a gyakorlati jelentősége, mint a FAE mintavételnek.
- rétegzett: amennyiben rétegzett mintavételt szeretnénk végezni, a sokasági egyedekről más – lehetőleg a vizsgálandó ismérvvel összefüggő – úgynevezett rétegképző ismérveket is ismernünk kell, amiket a kiválasztás során figyelembe is veszünk, de maga a kiválasztás véletlenül történik. A rétegképző ismérvek segíthetnek abban, hogy a mintánk összetétele emlékeztessen a sokaság összetételére, illetve abban is, hogy bizonyos esetekben hatékonyabban[1] tudjuk a becslést elvégezni. A túl sok változó szempontjából történő rétegzésnek gátat szab, hogy a rétegek száma az ismérvváltozatok számának szorzata szerint alakul. A rétegzés oka lehet egyszerűen az is, hogy pl. nem csak országos, de régiós bontású adatokra is szükség van, és rétegzéssel biztosítjuk a megfelelő mintaelemszámot valamennyi régió (réteg) esetén.
- csoportos és többlépcsős: gyakran egy teljes sokaságról nem rendelkezünk listával, azonban a sokaság csoportokra bontható és a csoportokról rendeklezünk információval, más esetekben pedig olcsóbb, illetve gyorsabb a csoportos kiválasztás. Csoportos mintavétel esetén a csoportok közül választunk ki néhányat véletlenszerűen, majd a kiválasztott csoportban minden egyedet felmérünk. Ez a mintavételi módszer azonos mintaelemszám mellett kevésbé hatékony, azonban ahogy már említettük, sok esetben egyszerűbben kivitelezhető. Többlépcsős mintavételről akkor beszélünk, ha több, egymásba ágyazott csoportba sorolhatók a sokasági egyedeink.

Tegyük fel, hogy az általános iskolások matematikai képességeit szeretnénk felmérni. A Minisztérium rendelkezik egy listával, amiről véletlenül kiválaszthatjuk a tesztet kitöltő iskolásokat. Ebben a példában a visszatevéses mintavétel nem életszerű, vegyük azonban észre, hogy a visszatevés nélküli mintavétel sem

[1] Ez azt jelenti, hogy azonos mintaelemszám mellett szűkebb konfidencia intervallumot tudunk képezni az adott paraméterre, mint pl. a FAE mintavétel segítségével. Ez abban az esetben lehetséges, ha a rétegképző ismérv kapcsolatban áll a becsülni kívánt változóval.

változtatja meg túlságosan a sokaságot, feltéve, hogy a mintavételi arány nem túlságosan nagy. Rétegzett mintavétel esetén például nem és régió szerint rétegzést alkalmazva biztosítható, hogy valamennyi rétegre megfelelő mintánk, így adataink fognak rendelkezésre állni. Csoportos mintavételt például az iskolák közüli véletlen választás, majd az adott iskola teljes felmérése jelentene, míg a többlépcsős mintavétel esetén pl. az iskolákon belül osztályokat is választanánk. Ez utóbbi megoldások természetesen sokkal olcsóbbak és egyszerűbbek, statisztikai szempontból azonban kevésbé hatékonyak, így nagyobb mintára lehet szükség. A mintavételi módszer kiválasztása komoly gyakorlati feladat a statisztikai munkában.

A tankönyvben a független azonos eloszlású mintát eredményező és az egyszerű véletlen mintavétellel foglalkozunk részletesebben, mert ezek a leginkább elterjedtek, illetve matematikailag a legegyszerűbben kivitelezhetők. Ne feledjük ugyanakkor, hogy a komplexebb mintavételi módok esetén általában jobb eredményeket kapunk.

8.1.2. Nem véletlen mintavételi módszerek

A nem véletlen mintavételi módszerek esetén a mintába kerülés valószínűsége nem határozható meg, így a mintavételből fakadó hiba számszerűsítése sem oldható meg, ezért a következő fejezetekben bemutatott módszerek sem alkalmazhatók. Ez sajnálatos módon azt jelenti, hogy a kapott eredmények megbízhatóságát nem tudjuk megítélni, így komoly kutatások nem épülnek ilyen technikákra.

Nem véletlen mintavételi módszerek közül az alábbiakat említjük meg:

- kvóta szerinti
- koncentrált
- önkényes
- hólabda

A fogyasztói árak változását a Központi Statisztikai Hivatal méri hónapról hónapra. Az egyes termékkategóriák (pl. pékáruk) árát olyan termékek alapján mérik, melyek a legfontosabbak (pl. egy kilogrammos fehér kenyér), azaz a kiválasztás nem véletlenszerű. Ennek megfelelően a számított index mintavételi bizonytalanságát nem is számszerűsíti a Hivatal. A hólabda módszer jellemzően az online felmérésekhez kötődik, melyben a résztvevőket arra kérik, hogy ismerőseiket, barátaikat is kérjék meg a kérdőív kitöltésére, ilyen módon felhízlalva a mintát.

Az adatok gyűjtésének technikai módszerei is igen eltérőek lehetnek. A személyes megkeresés sok szempontból nagyon hatékony (jobb válaszadási hajlandóság, a kérdés magyarázata szükség szerint, stb), azonban kifejezetten költséges módszer.

A telefonon, vagy levélen keresztül történő mintavételezés olcsóbb ugyan, azonban jellemzően magasabb a nemválaszolási arány, illetve lehetnek olyan rétegek (a társadalom alsó és felső peremén is), akiket nem lehetséges ilyen módokon elérni. Az informatikai eszközök térnyerésével az elektronikus módszerek is egyre könnyebben elérhetők, de ne felejtsük el, hogy ezekkel az eszközökkel sokkal inkább a nem véletlen mintavételi módszerek felé terelhetjük a kutatásunkat. A kérdőív online megosztásánál nagyon gyakori torzítás az önkiválasztás, azaz maga a kitöltő dönti el, hogy kitölti-e a kérdőívet. Könnyen belátható, hogy az adott témában érdeklődő egyén nagyobb valószínűséggel tölti ki azt, így a levonható összkép nem lesz valós.

8.1.3. Reprezentativitás

Az 1. fejezetben már szóba került a szelekciós torzítás fogalma, mely alatt azt értjük, hogy a sokaság bizonyos részei nem – vagy nem megfelelő valószínűséggel – kerülnek a mintába, vagy fordítva, olyanok is a mintánkba kerülnek, akik a vizsgálni kívánt sokaságban nincsenek benne. Reprezentativitás alatt általánosabb értelemben azt értjük, hogy a mintavételi módszerünk biztosítja, hogy a minta jól leírja a sokaságot egy adott vizsgálat céljainak szempontjából. Kiemelendő tehát, hogy önmagában nem beszélhetünk reprezentativitásról, az mindig valamilyen változó, vagy változók szempontjából értelmezhető. A minta mérete és reprezentativitása nem keverendő össze.

8.2. Mintavételi eloszlások

Az üzleti életben és más, statisztikai módszereket alkalmazó területeken is gyakran előfordul, hogy nem ismerjük a teljes sokaságot, így minta alapján kell döntéseket hoznunk. A következtetéses statisztika egyik nagy területe a becslés, melynek során egy sokasági paraméter közelítő értékét kívánjuk meghatározni egy mintabeli statisztika segítségével. A statisztikai becslés elvégzéséhez azonban elsőként azt kell megértenünk, hogy egy adott sokaságból kiválasztható összes minta hogy viselkedik. Becslést bármelyik, a 2. és 3. fejezetekben megismert sokasági paraméterre vonatkozóan készíthetnénk, de jelen tananyagunkban elsősorban két gyakran vizsgált paraméterre, a sokasági átlagra és a sokasági arányra koncentrálunk. A gyakorlatban tehát egyetlen minta átlaga alapján fogunk következtetéseket levonni a sokasági átlagra vonatkozóan, illetve egyetlen mintában tapasztalt arány alapján következtetünk a sokasági arányra. Ahhoz, hogy ezt el tudjuk végezni, először végiggondoljuk, hogy adott sokaságból milyen lehetséges mintaátlagú minták, illetve milyen mintabeli aránnyal rendelkező minták keletkezhetnek. Az összes lehetséges minta alapján kiszámított átlagokat (melyek különbözők lesznek) az átlag mintavételi eloszlásának, analóg módon az arányokat az arány mintavételi eloszlásának nevezzük. Természetesen az összes többi paraméter (medián, szórás, variancia, kvartilisek, stb.) is rendelkezik mintavételi eloszlással, ezekről azonban részletesen nem beszélünk.

Mivel a következtetéses statisztika területére lépünk, át kell ismételnünk a már alkalmazott jelöléseket, illetve újakat is be kell vezetnünk:

- $N, n, \frac{n}{N}$: rendre alapsokaság elemszáma, minta elemszáma és kiválasztási arány,
- μ, σ, π: rendre alapsokasági átlag, szórás és arány,
- $\overline{X}, S, P$: rendre a mintabeli átlag, szórás és arány valószínűségi változója,
- $\overline{x}, s, p$: rendre a mintabeli átlag, szórás és arány egy adott mintából számított realizációja, amit a gyakorlatban megfigyelhetünk.

Fontos továbbá, hogy valamennyi mintaelem valószínűségi változó, azaz őket általánosan X_i-vel fogjuk jelölni.[2] A mintalemek tehát valószínűségi változók, mivel – a mintavétel előtt – értékük ismeretlen, a mintaelemek realizációját x_i-vel fogjuk jelölni ($i = 1, 2, \ldots, n$). Az $X_i = x_i$ esemény tehát azt jelenti, hogy az i. mintaelem épp x_i értéket vesz fel, hasonlóan a diszkrét és folytonos valószínűségi változóknál alkalmazott jelöléshez. Az általános mintaelem eloszlása megegyezik a sokaság eloszlásával, így $\mathbf{E}(X_i) = \mu$, valamint $\mathbf{D}^2(X_i) = \sigma^2$

8.2.1. Az átlag mintavételi eloszlása

Az átlag mintavételi eloszlásának vizsgálatát egy egyszerű példán keresztül vezetjük be, levonunk néhány következtetést, majd innen haladunk az összetettebb esetek felé.

Legyenek egy $N = 4$ elemű sokaság, elemei 33, 36, 49 és 62 (életkorok évben). Könnyen kiszámíthatjuk, hogy ekkor az átlagéletkor $\mu = 45$ év és $\sigma = 11{,}51$ év a sokasági szórás. Vizsgáljuk meg az összes $n = 2$ elemű visszatevéssel húzott mintát. Mivel csupán 16 különböző lehetséges minta van, ezeket fel tudjuk sorolni, majd valamennyi esetben meg tudjuk határozni a mintaátlag adott mintára érvényes realizációját. Ezeket a 8.1. táblázat tartalmazza.

8.1. táblázat: A kételemű sokaság összes lehetséges visszatevéses mintája és mintaátlaga

#	x_1	x_2	$\overline{x}$	#	x_1	x_2	$\overline{x}$
1	33	33	33,0	9	49	33	41,0
2	33	36	34,5	10	49	36	42,5
3	33	49	41,0	11	49	49	49,0
4	33	62	47,5	12	49	62	55,5
5	36	33	34,5	13	62	33	47,5
6	36	36	36,0	14	62	36	49,0
7	36	49	42,5	15	62	49	55,5
8	36	62	49,0	16	62	62	62,0

[2] A jelölés nem keverendő össze az i. sokasági értékkel, sajnos a rengeteg fogalom mellett nehéz elkerülni az azonos jelölést. Törekszünk rá, hogy mindig egyértelmű maradjon a jelölésrendszer.

A fenti példára vonatkozóan a következő megállapításokat tehetjük:

- egyik mintaátlag sem „találja el” a sokasági átlagot, de jellemzően közel esnek a $\mu = 45$ értékhez
- a mintaátlagok átlaga (jelöljük $\mathbf{E}(\overline{X}) = \mu_{\overline{X}}$ módon) épp a sokasági átlaggal egyezik meg, hiszen (a számításnál azt is felhasználtuk, hogy a különböző minták bekövetkezési valószínűsége egyenlő, így nincs szükség súlyozásra)

$$\mu_{\overline{X}} = \frac{33 + 34{,}5 + \cdots + 62}{16} = 45$$

- a mintaátlagok szórása (jelöljük $\mathbf{D}(\overline{X}) = \sigma_{\overline{X}}$ módon) a sokasági szórásnál kisebb, hiszen

$$\sigma_{\overline{X}} = \sqrt{\frac{(33-45)^2 + (34{,}5-45)^2 + \cdots + (62-45)^2}{16}} = 8{,}14$$

A fenti példában azt tapasztaltuk, hogy az összes visszatevéssel vett mintaátlag várható értéke pontosan megegyezik a sokasági átlaggal. Az átlag ezen könnyen belátható tulajdonságát torzítatlanságnak nevezzük, amely természetesen nem csak a fenti példában állja meg a helyét. Tudjuk tehát, hogy ugyan valószínűleg az aktuálisan kiválasztott mintánk átlaga nem egyezik meg pontosan a keresett μ sokasági paraméterrel, de az összes kiválasztható minta esetén az egyezőség várható értékben teljesül, azaz

$$\mu_{\overline{X}} = \mu \tag{8.1}$$

A következő felismerésünk talán még fontosabb, hiszen pontosan arra vagyunk kíváncsiak, hogy az egyes lehetséges mintaátlagok milyen távol esnek a valós sokasági átlagtól. A mintaátlagok szórása pedig épp ezt mutatja meg, hiszen a szórás definíciója alapján az egyes egyedek átlagtól vett átlagos távolságát mutatja meg. Jelen esetben az egyedek az egyes lehetséges mintaátlagok, a mintaátlagok átlagáról pedig tudjuk, hogy az épp a sokasági átlag (lásd torzítatlanság). A mintaátlagok szórását az átlag standard hibájának nevezzük, ami épp a mintavétel átlagos hibáját mutatja meg.

A standard hiba n elemű, független azonos eloszlású mintavétel esetén, kihasználva, hogy az egyes mintaelemek varianciája azonos

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{8.2}$$

azaz azt látjuk, hogy a standard hiba (a mintavétel által okozott átlagos hiba, az abból eredő átlagos hiba, hogy nem figyeljük meg a teljes sokaságot) egyrészt függ a sokaság heterogenitásától, másrészt a minta méretétől. Minél nagyobb a sokasági szórás, annál heterogénebbek a lehetséges mintaátlagok is, illetve ami talán még fontosabb:

a mintaelemszám növelésével az adott sokaságból levonható következtetéseink egyre pontosabbak lesznek.

Az előző példa adatai alapján ellenőrizhető, hogy az összefüggés valóban teljesül-e

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{11{,}51}{\sqrt{2}} = 8{,}14$$

azaz ugyanazt az eredményt kaptuk pusztán a sokasági szórás és a mintaelemszám felhasználásával, mint az összes mintaátlag szórásának kiszámításával. A standard hiba tehát azt jelenti ebben az esetben, hogy átlagosan 8,14 évvel térnek el a mintaátlagok a saját átlaguktól, azaz a sokasági átlagtól. Átlagosan 8,14 évet tévedünk, ha egy kételemű minta alapján becsüljük meg a sokasági életkort.

Természetesen nagyobb N és n esetén az összes mintaátlag meghatározása már nem járható út, de a (8.1) és (8.2) összefüggések természetesen továbbra is érvényesek. Ezek az azonosságok megadják a mintavételi eloszlás várható értékét és szórását, azonban arra nem adnak választ, hogy milyen az eloszlás típusa.

Amennyiben a sokaság normális eloszlású, belátható, hogy a mintavételi eloszlás is normális eloszlást követ és a (8.1), (8.2) összefüggések miatt a különböző n-ek esetére érvényes mintavételi eloszlások várható értéke μ, szórása pedig egyre kisebb, hiszen σ-t egyre nagyobb $\sqrt{n}$ nevezővel kell osztanunk. Ez azt jelenti, hogy egyre nagyobb minták esetén a mintaátlagok egyre inkább a sokasági átlag körül tömörülnek. Mivel a mintavételi eloszlás normális, ezért olyan kérdésekre is választ tudunk adni, hogy mi a valószínűsége, hogy egy adott sokaságból n elemű mintát véve a mintaátlag adott intervallumba esik.

Legyen $X \sim \mathcal{N}(100, 15)$, ami az IQ eloszlását jól leírja a sokaságban. Számítsuk ki az alábbi valószínűségeket:

- egy véletlenül kiválasztott egyén IQ-ja 90 és 110 között van

$$\mathbf{P}(90 < X < 110) = 0{,}495$$

- öt véletlenül kiválasztott egyén átlagos IQ-ja 90 és 110 között van

$$\mathbf{P}(90 < \overline{X} < 110) = 0{,}864$$

- tizenöt véletlenül kiválasztott egyén átlagos IQ-ja 90 és 110 között van

$$\mathbf{P}(90 < \overline{X} < 110) = 0{,}990$$

Azaz az emberek közel fele 90 és 110 közötti IQ-val rendelkezik, amit korábbi ismereteink alapján is ki tudunk számítani. A normalitás és a (8.1) és (8.2) összefüggések alapján a másik két valószínűség is kiszámítható, mindössze a

szórást kell a megfelelő standard hibára módosítani a számítás során. Azt tapasztaljuk, hogy a lehetséges minták egyre nagyobb arányban vannak a sokasági várható értékhez ($\mu = 100$) közel.

A legtöbb gyakorlati esetben azonban a sokaság ismeretlen eloszlású, illetve a sokaság normalitását nem is tételezhetjük fel. Ebben az esetben a mintavételi eloszlás alakjának meghatározása összetett feladat. A Centrális Határeloszlás-Tétel (CHT), vagy központi határeloszlás-tétel azonban kimondja, hogy elégségesen nagy számú független mintavétel esetén a mintavételi eloszlás közelítőleg normálissá válik.[3] Természetesen az elégségesen nagy minta méretének meghatározása sem egyszerű, az függ a sokaság alakjától, de a különböző tankönyvek $n = 30$ és $n = 100$ közötti általános határértékről beszélnek. Természetesen léteznek olyan extrém eloszlások, melyek esetén akár a több ezer elemű minták mintavételi eloszlása sem közelíthető jól a normális eloszlással, ezek tárgyalása azonban meghaladja tananyagunk kereteit.

A 8.1. ábrán öt különböző sokasági eloszlásból vett véletlen minták mintaátlagainak eloszlását tüntettük fel. A sokaságok sűrűségfüggvényei a felső sorban láthatók, a szaggatott vonalak az eloszlások várható értékét mutatják. A következő sorokban egyre nagyobb minta alapján rajzoltuk meg a mintavételi eloszlást, valamint a mintavételi eloszlás várható értékét.[4]

A 8.1. ábra alapján levonható következtetések

- a mintavételi eloszlások várható értéke minden esetben megegyezik a sokasági átlaggal
- a mintavételi eloszlások szórása a minta nagyságának növekedésével csökken
- a mintavételi eloszlások a minta nagyságának növekedésével egyre inkább hasonlítanak a normális eloszlásra

A Centrális határeloszlás-tétel szerint tehát ha $X_1, X_2, \ldots, X_n$ független, azonos eloszlású véletlen változók μ várható értékkel és σ^2 varianciával, akkor

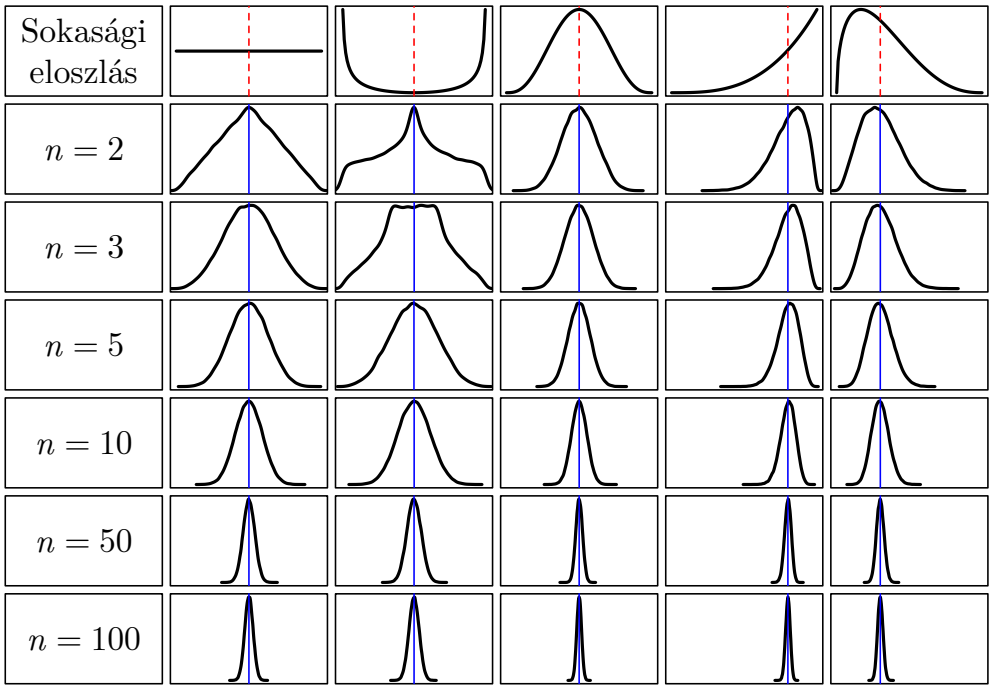
$$\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = \overline{X} \sim \mathcal{N}\left(\mu_{\overline{X}}, \sigma_{\overline{X}}\right), \tag{8.3}$$

ahol $\overline{X}$ a mintaátlag valószínűségi változója, melynek várható értéke $\mu_{\overline{X}} = \mu$, szórása pedig $\sigma_{\overline{X}} = \sigma/\sqrt{n}$.

Ahogy azt már a 6.3.3. fejezetben is láttuk, gyakran a standard normális eloszlás használata a kézenfekvőbb, a következtetéses statisztika sok területe is ezt használja. A fenti állítás standardizálás után úgy is megfogalmazható, hogy elegendően nagy minta esetén

[3] Feltétel továbbá, hogy a sokaság varianciája véges, de ez a gyakorlati eseteinkben minden esetben teljesül.

[4] A pontossághoz hozzátartozik, hogy mivel az alapsokaságok végtelen sok értéket tartalmaznak, ezért a lehetséges minták száma is végtelen. A végtelen sok minta kiválasztására nem vállalkoztunk, ábránként 100000 mintavétel történt.



8.1. ábra. A CHT működése öt különböző sokaság esetén

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1) \tag{8.4}$$

ami a következtetéses statisztika egyik legfontosabb összefüggése, segítségével valószínűségi állításokat tehetünk a mintaátlaggal kapcsolatosan.

Amennyiben a mintavétel nem függetlenül történik, azaz egyszerű véletlen mintavételről beszélünk (véges alapsokaságból, visszatevés nélküli minta), akkor bizonyítható, hogy a standard hiba megváltozik

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{8.5}$$

ahol a második tényező neve véges szorzó, utalva arra, hogy véges alapsokaságból választottuk a mintát. Ezzel kapcsolatban az alábbi megállapításokat tehetjük:

- A véges szorzó csökkenti a mintaátlag bizonytalanságát, hiszen értéke 1 alatti.
- $n \to N$ esetben $\sigma_{\overline{X}} \to 0$, azaz ha a minta mérete megközelíti a sokaság méretét, akkor a bizonytalanság 0-hoz tart.
- $N \to \infty$ esetben a véges szorzó hatása jelentéktelen, végtelen nagy (a gyakorlatban nagy) sokaságok esetén használatától eltekinthetünk.
- Gyakran közelítjük az egyszerűbb $\sqrt{1 - \frac{n}{N}}$ formulával, ami némileg beszédesebb: minél nagyobb a mintavételi arány, annál inkább csökkenthető a standard hiba. Nagyon kicsi, néhány százalékos mintavételi arány esetén a tényező értéke közel egy, így érdemben nem befolyásolja a véges szorzó a standard hibát.
- A binomiális és hipergeometriai eloszlások varianciái épp ebben a véges szorzóban különböznek, lásd (5.12) és (5.14).

Legyen a vizsgált sokaságunk mérete $N = 10\,000$. Amennyiben $n = 100$ elemű a mintánk, azaz a n/N kiválasztási arány 1%-os, a véges szorzó értéke

$$\sqrt{\frac{N-n}{N-1}} = 0{,}99504$$

azaz a hibahatár mindössze fél százalékkal csökken, a közelítő képletet alkalmazva pedig 0,99499 értéket kapunk. A gyakorlatban ezek az értékek elhanyagolhatóan kicsik. Abban az esetben azonban, ha a sokaság mérete mindössze $N = 500$, a véges szorzó értéke 0,895 körül alakul.

8.2.2. Az arány mintavételi eloszlása

Sok esetben nem a sokasági átlagra, hanem a korábban[5] π-vel jelölt, adott tulajdonsággal rendelkező egyedek sokasági arányára szeretnénk következtetéseket levonni. A sokasági arányra vonatkozó következtetésekben a mintabeli arány lesz segítségünkre, így a mintabeli arány mintavételi eloszlását is meg kell ismernünk. Jelölje P a mintabeli arány valószínűségi változóját, p pedig a mintabeli arány egy realizációját. Belátható, hogy független mintavételezés esetén a mintabeli arány is torzítatlan és normális eloszlást követ, méghozzá

$$P \sim \mathcal{N}(\pi, \sigma_P) \tag{8.6}$$

paraméterekkel, ahol $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$ az arány standard hibája. A normális eloszlással történő közelítés abban az esetben alkalmazható, ha a mintaelemszám elegendően nagy. A legtöbb tankönyv feltételként azt követeli meg, hogy $n\pi > 5$ és $n(1-\pi) > 5$ is teljesüljenek, néhol a szigorúbb $n\pi > 10$ és $n(1-\pi) > 10$, vagy az $n\pi(1-\pi) > 5$ feltételeket adják meg. Valamennyi feltétel tartalma azonos: amennyiben π közel 0,5, elegendő viszonylag kis, 10-20 elemű minta a normalitáshoz, míg a 0,5 értéktől távolodva egyre nagyobb mintára van ehhez szükség. Más megközelítésben a normális közelítés megfelelő, ha a keresett arány se nem túl kicsi, se nem túl nagy a mintaelemszámhoz képest.

Mindez azt jelenti, hogy egy π sokasági aránnyal jellemezhető sokaságból minden n elemű mintát kiválasztva a p mintabeli arányok átlaga pontosan π, valamint a mintaelemszám növekedésével a mintabeli arányok egyre inkább megközelítik a sokasági arányt.

Abban az esetben, ha a mintát visszatevés nélkül vesszük és a sokaság viszonylag kicsi, az átlag standard hibája esetén megismert véges szorzó alkalmazható:

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}} \tag{8.7}$$

Tegyük fel, hogy egy párt támogatottsága a választók teljes körében $\pi = 0{,}2$, azaz 20%. Amennyiben $n = 100$ elemű mintát választunk, mit mondhatunk el a lehetséges mintabeli arányokról? Mi a valószínűsége, hogy az egyetlen kiválasztott mintában a mintabeli arány 25% feletti lesz? Hogyan változik a tudásunk, ha tudjuk, hogy $N = 500$ és a mintát visszatevés nélkül választjuk?

Mivel $n\pi = 20$ és $n(1-\pi) = 80$, ezért a mintavételi eloszlás normalitása feltételezhető, azaz (8.6) alapján tudjuk, hogy a lehetséges mintabeli arányok $P \sim \mathcal{N}(0{,}2, 0{,}04)$ normális eloszlást követnek. A 6.3.3. fejezetben tanultak alapján pedig

$$\mathbf{P}(P > 0{,}25) = 0{,}1056$$

[5] A binomiális eloszlás tárgyalása során.

annak a valószínűsége, hogy egy véletlenül választott $n = 100$ elemű mintában a mintabeli arány 25% feletti, feltéve hogy $\pi = 0{,}2$

Amennyiben véges sokaságból visszatevés nélkül választjuk a mintát, úgy a standard hiba a véges szorzóval módosul, azaz $P \sim \mathcal{N}(0{,}2, 0{,}0358)$. Az azonos módon számított valószínűség $\mathbf{P}(P > 0{,}25) = 0{,}0813$, tehát csökkent a valószínűsége, hogy viszonylag nagy tévedést követünk el.

8.2.3. A variancia mintavételi eloszlása

A mintaátlag és a mintabeli arány mintavételi eloszlásának megismerése után foglalkozzunk röviden a mintabeli variancia eloszlásával. Gyakran ugyanis nem csak az a fontos, hogy a termékünk átlagosan megfelelő legyen, hanem ezt a magas minőséget alacsony varianciával is kell előállítanunk, hiszen az egyedi vásárlóinkat nem az átlagos termék, hanem az az egyetlen termék érdekli, amit megvásárolt.

A variancia esetén a vizsgálandó valószínűségi változó meghatározása nem annyira magától értetődő, mint az átlag és az arány esetén (ahol a sokasági paramétert az azonosan képzett mintabeli statisztikával vizsgáltuk). Tekintsük a

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1} \tag{8.8}$$

valószínűségi változót, ahol a szokásos jelölések mellett S^2 jelöli a mintabeli korrigált variancia, így S a mintabeli korrigált szórás valószínűségi változóját. Első ránézésre a formula hasonlít a (2.13) képletre, azzal a különbséggel, hogy N helyett n négyzetösszeget adunk össze, valamint a sokasági átlag helyett a mintaátlag szerepel. Különbség ugyanakkor az is, hogy a nevezőben $n-1$ szerepel a „várt” n helyett. Ennek az az oka, hogy ugyan n független mintaelemet veszünk, azonban miután már ismerjük a mintaátlag értékét, ebből csak $n-1$ darab határozható meg szabadon. A jelenséget szabadságfoknak nevezi a statisztika. A mintabeli varianciát azért számoljuk így, mert ebben az esetben igaz, hogy

$$\mathbf{E}(S^2) = \sigma^2 \tag{8.9}$$

Azt már tudjuk tehát, hogy a mintabeli variancia várható értéke a sokasági variancia, és ez az eredmény viszonylag általános, azonban a mintavételi eloszlását még nem ismerjük. Amennyiben a sokaság eloszlása normális, a mintavételi eloszlásnak egy – eddig ismeretlen – nevezetes eloszláshoz van köze, ami a χ^2 eloszlás[6]. Legyen tehát X egy normális eloszlású sokaság, ismert σ^2 varianciával, ekkor

$$\frac{(n-1)\,S^2}{\sigma^2} \sim {}_{n-1}\chi^2 \tag{8.10}$$

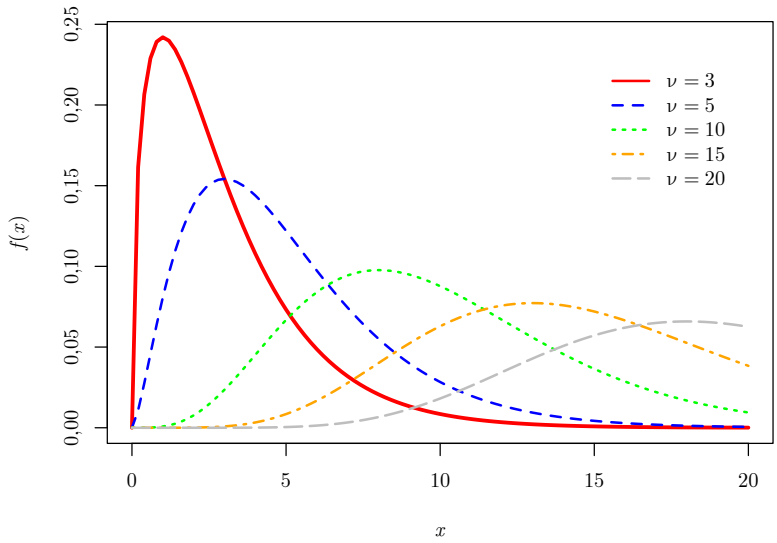
[6]Ejtsd: khí-négyzet eloszlás

azaz $n-1$ szabadságfokú χ^2 eloszlást követ. A χ^2 eloszlást egyetlen paraméter, a szabadságfok segítséggel határozzuk meg. Az $n-1$ szabadságfokú χ^2 eloszlás egyébként $n-1$ darab, egymástól független standard normális eloszlású véletlen változó négyzetösszegeként állítható elő.

A mintabeli átlag és arány mintavételi eloszlása – ahogy láttuk, bizonyos feltételek mellett – a normális eloszláshoz kötődik, a variancia esetén azonban a csak pozitív értékeket felvevő χ^2 eloszlásra kell támaszkodnunk. A $X \sim {}_{\nu}\chi^2$ eloszlás momentumai

$$\mathbf{E}(X) = \nu \qquad \mathbf{D}^2(X) = 2\nu \tag{8.11}$$

A 8.2. ábrán a χ^2 eloszlás sűrűségfüggvénye látható különböző ν szabadságfokok mellett.



8.2. ábra. Különböző szabadságfokú khí-négyzet eloszlások

A χ^2 eloszlás különböző szabadságfokokhoz tartozó kvantilisei erre a célra készített táblázatokból, vagy statisztikai szoftverekből is kinyerhetők. Az eloszlás nem szimmetrikus, amit a kvantilisek meghatározásakor figyelembe kell vennünk.

A $\nu = 15$ szabadságfokú χ^2 eloszlás esetén keressük azt a két kvantilist, melyek 5-5%-ot vágnak le az eloszlás két oldaláról. Ekkor szoftver, vagy táblázat segítségével megállapítható, hogy a keresett alsó kvantilis

$$ {}_{15}\chi^2_{0,05} = 7{,}261$$

a felső kvantilis pedig

$$ {}_{15}\chi^2_{0,95} = 24{,}996$$

Azaz 7,261 és 24,996 közötti található a ${}_{15}\chi^2$ eloszlás középső 90%-a. Itt kell megjegyeznünk, hogy ez az intervallum olyan értelemben középső, hogy balra és jobbra is 5-5%-ot vágunk le, ez azonban nem a legrövidebb intervallum, a gyakorlatban azonban ezt szokás alkalmazni. A legrövidebb intervallum megkeresése túlmutat tananyagunk keretein.

Összefoglalva tehát a σ^2 varianciájú sokaságból vett n elemű mintákból számított S^2 valószínűségi változóról azt tudjuk, hogy várható értéke pontosan a sokasági variancia (lásd (8.9)). Ezen felül, amennyiben a sokaság normális, úgy a (8.10) valószínűségi változó ${}_{n-1}\chi^2$ eloszlást követ. Fontos újfent hangsúlyozni, hogy a variancia mintavételi eloszlása esetén csak abban az esetben tudunk következtetéseket levonni (akkor ismerjük a mintavételi eloszlást), ha a sokaság jó közelítéssel normális eloszlású. Az átlag esetén látott Centrális határeloszlás-tételre itt nem támaszkodhatunk, azaz pusztán a nagy minta alapján nem feltételezhetjük a χ^2 eloszlást. Alkalmazása előtt a normalitást a minta alapján ellenőrizni kell.

Legyen X normális eloszlású sokaság, $\sigma = 5$ szórással. A mintavételi eloszlás megértéséhez készítettük a 8.3. ábrát, ami egyszerűen úgy készült, hogy egy $\sigma = 5$ szórású normális eloszlásból $n = 16$ elemű mintákat vettünk 100000-szer, majd kiszámítottuk a (8.10) változó mintáról mintára ingadozó értékét. A 100000 realizáció ugyan nem az összes lehetséges minta, azonban az ezek alapján rajzolt hisztogram és az elméleti görbe láthatóan jól illeszkednek egymásra. Az ábrán jelöltük az előző példában kiszámított két kvantilis értékét (7,261 és 24,996) is.

Tudjuk tehát, (8.10) alapján, hogy

$$\mathbf{P}\left(7{,}261 < \frac{15S^2}{25} < 24{,}996\right) = 0{,}9$$

amiből egyszerű algebrai átalakítások után adódik, hogy

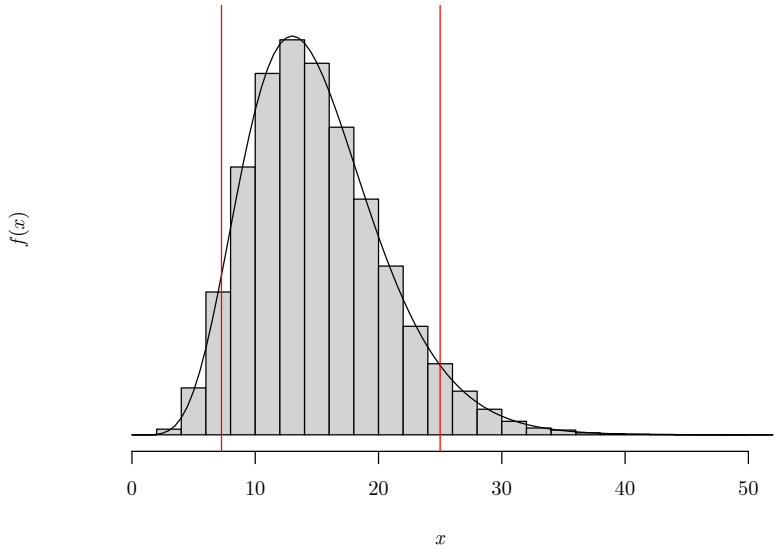
$$\mathbf{P}\left(3{,}479 < S < 6{,}454\right) = 0{,}9$$

azaz a mintabeli szórás az esetek 90%-ában 3,479 és 6,454 közé esik, ha $n = 16$ elemű mintát veszünk egy $\sigma = 5$ szórású normális eloszlásból.

8.3. A Student t eloszlás

Az átlag mintavételi eloszlásáról elegendően nagy, n elemű minta esetén azt találtuk a (8.4) összefüggésben, hogy a

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$



8.3. ábra. A varianciához kapcsolódó valószínűségi változó mintavételi eloszlása

valószínűségi változó standard normális eloszlást követ. A valószínűségi változóban a véletlen elem az $\overline{X}$ változóban rejlik, hisz a mintaátlag mintáról mintára eltérő lehet. A centrális határeloszlás tétel miatt ez a standardizált eredmény széles körben használható. Sok gyakorlati esetben azonban a σ sokasági szórás ismeretlen, így az összefüggés közvetlenül nem használható. Ilyen körülmények között természetes gondolat, hogy σ helyett a mintabeli korrigált szórást alkalmazzuk a képletben, azaz a

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim\sim {}_{n-1}t \tag{8.12}$$

valószínűségi változót kell alkalmaznunk. Ez a valószínűségi változó már nem standard normális eloszlású, illetve a mintabeli átlag mellett a mintabeli szórás is megjelenik, mint valószínűségi változó. A T valószínűségi változót Student t eloszlásúnak nevezzük, aminek szintén egy paramétere van (ν) és szintén szabadságfoknak nevezzük.[7]

A Student t eloszlás nevét nem egy hallgatóról, vagy egy Student nevű tudósról kapta. Az eloszláshoz kapcsolódó matematikai problémát William Sealy Gosset publikálta Student álnéven, aki a Guinness sörfőzde alkalmazásában állt. Az eloszlás kvantiliseit tartalmazó táblázatot azzal a kommenttel küldte el a 20. század talán legnagyobb hatású statisztikusának, Ronald Fishernek, hogy talán

[7] A hasonlóság nem véletlen, a t eloszlás egy standard normális eloszlás és egy χ^2 eloszlás hányadosához kapcsolódik, emiatt azonos a paraméterek száma és elnevezése is.

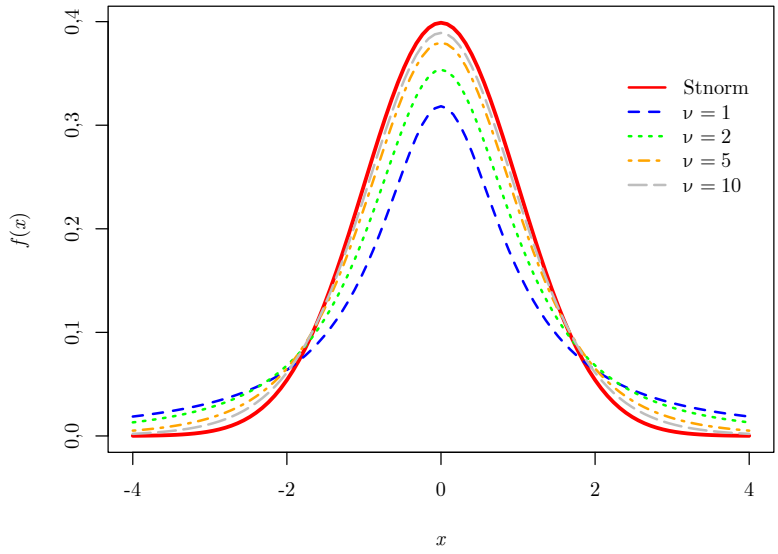
Fisher az egyetlen ember a világon, aki azt használni fogja a megalkotóján kívül. Ennél nagyobbat nehezen tévedhetett volna.

A t eloszlás tehát figyelembe veszi azt a többlet bizonytalanságot is a mintavétel során, hogy adott esetben nem csak a sokasági átlag, hanem a sokasági variancia is ismeretlen. Az eloszlás alakja ettől függetlenül nagyon hasonlít a standard normális eloszláséra, szimmetrikus, azonban az eloszlás farkaiban nagyobb, míg a várható érték körül relatíve kisebb a sűrűségfüggvény értéke. A ν szabadságfokú t eloszlás paraméterei

$$\mathbf{E}(X) = 0\,(\nu > 1)\,, \quad \mathbf{D}^2(X) = \frac{\nu}{\nu - 2}\,(\nu > 2) \tag{8.13}$$

azaz a standard normális eloszláshoz hasonlóan a várható érték 0, a variancia azonban egynél nagyobb, ami pontosan a többlet bizonytalanságot testesíti meg. $\nu = 1$ esetén a $\mathbf{E}(X)$ várható érték nem létezik, valamint $\nu \leq 2$ esetén $\mathbf{D}^2(X)$ sem létezik. A nem létező momentumok esetére jelen tananyagban nem tértünk ki részletesen, de a fentiek a nagyon kis minták esetén problémákat okoznak.

Az is megfigyelhető, hogy a szabadságfok (ami közvetlenül a mintalemszámhoz kapcsolódik) növekedésével a variancia egyre közelebb kerül felülről a standard normális eloszlás 1 értékéhez, hiszen az egyre nagyobb mintából egyre közelebb kerül a mintabeli és a sokasági szórás. Ugyanez figyelhető meg az eloszlások alakján is, a 8.4. ábrán a standard normális eloszlás és különböző szabadságfokú t eloszlások láthatók.



8.4. ábra. A standard normális és különböző szabadságfokú t eloszlások

A t eloszlásból származó kvantilisek táblázata megtalálható a Képletgyűjteményben és természetesen szoftveres úton is meghatározhatók a szükséges értékek. Amennyiben

különböző szabadságfokú t eloszlások és a standard normális eloszlás α kvantiliseit (azokat az értékeket, melyektől α terület van jobbra) hasonlítjuk össze, azt láthatjuk, hogy

- a t eloszlások kvantilisei magasabbak a standard normális értéknél,
- a szabadságfok növekedésével a t eloszlás kvantilisei a normális eloszlásból származó értékhez tartanak, azt minden határon túl megközelítik,
- a fentiek miatt nagy minta esetén tulajdonképpen a standard normális eloszlásból származó érték is használható a gyakorlatban.

Keressük meg a 8.4. ábrán látható t eloszlások esetére azokat a kvantiliseket, melyek 5-5%-ot vágnak le az eloszlás alján és tetején. A szimmetria miatt tudjuk, hogy minden szabadságfok mellett a két keresett érték egymás ellentéte, ezért elegendő az egyiket megkeresnünk. Szoftver, vagy a t eloszlás táblázata alapján azt kapjuk, hogy a keresett pozitív kvantilisek

8.2. táblázat: A t eloszlás 95%-os kvantilisei néhány szabadságfok esetén

ν	${}_{\nu}t_{0,95}$
1	6,314
2	2,920
5	2,015
10	1,812

Azt látjuk, hogy ezek a kvantilisek egyre kisebbek, ami mögött az a tény húzódik meg, hogy nagyobb minta alapján becsült sokasági szórás esetén egyre kevésbé vagyunk bizonytalanok. A szabadságok növekedésével a kvantilisek a normális eloszlásból származó $z_{0,95} = 1{,}645$ értékhez közelítenek.

8.4. Excel tippek

Hasznos Excel függvények:

- GYÖK
- SZÓR.M

Figyeljünk, hogy a SZÓR.M függvény nem azonos a SZÓR.S függvénnyel, amely a sokasági szórás kiszámolására alkalmas! A SZÓR.M függvény ezzel szemben a mintabeli korrigált szórást adja meg.

- T.ELOSZL, T.INVERZ
- KHINÉGYZET.ELOSZLÁS, KHINÉGYZET.INVERZ

Hasznos Excel funkciók:

- Az Adatelemzés menü Mintavétel eszköze 

9. fejezet

Intervallum becslés alapjai

Az előző, 8.2. fejezetben néhány mintabeli statisztika (átlag, arány és korrigált variancia) mintavételi eloszlását ismertük meg, azaz azt tekintettük át, hogy egy adott sokaságból származó összes lehetséges mintabeli statisztika milyen eloszlással jellemezhető. Erre az ismeretre szükségünk van ahhoz, hogy a gyakorlatban sokkal fontosabb kérdéskörrel, a becsléssel foglalkozzunk. Feladatunk egy ismeretlen sokasági paraméter közelítő meghatározása lesz egyetlen minta alapján. Ebben a fejezetben elsőként az ún. pontbecslést, valamint annak tulajdonságait tekintjük át. A pontbecslés értékét az aktuális minta határozza meg, így a mintavételi eloszlásról tanultak alapján tudhatjuk, hogy valószínűtlen, hogy ez éppen eltalálja a sokasági paramétert, ezért a végső célunk egy olyan intervallum meghatározása lesz, mely nagy valószínűséggel tartalmazza a sokasági értéket. Ezt az intervallumot konfidencia intervallumnak, az eljárást intervallum becslésnek nevezzük. Fontos újfent rögzíteni, hogy ettől a fejezettől kezdődően a következtetéses statisztika területére lépünk, azaz a sokasági paraméterek fix, ámde ismeretlen értékűek, rájuk csak egy minta alapján tudunk következtetéseket levonni. Ennek megfelelően a továbbiakban nem beszélünk átlagról, vagy szórásról, ennél pontosabban kell fogalmaznunk, azt is minden esetben jeleznünk kell, hogy mintabeli, vagy sokasági értékről van szó.

9.1. A becslőfüggvény tulajdonságai

Az ismeretlen sokaság paraméterére vonatkozó következtetésünk minden esetben mintabeli információn fog alapulni, azaz egy valószínűségi változón, hiszen a mintánk minden mintavétel esetén más és más lenne. Ezt a valószínűségi változót becslőfüggvénynek, az adott minta esetén kiszámított értékét pedig becsült értéknek nevezzük.[1] A becslőfüggvény tehát egy mintaelemektől mint valószínűségi változóktól

[1] A továbbiakban is próbáljuk a valószínűségszámítás során bevezetett konvenciót követni. A valószínűségi változót nagybetűvel, míg a mintából származó becsült értéket kisbetűvel fogjuk jelölni. Ahol ez már nehezen tartható, gyakran használt jelölés a $\hat{}$. Például a θ paraméter becslőfüggvényét

függő matematikai formula, míg a becsült érték ennek a formulának az adott mintára vonatkozó kiszámított értéke, a valószínűségi változó egy realizációja, egy szám.

Első közelítésben tehát becslőfüggvényt kell meghatároznunk, amivel a becslés végrehajtható. Ez ugyan egyszerűnek tűnik, de mégsem létezik egyetlen legjobb módszer a függvény meghatározására. Az ebben a tananyagban tárgyalt egyszerű esetekre a becslőfüggvény többnyire kézenfekvő, azonban összetettebb esetekre ez már nem igaz. Annak érdekében, hogy a különböző becslőfüggvényeket össze tudjuk hasonlítani, érdemes néhány kritériumot felállítani, a két legfontosabb tulajdonság a torzítatlanság és a hatásosság.

9.1.1. Torzítatlanság

A becslőfüggvények egyik kívánatos tulajdonsága a torzítatlanság, ami alatt a következő tulajdonságot értjük (ahogy azt már az előző fejezetben is láttuk):

$$\mathbf{E}(\hat{\theta}) = \theta \tag{9.1}$$

Ekkor azt mondjuk tehát, hogy a $\hat{\theta}$ becslőfüggvény torzítatlan becslőfüggvénye a θ sokasági paraméternek, a becslőfüggvény minden lehetséges mintán vett várható értéke pontosan a keresett sokasági paraméterrel egyenlő. A torzítatlanság tehát nem azt jelenti, hogy akár csak egyetlen mintabeli érték eltalálja a sokasági értéket, hanem azt, hogy a mintabeli értékek várható értékben a sokasági paramétert adják.

Ahogy azt a (8.1) összefüggésünk mutatta, a mintabeli átlagok átlaga (várható értéke) megegyezik a sokasági átlaggal, azaz a mintaátlag $\hat{\theta} = \overline{X}$ torzítatlan becslőfüggvénye a sokasági átlagnak $\theta = \mu$. A mintabeli átlag azonban nem az egyetlen lehetséges – és nem is minden esetben a legjobb – becslőfüggvénye a sokasági átlagnak. A mintabeli medián is egy lehetséges, de nem minden esetben jó becslőfüggvény.

Amennyiben a σ^2 sokasági variancia becslőfüggvényét keressük, a torzítatlan becslőfüggvényt a mintabeli korrigált variancia adja, de itt is elképzelhető lenne más becslőfüggvény, pl. a mintabeli terjedelem osztva hárommal. Természetesen elméleti megalapozás nélkül ez a becslőfüggvény valószínűleg nem teljesítene túlságosan jól.

Vannak olyan esetek, amikor nem található torzítatlan becslőfüggvény, azaz nem teljesül a (9.1) összefüggés. A torzítás mértékét a

$$\mathbf{E}(\hat{\theta}) - \theta \tag{9.2}$$

$\hat{\theta}$ jelöli, általános esetben, amikor egy tetszőleges paraméterről beszélünk, mi is ezt a jelölést fogjuk használni.

különbség méri, ami jellemzően n, a mintaelemszám változásával változik[2]. Amennyiben $n \to \infty$ esetén a torzítás a 0-hoz tart, a tulajdonságot aszimptotikus torzítatlanságnak nevezzük.

A 9.1. táblázatban az általunk eddig ismert és leggyakrabban használt paramétereket és becslőfüggvényeket foglaljuk össze. Valamennyi becslőfüggvényről elmondható, hogy azok torzítatlanok, azaz a hozzájuk tartozó torzítás 0.

9.1. táblázat: Sokasági paraméterek és becslőfüggvényeik

paraméter	θ	becslőfüggvény	$\hat{\theta}$
sokasági átlag	μ	mintabeli átlag	$\overline{X}$
sokasági variancia	σ^2	mintabeli korrigált variancia	S^2
sokasági arány	π	mintabeli arány	P

9.1.2. Hatásosság

A gyakorlatban sok esetben több torzítatlan becslőfüggvény is található, közülük segíthet választani a hatásosság. Azt a becslőfüggvényt preferáljuk a torzítatlan becslőfüggvények közül, melyek jobban koncentrálódnak a becsülni kívánt paraméter körül. Ez azt jelenti, hogy a hatásosabb becslőfüggvény esetén a mintabeli értékek átlagosan közelebb vannak a sokasági értékhez, mint a kevésbé hatékony alternatíva esetén.

Legyen $\hat{\theta_1}$ és $\hat{\theta_2}$ két, θ-ra vonatkozó, ugyanakkora mintából számított becslőfüggvény, akkor $\hat{\theta_1}$ hatásosabb mint $\hat{\theta_2}$, ha $\mathrm{Var}(\hat{\theta_1}) < \mathrm{Var}(\hat{\theta_2})$.

Léteznek ún. abszolút hatásos becslőfüggvények, melyek adott feltételek mellett bizonyíthatóan a legkisebb varianciájú becslést eredményezik.

Legyen X normális eloszlású sokaság, melyből függetlenül vett nagy, n elemű minta átlaga torzítatlan, és azt is tudjuk, hogy

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

Belátható, hogy a mintabeli medián is torzítatlan becslőfüggvénye a sokasági átlagnak nagy n esetén, valamint

$$\mathrm{Var}(\hat{\mathrm{me}}) = \frac{\pi}{2}\frac{\sigma^2}{n}$$

[2]Amennyiben nem változna, elegendő lenne a torzítást levonni a becslőfüggvényből, hogy torzítatlan becslőfüggvényt kapjunk.

Mivel a mintaátlag varianciája alacsonyabb a mintabeli medián varianciájánál ($\mathrm{Var}(\overline{X}) < \mathrm{Var}(\hat{\mathrm{me}})$), ezért a fenti feltételek mellett az átlag hatásosabb becslőfüggvény, mint a medián. Nem szabad azonban elfelejtenünk, hogy lehetnek olyan esetek, például kiugró értékek esetén, amikor a medián hatásosabb becslőfüggvény a sokasági átlagra vonatkozóan. Ennek vizsgálata meghaladja tankönyvünk kereteit.

9.2. Konfidencia intervallum becslés

Ebben a fejezetben elsőként a konfidencia intervallum fogalmáról általánosan beszélünk, majd a sokasági átlag (várható érték) becslés esetén több lépésen keresztül érkezünk el a gyakorlatban leggyakrabban alkalmazott formuláig. Az arányra és a varianciára vonatkozó intervallum becslés logikája hasonló, így ezeket jóval kevésbé részletesen tárgyaljuk.

A 9.1. fejezetben azt vizsgáltuk, hogy a pontbecslések minőségét milyen tulajdonságokkal jellemezhetjük. A torzítatlanság egy kívánatos tulajdonság, de pusztán annyit jelent, hogy a minták összességében a becslésünk eltalálja a sokasági értéket. A gyakorlatban ezzel szemben nincs lehetőségünk az összes lehetséges mintát kiválasztani, jellemzően egy minta alapján szeretnénk a sokaságra következtetni. Ekkor tudjuk, hogy az egyetlen mintánkból származó pontbecslésünk nagy valószínűséggel nem egyezik meg pontosan a sokasági paraméterrel. A célunk ezért egy olyan intervallum meghatározása, ami a sokasági paramétert már kellően nagy megbízhatósággal tartalmazza. Még inkább alátámasztja ennek a célnak a jogosságát az a tény, hogy a pontbecslés nem veszi figyelembe a minta nagyságát, annak ellenére, hogy tudjuk, egy kis mintából nyert pontbecslés nem olyan hasznos, mint egy nagy mintából nyert. A konfidencia intervallum becslés erre a problémára is megoldást nyújt.

A konfidencia intervallum meghatározásához tehát két valószínűségi változót kell meghatároznunk, melyek az intervallum alsó és felső határát jelölik ki. Eddig csak azt mondtuk, hogy azt szeretnénk, hogy nagy megbízhatósággal tartalmazza ez az intervallum a keresett sokasági paramétert. Nagy megbízhatóság alatt azt értjük, hogy amennyiben újra és újra mintát vennénk és újra és újra kiszámítanánk a (később meghatározandó) konfidencia intervallumot, akkor az esetek pl. 95%-ában az intervallum fedje le a sokasági paramétert. Az egyetlen minta alapján kiszámított alsó és felső határ vagy lefedi a sokasági paramétert, vagy sem, ezt nem tudjuk, de azt állíthatjuk, hogy a hasonló eljárással készült intervallumok 95%-a (vagy más százalékos értéke) fedi a paramétert. Az ilyen intervallumot 95%-os megbízhatóságú konfidencia intervallumnak nevezzük.

Legyen a keresett sokasági paraméter θ, amire mintabeli információ alapján szeretnénk becslést készíteni. Jelölje a két, mintabeli adatoktól függő valószínűségi változót X_a és X_f, melyek az intervallum alsó és felső határát adják. Célunk X_a és X_f meghatározása úgy, hogy $\mathbf{P}\left(X_a < \theta < X_f\right) = 1-\alpha$, ahol α tetszőleges 0 és 1 közötti érték (általában 0 közeli). A minta alapján kiszámított értékek legyenek x_a és x_f, ekkor az általuk meghatározott intervallumot θ-ra vonatkozó $100(1-\alpha)\%$ megbízhatóságú konfidencia

intervallumnak nevezzük, $100(1-\alpha)\%$ pedig a konfidencia szint, vagy megbízhatósági szint.

Amikor mintavétel történik, akkor a mintabeli statisztikától azt várjuk, hogy a sokasági paraméter környezetében lesz az értéke, némely minta esetén annál nagyobb, némely minta esetén annál kisebb értéket fogunk kapni, ezért természetes, hogy a legtöbb (de nem minden) konfidencia intervallum a

$$\hat{\theta} \pm \Delta \tag{9.3}$$

alakot ölti, azaz a pontbecslésre szimmetrikus a konfidencia intervallum alsó és felső határa. A Δ neve hibahatár, meghatároza a különböző paraméterek és feltevések mellett a következő fejezetekben történik.

9.2.1. Várható értékre vonatkozó intervallum becslés

A leggyakrabban becsült sokasági jellemző az átlag, más kifejezéssel a várható érték. Matematikailag a legegyszerűbb eset, ha egy normális eloszlást tekintünk, aminek μ várható értéke ismeretlen, σ^2 varianciája azonban ismert. A várható értéket egy n elemű minta alapján kívánjuk megbecsülni. Tudjuk, hogy

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

Jelölje továbbá Z egy standard normális eloszlású véletlen változót, $z_{1-\alpha/2}$ pedig azt a standard normális eloszlásból származó kvantilist, melytől balra eső terület $1-\alpha/2$, azaz a jobbra eső terület $\alpha/2$. Az eloszlás szimmetriája alapján azt is tudjuk, hogy

$$\mathbf{P}\left(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}\right) = 1-\alpha$$

ami épp a kívánt konfidencia szint. Ebből egyszerű algebrai átalakítások segítségével adódik a konfidencia intervallum

$$\begin{aligned}
\mathbf{P}\left(-z_{1-\alpha/2} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha/2}\right) &= 1-\alpha \\
\mathbf{P}\left(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha \\
\mathbf{P}\left(\overline{X}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha
\end{aligned} \tag{9.4}$$

azaz egy μ-re vonatkozó valószínűségi állítást kaptunk, épp a (9.3) egyenletben definiált formában. A hasonlóság még inkább látható, ha az utolsó sort más formában írjuk. Az $1-\alpha$ megbízhatóságú konfidencia intervallum a minta alapján[3]

$$\overline{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = \overline{x} \pm \Delta_{\overline{X}} \tag{9.5}$$

azaz az átlagra vonatkozó konfidencia intervallumot a mintaátlag (pontbecslés) köré képezzük, $\Delta_{\overline{X}}$ pedig az átlagra vonatkozó hibahatár. A keletkező konfidencia intervallum hossza tehát a hibahatár kétszerese, értékét a sokasági szórás, a mintaelemszám, illetve a z-értéken keresztül a megbízhatóság befolyásolja. Amennyiben a sokasági szórás magas, úgy a lehetséges mintaátlagok is jelentősen eltérhetnek, ezért a becslésünk is pontatlanabb lesz. Természetesen a mintaelemszám növekedése épp ellentétes hatású, azonban vegyük észre, hogy a gyökös tag miatt a mintaelemszám növelése egyre kevésbé hatékony. Ha a becsült intervallum hosszát a felére szeretnénk csökkenteni, négyszer akkora mintára van szükségünk. Ha a negyedére, akkor már 16-szoros minta kiválasztására van szükség.

A $z_{1-\alpha/2}$ kvantilisek táblázatból, vagy szoftver segítségével könnyen megtalálhatók, azonban néhány gyakran használt megbízhatósági szint esetére érdemes őket megjegyezni.

9.2. táblázat: Gyakran használt megbízhatósági szintek és normális kvantilisek

Konfidencia szint $(1-\alpha)$	90%	95%	95,5%	99%
$z_{1-\alpha/2}$	1,645	1,96	2,00	2,576

A táblázat alapján jól látható, hogy a megbízhatóság növekedésével a z-értékek is emelkednek, ezt a jelenséget a statisztika úgy fogalmazza meg, hogy adott mintaelemszám (és mintavételi módszer) mellett a pontosság (az intervallum hossza) és a megbízhatóság $(1-\alpha)$ csak egymás rovására javíthatók.

Egy üzemben a régóta működő gépről ismert, hogy a legyártott alkatrészek számunkra fontos hossza jó közelítéssel normális eloszlású és a várható értéktől függetlenül a szórás $\sigma = 5$ mm. Az új megrendeléshez arra van szükség, hogy az alkatérsz hossza 1000 mm legyen. Az új beállításokkal próbagyártást végeztek az üzemben, amiből $n = 10$ elemű mintát vettünk, a mintaátlag $\overline{x} = 1002{,}74$. Becsüljünk 95%-os megbízhatóságú konfidencia intervallumot a sokasági átlagra (μ) vonatkozóan!

[3] A levezetésben valószínűségi állítást tettünk, amit csak valószínűségi változókkal tehetünk meg, ezért használtuk a $\overline{X}$ jelölést. Az adott mintából kiszámított érték, az adott realizáció jele $\overline{x}$, az egyszerűség kedvéért innentől gyakran fogjuk ezt a kisbetűs jelölést is alkalmazni.

Mivel a sokaságról feltételezhetjük, hogy normális eloszlást követ, alkalmazhatjuk (9.5) formulát, azaz a 95%-os megbízhatóságú konfidencia intervallumra

$$1002{,}743 \pm 1{,}96\frac{5}{\sqrt{10}} = 1002{,}743 \pm 3{,}099$$

adódik. A mintánk alapján 95%-os megbízhatósággal állíthatjuk, hogy az ismeretlen sokasági átlag 999,644 és 1005,842 mm között van.

A (9.5) formula abban az esetben alkalmazható, ha a sokaság normális, ismeretlen μ, de ismert σ paraméterekkel. A gyakorlatban ugyan előfordulhat ilyen eset (pl. korábbi tanulmányokból ismerhetjük a varianciát), azonban érezhetően nem túl gyakori eset. Tegyük most fel, hogy a sokasági variancia nem ismert, ezért (9.5) formula nem alkalmazható. A 8.3. fejezetben megismert mintavételi eloszlás azonban segítségünkre lehet. (8.12) formula szerint a

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim {}_{n-1}t$$

valószínűségi változó $n-1$ szabadságfokú t eloszlást követ, amiből (9.5) képlethez hasonló módon levezethető az átlagra vonatkozó konfidencia intervallum becslés mintából számítandó értéke ismeretlen sokasági variancia esetére is

$$\overline{x} \pm {}_{n-1}t_{1-\alpha/2}\frac{s}{\sqrt{n}} = \overline{X} \pm \Delta_{\overline{X}} \tag{9.6}$$

ahol ${}_{n-1}t_{1-\alpha/2}$ az $n-1$ szabadságfokú t eloszlás $1-\alpha/2$ kvantilise. A t eloszlásból származó kvantilisek (azonos megbízhatóság esetén) szabadságfoktól függetlenül magasabbak a 9.2. táblázatban látható értékektől, a szabadságfok növekedésével felülről egyre inkább megközelítve azokat.

A (9.6) formula igazi jelentőségét az adja, hogy nem csak abban az esetben alkalmazható, ha a sokaság normális eloszlású, hanem abban az esetben is, ha a minta elegendően nagy. Ekkor a centrális határeloszlás tétel az, ami megengedi a képlet alkalmazását. Annak meghatározása, hogy mi számít elegendően nagynak nem egyszerű matematikai feladat, alapvetően a sokaság normalitástól való eltérésének mértéke határozza ezt meg. A legtöbb tankönyvben a nagy minta határát valahol 30 és 100 között szokás meghúzni. Ebben a tananyagban a 30 feletti mintaelemszámot elegendően nagy mintának fogjuk tekinteni ahhoz, hogy a központi határeloszlás-tétel működésében már bízzunk, kiemelve azt, hogy extrém eloszlások esetén akár több ezer elemű minták sem biztosítják a (9.6) formula alkalmazhatóságát. A gyakorlatban (9.6) alkalmazása előtt ezért mindenképp javasoljuk a sokaság normalitásának vizsgálatát.

Meg szeretnénk becsülni 90%-os megbízhatósággal egy adott településen az egyetemisták által havonta lakhatásra költött összeg átlagát. Ehhez $n = 50$ elemű mintát választottunk. A sokaság eloszlásának típusáról ugyan nem rendelkezünk ismeretekkel, de mivel a minta elegendően nagy, illetve a készített hisztogram

alapján a sokaság jelentősen nem tér el a normálistól, alkalmazhatjuk a (9.6) formulát. Tegyük fel, hogy a mintabeli átlagra $\overline{x} = 47\,543$ Ft, míg a mintabeli korrigált szórásra $s = 13\,342$ Ft adódott. Ekkor

$$\overline{x} \pm {}_{49}t_{0,95}\frac{s}{\sqrt{n}} = 47543 \pm 1{,}6765\frac{13342}{\sqrt{50}} = 47543 \pm 3163{,}4$$

azaz 90%-os megbízhatósággal állíthatjuk, hogy az átlagos lakhatásra költött összeg a sokaságban 44,380 és 50,706 forint közötti.

A fejezetben eddig bemutatott becslési eljárások ((9.5) és (9.6)) független azonos eloszlású mintavételt tételeznek fel, azaz praktikusan azt, hogy N, a sokaság elemszáma végtelen, vagy annyira nagy, hogy a visszatevés nélküli mintavétel sem változtatja meg jelentősen a sokaság összetételét. Abban az esetben, ha N ismert, úgy a (9.5) és (9.6) képletben alkalmazott standard hiba módosul,

$$\overline{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{9.7}$$

valamint

$$\overline{x} \pm {}_{n-1}t_{1-\alpha/2}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{9.8}$$

adja a konfidencia intervallum becslés alsó és felső határát. A formulákban megjelenő, a (8.5) összefüggésben már megismert véges szorzóról megállapítottuk, hogy a standard hiba értékét csökkenti, azaz a visszatevés nélküli mintavétellel pontosabb becslést végezhetünk.

Sok gyakorlati esetben nem a sokasági átlagra, hanem az értékösszegre vonatkozó becslést szeretnénk elvégezni, azaz a $\sum X_i = N\mu$ sokasági értékét szeretnénk közelítőleg ismerni. Belátható, hogy ebben az esetben a becslés elvégezhető két lépésben:

- a szituációnak megfelelő becslés elvégzése μ-re vonatkozóan (9.7), vagy (9.8) segítségével, majd
- a konfidencia intervallum alsó és felső határának N-nel történő szorzása.

Egészítsük ki az előző, lakhatásra fordított összeggel foglalkozó példánkat azzal, hogy tudjuk, $N = 20\,000$ hallgató tanul az adott városban. A standard hiba ekkor a véges szorzóval módosul, azonban a hatás elenyésző, hiszen a sokaság viszonylag nagy a minta méretéhez képest, így a valamivel szűkebb

$$47\,543 \pm 3155{,}6$$

intervallum adódik. A sokaság elemszámának ismerete azonban lehetőséget ad az értékösszeg becslésére, ezek szerint 90%-os megbízhatósággal az egyetemisták legalább 887 748 000, legfeljebb 1 013 972 000 forintot költenek összesen lakhatásra havonta.

9.2.2. Arányra vonatkozó intervallum becslés

Az átlagra vonatkozó intervallum becsléshez nagyon hasonló megfontolások és a (8.6) összefüggés alapján belátható, hogy

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} = p \pm \Delta_P \tag{9.9}$$

az arányra vonatkozó $1-\alpha$ megbízhatósági szintű konfidencia intervallum. Ahogyan azt az arány mintavételi eloszlásánál is láttuk a 8.2.2. fejezetben, a normális eloszlással történő közelítés, tehát (9.9) abban az esetben alkalmazható, ha $n\pi > 5$ és $n(1-\pi) > 5$ is teljesül. Az $n\pi > 5$ és $n(1-\pi) > 5$ feltételeket természetesen nem tudjuk közvetlenül ellenőrizni, hisz feladatunk épp a π binomiális paraméter becslése, azonban a mintabeli arány segítségével az ellenőrzése elvégezhető. Léteznek becslési eljárások a kis minta esetére is, ezekkel azonban jelen tananyagunkban nem foglalkozunk részletesen.

A sokasági átlagra vonatkozó konfidencia intervallumot tehát a mintabeli arány mint pontbecslés köré rajzoljuk. Általánosságban elmondható, hogy n növekedésével a hibahatár csökken és ezzel együtt a becslés pontossága javul. A $p(1-p)$ kifejezés értéke a 0-1 tartományon $p = 0{,}5$ esetén maximális, azaz adott mintaelemszám mellett a 0,5 körüli mintabeli arány adja a leginkább pontatlan becslést, illetve a $p(1-p)$ szorzat szimmetriája miatt a 0,5 értéktől távolodva a becslés egyre pontosabb.

A sokasági átlag becsléséhez hasonlóan belátható, hogy véges N sokasági elemszám és visszatevés nélküli mintavétel esetén (9.9) helyett a

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}} \tag{9.10}$$

formula használatos, azaz a standard hiba ebben az esetben is a véges szorzóval módosul.

A gyakorlatban sok esetben nem az adott tulajdonsággal rendelkező megfigyelések sokasági arányára, hanem azok sokasági számára vagyunk kíváncsiak. Az értékösszeg becsléséhez hasonlóan két lépésben oldható meg a feladat:

- a feltételek ellenőrzése, majd becslés végzése π-re vonatkozóan, majd
- a konfidencia intervallum alsó és felső határának N-nel történő szorzása.

Egy egyszerű véletlen mintavétel adatai alapján a megkérdezett $n = 1000$ választó közül $k = 645$ válaszolta azt, hogy a következő hétvégén részt venne a választásokon. Készítsünk 99%-os megbízhatóságú konfidencia intervallumot a π sokasági arányra vonatkozóan! Mivel a mintánk meglehetősen nagy, ezért alkalmazhatjuk a (9.9) formulát.

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0{,}645 \pm 2{,}576\sqrt{\frac{0{,}645(1-0{,}645)}{1000}} = 0{,}645 \pm 0{,}039$$

azaz a 99%os megbízhatóságú konfidencia intervallumunk alsó határa 60,6%, felső határa 68,4%, becslésünk szerint a választók összességében a két érték közötti lesz a részvételi arány.

Amennyiben tudjuk, hogy $N = 4\,000\,000$ szavazásra jogosult van, alkalmazhatjuk a (9.10) formulát is, azonban a véges szorzó alkalmazása érdemben nem fogja megváltoztatni a végeredményt. Azt azonban tudhatjuk, ugyancsak 99%-os megbízhatóság mellett, hogy a választás 2 millió 424 ezer és 2 millió 736 ezer főt fog várhatóan mozgósítani.

9.3. Mintaelemszám tervezés

Az előző, 9.2. fejezetben néhány sokasági paraméterre vonatkozó becslési eljárást tekintettünk át. Néha az adott minta alapján kapott intervallum túl széles, azaz a becslésünk túl pontatlan. Ezen többek között további mintavételezés segíthet. Más esetekben előre, a kísérlet megkezdése előtt rögzített a tolerálható hiba mértéke. Az ebben a fejezetben taglalt formulák abban segítenek, hogy ezekben a szituációkban hogyan számítsuk ki a szükséges minta nagyságát. A formulák természetesen különbözők abban az esetben, ha várható értékre és ha arányra vonatkozó a feladat, illetve az is fontos szempont, hogy FAE (végtelen nagy sokaság és/vagy visszatevéses minta), vagy EV mintavétel (véges sokaság, visszatevés nélkül) történik-e.

Egy normális eloszlású, ismert σ^2 varianciájú, ismeretlen μ várható értékű alapsokaságból vett n elemű minta esetén az $1-\alpha$ megbízhatóságú konfidencia intervallum (9.5) alapján

$$\overline{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = \overline{x} \pm \Delta_{\overline{X}}$$

amiből ha ismert $\Delta_{\overline{X}}$, akkor egyszerű algebrai átalakítások segítségével azt kapjuk, hogy

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{\Delta^2_{\overline{X}}} \tag{9.11}$$

Ez a mintaelemszám választás biztosítja, hogy a konfidencia intervallum hossza a kívánt határértéken belül marad. A képlet nem garantálja, hogy az eredmény egész szám legyen, így a gyakorlatban a (9.11) formula eredményét felfelé kerekítjük.

Az arányra vonatkozó mintaelemszám hasonló logika mentén határozható meg, felhasználva az arányra vonatkozó standard hibát

$$n = \frac{z^2_{1-\alpha/2}\pi(1-\pi)}{\Delta_P^2} \tag{9.12}$$

ami azonban közvetlenül nem használható, hiszen a képlet tartalmazza a keresett $0 < \pi < 1$ binomiális paramétert, ami a mintaelemszám tervezésekor még nem lehet ismert. A π paraméter helyett használhatunk a sokasági arányra vonatkozó korábbi információt, vagy szakértői becslést, ami természetesen nem lesz tökéletesen pontos, azonban a mintaelemszám tekintetében hozzávetőleges értéket adhat. A másik lehetőség a $\pi(1-\pi)$ függvény már említett tulajdonságának kihasználása, amiről tudjuk, hogy értékének maximuma a 0-1 intervallumon 0,25. Amennyiben tehát nem rendelkezünk információval a sokasági arányról, a legrosszabb esetet feltételezhetjük és a $\pi(1-\pi) = 0{,}25$ helyettesítéssel élhetünk. Amennyiben a megbízhatóságot 95,5%-osnak választjuk, a (9.12) formula még tovább egyszerűsödik, hisz ekkor $z_{1-\alpha/2}$ értéke pontosan 2. Ebben a speciális esetben tehát a klasszikus, könnyen megjegyezhető

$$n = \frac{1}{\Delta_P^2} \tag{9.13}$$

formulát alkalmazhatjuk.

A (9.11) és (9.12) formulák tehát abban az esetben alkalmazandók, ha N ismeretlen, vagy nagyon nagy. Ha ismert a sokaság elemszáma, akkor a mintaelemszám tervezésekor érdemes ezt is figyelembe venni, főként ha a kiválasztási arány várhatóan viszonylag nagy lesz. Ezekben az esetekben a mintaelemszám tervezés teljesen analóg módon a (9.7) és (9.10) formulákból indulunk ki, és a feladat n kifejezése adott hibahatár mellett. A feladatot az bonyolítja kissé, hogy n két helyen is megjelenik. Belátható azonban, hogy ha n_0 jelöli a FAE hibahatár alapján számított szükséges mintaelemszámot, akkor a véges szorzó miatti korrekció a

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \tag{9.14}$$

módon végezhető el[4]. A korábbiakhoz hasonlóan (9.13) sem feltétlenül egész számot ad eredményül, így a végeredményt felfelé kerekítjük a gyakorlatban. Vegyük észre, hogy a módosított mintaelemszám n_0 értékénél releváns esetben kisebb, hiszen n_0-t egy 1-nél nagyobb számmal osztjuk. A korrekció annál nagyobb, minél nagyobb a tervezett mintavételi arány.

[4] Ahogy említettük, a véges szorzó közelítő értéke, $1 - \frac{n}{N}$ a korrekcióra ezzel $n = \frac{n_0}{1+\frac{n_0}{N}}$ adódik. Sok tankönyvben ez szerepel n_0 behelyettesítésével az átlag és az arány esetére.

Egy közvéleménykutatás során 2 százalékpontos pontossággal szeretnénk meghatározni egy adott párt támogatottságát egy kisvárosban, 90%-os megbízhatóság mellett. Mivel egy arány meghatározásáról van szó, ezért (9.12) formulából indulunk ki. Tegyük fel továbbá, hogy semmiféle információval nem rendelkezünk π értékéről, így a legrosszabb forgatókönyvvel, $\pi = 0{,}5$-tel számolunk. Ekkor

$$n = \frac{z^2_{1-\alpha/2}\pi(1-\pi)}{\Delta_P^2} = \frac{1{,}645^2 \times 0{,}5(1-0{,}5)}{0{,}02^2} = 1690{,}96$$

azaz 1691 elemű mintát javasolhatunk. Mivel a $\pi(1-\pi)$ kifejezés maximumával számoltunk, ha a párt támogatottsága nagyon alacsony, vagy nagyon magas, akkor ennél kisebb minta is elegendő lehet, vagy a másik oldalról megközelítve: ez a mintaelemszám 2 százalékpontnál alacsonyabb hibahatárt is eredményezhez. Abban az esetben, ha rendelkezünk megbízható szakértői becsléssel π értékéről, természetesen azt is alkalmazhatjuk.

Tegyük fel továbbá, hogy tudjuk, a kisvárosban $N = 20\,000$ választópolgár él. Ekkor a szükséges mintaelemszám az előző eredmény és (9.14) korrekció segítségével

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} = \frac{1690{,}96}{1 + \frac{1690{,}96-1}{20000}} = 1559{,}21$$

azaz a sokaság elemszámát figyelembe véve azt kapjuk, hogy visszatevés nélkül elegendő 1560 választópolgár megkérdezése, még a legrosszabb esetben is.

9.4. Excel tippek

Hasznos Excel függvények:

- SZÓR.M
- NORM.S.INVERZ, T.INVERZ (z érték és t érték meghatározásához)

Hasznos Excel funkciók:

- Az Adatelemzés menü Leíró statisztika eszköze 

10. fejezet

Összetett becslések

A 9. fejezetben rövid elméleti áttekintés után egyetlen sokaság egy kiválasztott paraméterének értékére adtunk konfidencia intervallum becslést. Gyakran szükséges azonban két sokaság összehasonlítása. A statisztikai következtetés két fontos területe a két sokaságban megfigyelhető várható érték/átlag, illetve a két sokasági arány összehasonlítása. A két sokasági paraméter összehasonlítása technikailag a különbségük becslésén keresztül fog megvalósulni, ilyen értelemben az ebben a fejezetben bemutatandó becslési eljárások a 9. fejezetben bemutatott eszközök általánosításaként tekinthetők. Ennek megfelelően formátumuk a (9.3) formulának megfelelő lesz.

A két sokasági várható érték különbségére vonatkozó konfidencia intervallum becslés ennek megfelelően a

$$\overline{x} - \overline{y} \pm \Delta \tag{10.1}$$

alakú lesz, ahol $\overline{x}$ az egyik, $\overline{y}$ pedig a másik sokaságból vett minta átlaga. A Δ hibahatár kiszámításának pontos képlete attól függ, hogy milyen szituációban történt a mintavétel, illetve milyen feltételezésekkel élhetünk a sokaságokról. A különböző szituációkat és a hibahatárok képletét mutatja be a 10.1. fejezet.

Hasonló logika mentén képezhetjük a két sokasági arány különbségére vonatkozó konfidencia intervallumot is

$$p_X - p_Y \pm \Delta \tag{10.2}$$

ahol p_X és p_Y a két sokaságból vett független mintabeli arány. A hibahatár kiszámítási módját és a szükséges feltételeket a 10.2. fejezet mutatja be.

10.1. Sokasági várható értékek különbségének becslése

A sokasági várható értékek, vagy sokasági átlagok különbségének becslése során két nagy csoportot különböztetünk meg a mintavételi szituáció alapján:

- párosított (függő) minták és
- független minták segítségével történő becslés.

Párosított mintáról (10.1.1. fejezet) akkor beszélünk, ha az egyik mintában található mintaelemek befolyásolják a másik mintába kerülő mintaelemek értékét. Két alapvető szituációban fordul elő ez:

- ugyanazokat a megfigyeléseket mérjük két időpontban (előtte-utána tanulmányok), vagy
- mesterségesen párosított minták esetén.

A leggyakrabban klinikai vizsgálatok esetén fordul elő, hogy páciensek kezelés előtti és utáni átlagos jellemzőinek változását szeretnénk vizsgálni, de a gazdaságtudományok területén is gyakori ez az eset, pl. a fogyasztók marketingkampány előtti és utáni elégedettségét, vagy a beosztottak tréning előtti és utáni hatékonyságát kívánjuk vizsgálni. Ilyen esetekben természetes párokat alkotnak a megfigyeléseink, hiszen ugyanazokat a személyeket vizsgáltuk a „beavatkozás" előtt és után. Néhány esetben ez a természetes párosítás nem kivitelezhető, ekkor gyakran mesterségesen párosított mintákat alkalmaz a statisztikai gyakorlat: például azonos korú, iskolai végzettségű, érdeklődésű fogyasztót választunk az egyik és másik csoportba is és a válaszaikat párosítva elemezzük. Ilyen módszert választhatunk például két konkurens reklámfilm közüli választásra is fókuszcsoportos fogyasztói visszajelzések alapján.

Abban az esetben, ha a két sokaságból egymástól függetlenül veszünk mintát, azaz a mintába kerülő elemek nem befolyásolják a másik minta értékeit, független mintás becslésről beszélünk (10.1.2. fejezet). Ebben az esetben a sokaságra vonatkozó ismereteink és feltevéseink fogják (10.1) pontos formuláját megadni.

A jelölésrendszerünket annyiban kell módosítanunk a korábbiakhoz képest, hogy most már nem egy, hanem két sokaságról, illetve két mintáról kell beszélnünk. A sokaságot eddig X jelölte, legyen ez most az egyik sokaság, míg a másikat jelölje Y. Ennek megfelelően a[1] a következő jelöléseket vezetjük be:

- sokasági átlagok: μ_X és μ_Y

[1] A fejezetben néhány esetben valószínűségi változókról fogunk beszélni, ezért a pontosság kedvéért itt bevezetünk nagybetűs jelöléseket is.

- sokasági átlagok különbsége: $\delta = \mu_X - \mu_Y$
- sokasági varianciák: σ_X^2 és σ_Y^2
- sokasági elemszámok: N_X és N_Y
- mintaelemszámok: n_X és n_Y
- mintabeli átlagok valószínűségi változói: $\overline{X}$ és $\overline{Y}$
- mintabeli átlagok realizációi: $\overline{x}$ és $\overline{y}$
- mintabeli varianciák valószínűségi változói: S_X^2 és S_Y^2
- mintabeli varianciák realizációi: s_X^2 és s_Y^2

10.1.1. Párosított mintás becslés

Tegyük fel, hogy két párosított mintából rendelkezünk $n_X = n_Y = n$ megfigyelléspárral és feladatunk $\delta = \mu_X - \mu_Y$ sokasági különbségre vonatkozó konfidencia intervallum készítése. Amenniyben feltételezzük, hogy a különbség normális eloszlású (vagy elegendően nagy mintával rendelkezünk), úgy $1 - \alpha$ megbízhatóságú konfidencia intervallum becslést a

$$\overline{x} - \overline{y} \pm {}_{n-1}t_{1-\alpha/2}\frac{s_D}{\sqrt{n}} \tag{10.3}$$

formulával nyerhetjük, ahol a már megismert jelölések mellett s_D az n darab párból képzett $D = X - Y$ változóból számított mintabeli szórást jelenti.

A (10.3) formula az előző fejezetből ismerős, hiszen nagyon hasonlít a (9.6) képletre, különösen akkor, ha észrevesszük, hogy $\overline{x} - \overline{y} = \overline{d}$, azaz a pontbecslés az újonnan létrehozott $D = X - Y$ változó mintabeli átlaga. Ez egyben azt is jelenti, hogy a párosított mintás becslést visszavezettük a 9.2.1. fejezetben megismert módszerre. A becsléshez ugyanúgy az $n - 1$ szabadságfokú t eloszlás kvantilisére van szükségünk, ahol n a megfigyelléspárok, azaz a képezhető különbségek száma.

Egy nagy cég HR osztályán feladatunk egy drága kompetenciafejlesztő tréning hatásának mérése. Az első tréningen 30 fő vett részt, akik a cégnél dolgozók véletlen mintájaként tekinthetők. A résztvevők kitöltöttek egy-egy tesztet a tréning előtt és után, melyek pontszámát a 10.1. táblázat tartalmazza. Becsüljük meg 99%-os megbízhatósággal a teszten mért képesség várható javulását a sokaságban!

10.1. táblázat: A tréningen elért eredmények

sorszám	előtte (Y)	utána (X)	változás (D)	sorszám	előtte (Y)	utána (X)	változás (D)
1.	57,5	63,4	5,9	16.	45,4	44,3	-1,1
2.	52,0	54,9	2,9	17.	42,7	46,3	3,6
3.	72,1	76,3	4,2	18.	37,1	34,5	-2,6
4.	47,2	50,5	3,3	19.	48,1	49,0	0,9
5.	49,9	54,9	5,0	20.	26,0	28,2	2,2
6.	39,7	44,8	5,1	21.	37,4	37,9	0,5
7.	72,1	71,9	-0,2	22.	40,6	42,5	1,9
8.	48,3	47,8	-0,5	23.	55,5	62,2	6,7
9.	55,0	56,0	1,0	24.	65,1	66,9	1,8
10.	42,6	42,3	-0,3	25.	34,9	37,1	2,2
11.	61,0	60,1	-0,9	26.	59,6	59,4	-0,2
12.	39,9	43,7	3,8	27.	59,7	63,2	3,5
13.	45,3	51,6	6,3	28.	37,6	39,3	1,7
14.	46,9	51,0	4,1	29.	47,2	51,1	3,9
15.	35,3	41,2	5,9	30.	60,9	66,9	6,0

Mivel ugyanazok a személyek töltötték ki a tesztet, így párosított mintáról beszélhetünk. Első lépésként az elért pontszámok különbségét számítjuk ki egyénenként (a táblázatban a változás oszlopban látható). A pontszám változást hisztogram segítségével ábrázolva nem látunk a normalitástól való jelentős eltérést, illetve a mintaelemszámunk is viszonylag nagy, ezért alkalmazható (10.3):

$$51{,}307 - 48{,}753 \pm 2{,}756\frac{2{,}525}{\sqrt{30}} = 2{,}553 \pm 1{,}271$$

azaz 99%-os megbízhatóság mellett a sokasági változás (javulás) 1,282 és 3,824 pont közötti. A feladat megoldása során X változónak az előtte, Y változónak az utána pontszámot tekintettük. Fordított választás esetén az eredmény a mostani ellentettje, azaz $-2{,}553 \pm 1{,}271$ lenne, de szintén átlagos javulásként kellene értelmezni.

10.1.2. Független mintás becslés

Abban az esetben tehát, ha a két sokaságból egymástól függetlenül veszünk mintát, független mintás becslésről beszélünk. Követve a 9.2.1. fejezetben követett logikát elsőként a matematikailag legegyszerűbb, de a gyakorlatban kevéssé alkalmazható esettel kezdjük a tárgyalást, majd innen mozdulunk el a praktikusabb feltételezések felé. A fejezetben három különböző feltételezés mellett adunk (10.1) formulának konkrét alakot. Alapvetően mindhárom esetben a sokaságok normális eloszlását

tételezzük fel, ezt azonban a későbbiekben lazítjuk. A célunk tehát az ismeretlen $\delta = \mu_X - \mu_Y$ sokasági várható érték különbség becslése az alábbi feltételezésekkel:

- ismert σ_X^2 és σ_Y^2
- σ_X^2 és σ_Y^2 nem ismert, de egyezőségüket feltételezzük
- σ_X^2 és σ_Y^2 nem ismert, egyezőségüket nem tételezzük fel

Legyen tehát a két sokaság X és Y, melyekből nem feltétlenül azonos méretű, n_X és n_Y elemű mintákat veszünk. Ekkor könnyen belátható, hogy

$$Z = \frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0,1) \tag{10.4}$$

standard normális eloszlást követ. A fenti összefüggés alapján a (9.4) levezetéssel analóg módon kapjuk, hogy a minta alapján számított

$$\overline{x} - \overline{y} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \tag{10.5}$$

alsó és felső határok a δ ismeretlen sokasági átlag különbségre vonatkozó $1-\alpha$ megbízhatósági szintű konfidencia intervallumot adnak.

A gyakorlatban (10.5) alkalmazása nem túl gyakori, hiszen nem gyakran fordul elő, hogy a sokaságok normalitását és a sokasági varianciák ismeretét is feltételezhetjük, ezért itt példát sem hozunk. Amennyiben a választott minták nagyok, néhány tankönyvben mégis ez a formula szerepel, hiszen a nagyon nagy mintából már szinte biztosak lehetünk benne, hogy a mintabeli variancia jól közelíti a sokasági varianciát. Ebben a könyvben nem követjük ezt az elméletileg helytelen, gyakorlati szempontból viszont védhető gyakorlatot.

A második esetben a normalitás mellett már nem feltételezzük a sokasági varianciák ismeretét, azonban azok azonosságát igen, tehát $\sigma_X^2 = \sigma_Y^2$. Ez a feltételezés már gyakorlatiasabb, hiszen ismeretlen sokasági átlagok mellett igen ritka eset, hogy mindkét sokaság varianciáját ismerjük. Belátható, hogy ekkor (10.4) formula $\sigma^2 = \sigma_X^2 = \sigma_Y^2$ helyettesítéssel alkalmazható lenne, de a feltételezés szerint a közös σ^2 paraméter ismeretlen. Ahogy azt a korábbi fejezetekben már láttuk, ilyenkor a mintából származó becsléssel helyettesítjük a sokasági paramétert. Jelen esetben σ^2-re vonatkozóan mindkét minta tartalmaz információt, nem lenne jogos, ha pusztán az egyik mintabeli szórás alapján közelítenénk az értékét. A legjobb közelítést a két mintabeli korrigált variancia mintaelemszámmal[2] súlyozott átlagaként nyerhetjük. Az ismeretlen σ^2-re vonatkozó, mintán alapuló becslésünk neve mintabeli pooled (közös) variancia

[2]Pontosabban szabadságfokokkal, azaz mintaelemszám−1-gyel.

$$s_p^2 = \frac{(n_X - 1)\, s_X^2 + (n_Y - 1)\, s_Y^2}{n_X + n_Y - 2} \tag{10.6}$$

Ahogy azt már szintén megszokhattuk, a sokasági paraméter helyettesítése egy mintabeli becsült értékkel a mintavételi eloszlás normalitását felborít(hat)ja, ennek megfelelően az $1 - \alpha$ megbízhatóságú konfidencia intervallum δ-ra a sokaságok normalitását, valamint a sokasági varianciák egyezőségét feltételezve

$$\overline{x} - \overline{y} \pm {}_{n_X+n_Y-2}t_{1-\alpha/2} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \tag{10.7}$$

A formula szerkezete nagyon hasonlít (9.6) formuláéra, annak közvetlen általánosításaként tekinthető. A t eloszlás szabadságfoka $n_X - 1 + n_Y - 1$, vagy rövidebben $n_X + n_Y - 2$. A standard hiba helyén s_p és a mintaelemszámok szerepelnek.

A harmadik eset azt a szituációt vizsgálja, amikor a sokaságok normalitása mellett a sokasági varianciák egyezőségét nem tesszük fel. Ekkor (10.4)-ben sem σ_X^2, sem σ_Y^2 nem ismert. Ekkor nem a pooled mintabeli varianciával becsüljük ezek értékét, hanem külön-külön a megfelelő mintabeli varianciákkal. Belátható, hogy a normalitás ebben az esetben is sérül, a t eloszlás szabadságfoka pedig az alábbi formulával közelíthető adott minták mellett

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\left(\frac{s_X^2}{n_X}\right)^2 \frac{1}{n_X - 1} + \left(\frac{s_Y^2}{n_Y}\right)^2 \frac{1}{n_Y - 1}} \tag{10.8}$$

az $1 - \alpha$ megbízhatóságú konfidencia intervallum δ-ra a sokaságok normalitását, valamint a sokasági varianciák egyezőségét feltételezve

$$\overline{x} - \overline{y} \pm {}_{\nu}t_{1-\alpha/2} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} \tag{10.9}$$

ahol a t eloszlás ν szabadságfoka a (10.8) egyenletből adódik.

A (10.7) és a (10.9) formulák igazi jelentőségét az adja, hogy – az egymintás esethez hasonlóan – abban az esetben is alkalmazhatók, ha a sokaság eloszlása nem normális, de a minták mérete elegendően nagy. Ahogy azt a 9.2.1. fejezetben is tárgyaltuk, annak meghatározása, hogy mekkora az elegendően nagy, komoly matematikai-statisztikai feladat, a különböző szakirodalmak más-más értéket említenek. Jelen tananyagban a mindkét sokaságból 30 feletti mintaelemszámot már elegendően nagynak tekintjük. Amennyiben a minta alapján a sokaság jelentősen eltér a normális eloszlástól, ennél akár jóval nagyobb mintára is szükség lehet.

Egy felmérés a férfi és női ügyfelek pénzügyi tudatosságát vizsgálja, a közvetlenül a folyószámlán tartott összegen keresztül. Ez a forma ugyan likvid, azonban nem biztosít hozamot. A vizsgálat a bankszámlával rendelkező felnőtt lakosságból mint alapsokaságból rendelkezik egyszerű véletlen mintával. Becsüljük meg a férfiak és nők átlagos folyószámla egyenlegének különbségét 90%-os megbízhatóság mellett!

Jelölje X a férfiak sokasági egyenleg eloszlását, Y pedig a hölgyekét. A mintavétel során azt kaptuk, hogy a férfiak átlagos egyenlege 246 978 Ft, míg a nők esetén ez 332 739 Ft. A korrigált mintabeli szórás 132 843 Ft, illetve 141 434 Ft, ami 50 férfi és 40 női ügyfél megkérdezése alapján került kiszámításra.

Az egyenlegek eloszlása ugyan várhatóan nem normális, hanem valószínűleg jobboldali aszimmetriával rendelkezik (néhány extrém nagy egyenleg miatt), a viszonylag nagy mintaelemszám azonban lehetővé teszi, hogy (10.7), illetve (10.9) formulákat alkalmazzuk a sokasági varianciákra vonatkozó feltevésektől függően. A sokasági varianciák összehasonlítása a Statisztikai modellezés tárgy tananyaga, így feladatmegoldás során az alkalmazandó feltételezést explicit módon meg fogjuk adni. Ebben a példában illusztrációként mindkét esetet végigszámoljuk.

Tételezzük fel elsőként a sokasági varianciák egyezőségét, ami a mintabeli szórások hasonlósága miatt nem tűnik rossz feltételezésnek (természetesen a mintavétel miatt a mintabeli szórások akkor sem lesznek teljesen egyenlők, ha a sokasági szórások megegyeznek). Ekkor (10.7) alapján

$$\overline{x} - \overline{y} \pm {}_{n_X+n_Y-2}t_{1-\alpha/2} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} = 246\,978 - 332\,739 \pm 227\,272{,}1\sqrt{\frac{1}{50} + \frac{1}{40}}$$

ahol ${}_{88}t_{0,95} = 1{,}662$, illetve

$$s_p = \sqrt{\frac{(50-1)\,132\,843^2 + (40-1)\,141\,434^2}{50+40-2}} = 136\,717$$

Azaz a minta alapján számított konfidencia intervallum alsó határa $-133\,973$ Ft, felső határa $-37\,549$ Ft. Ez azt jelenti, hogy a sokasági különbségre vonatkozó becslésünk alapján a nők folyószáma egyenlege magasabb 90%-os megbízhatóság mellett legalább 37 549, legfeljebb 133 973 forinttal. Vegyük észre, hogy mivel a férfiak egyenlegét választottuk X és a nőkét Y változónak, a negatív előjelű becslés azt jelenti, hogy a nők egyenlege magasabb. Fordított választás esetén ugyancsak ellentett eredményt kapnánk.

Amennyiben a sokasági varianciák egyezőségét nem tételezzük fel, úgy (10.9) alapján

$$246\,978 - 332\,739 \pm {}_{81}t_{0,95}\sqrt{\frac{132\,843^2}{50} + \frac{141\,434^2}{40}}$$

ahol ${}_{81}t_{0,95} = 1{,}664$, illetve

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\left(\frac{s_X^2}{n_X}\right)^2 \frac{1}{n_X - 1} + \left(\frac{s_Y^2}{n_Y}\right)^2 \frac{1}{n_Y - 1}} = 81{,}26$$

Azaz a minta alapján számított konfidencia intervallum alsó határa $-134\,358$ Ft, felső határa $-37\,164$ Ft.

10.2. Sokasági arányok különbségének becslése

Gyakran nem két sokaság átlagát, hanem két sokaságban megfigyelhető arányt szeretnénk összehasonlítani. Az egymintás esethez hasonlóan megmutatható, hogy elegendően nagy minta esetén a

$$Z = \frac{(P_X - P_Y) - (\pi_X - \pi_Y)}{\sqrt{\frac{P_X(1-P_X)}{n_X} + \frac{P_Y(1-P_Y)}{n_Y}}} \sim \mathcal{N}(0,1) \tag{10.10}$$

valószínűségi változó standard normális eloszlást követ, ahol P_X és P_Y jelöli a mintabeli arányok valószínűségi változóját, π_X és π_Y a két sokasági arány, n_X és n_Y pedig a két mintaelemszám.

Ebből belátható, hogy $1 - \alpha$ megbízhatóságú konfidencia intervallum szerkeszthető a $\delta = \pi_X - \pi_Y$ sokasági arány különbségre

$$p_X - p_Y \pm z_{1-\alpha/2} \sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}} \tag{10.11}$$

módon, ahol p_x és p_y a mintából kiszámított arányt jelölik.

A termelési igazgató két beszállítótól származó termékek minőségét szeretné összehasonlítani, mintavételes módszerrel. Az A gyártó szállítmánya esetén a kivett 200 termékből 23 volt problémás, míg B gyártó esetén a kiválasztott 150 termékből 21 volt kifogásolható. Végezzünk 95%-os megbízhatóságú konfidencia intervallum becslést a sokasági arányok A és B gyártó közti különbségére vonatkozóan!

$$p_X - p_Y \pm z_{1-\alpha/2} \sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}} = 0{,}115 - 0{,}14 \pm 0{,}071$$

Azaz a 95%-os megbízhatóságú konfidencia intervallum alsó határa $-0{,}096$ és felső határa 0,046. Az eredmény alapján nem tudjuk eldönteni egyértelműen, hogy melyik sokasági arány a nagyobb, hiszen a különbség negatív és pozitív is lehet.

10.3. Excel tippek

Hasznos Excel függvények:

- SZÓR.M
- NORM.S.INVERZ, T.INVERZ (z érték és t érték meghatározásához)

A (10.8) egyenletből származó szabadságfokra jellemzően nem egész érték adódik. Az Excel azonban csak pozitív egész szabadságfokokra értelmezi a t eloszlást, amennyiben nem egész számot adunk meg, automatikusan lefelé kerekíti azt, ami konzervatív (nagyobb) t értéket ad. Egyéb szoftverekben a keresett érték pontosabban is megtalálható.