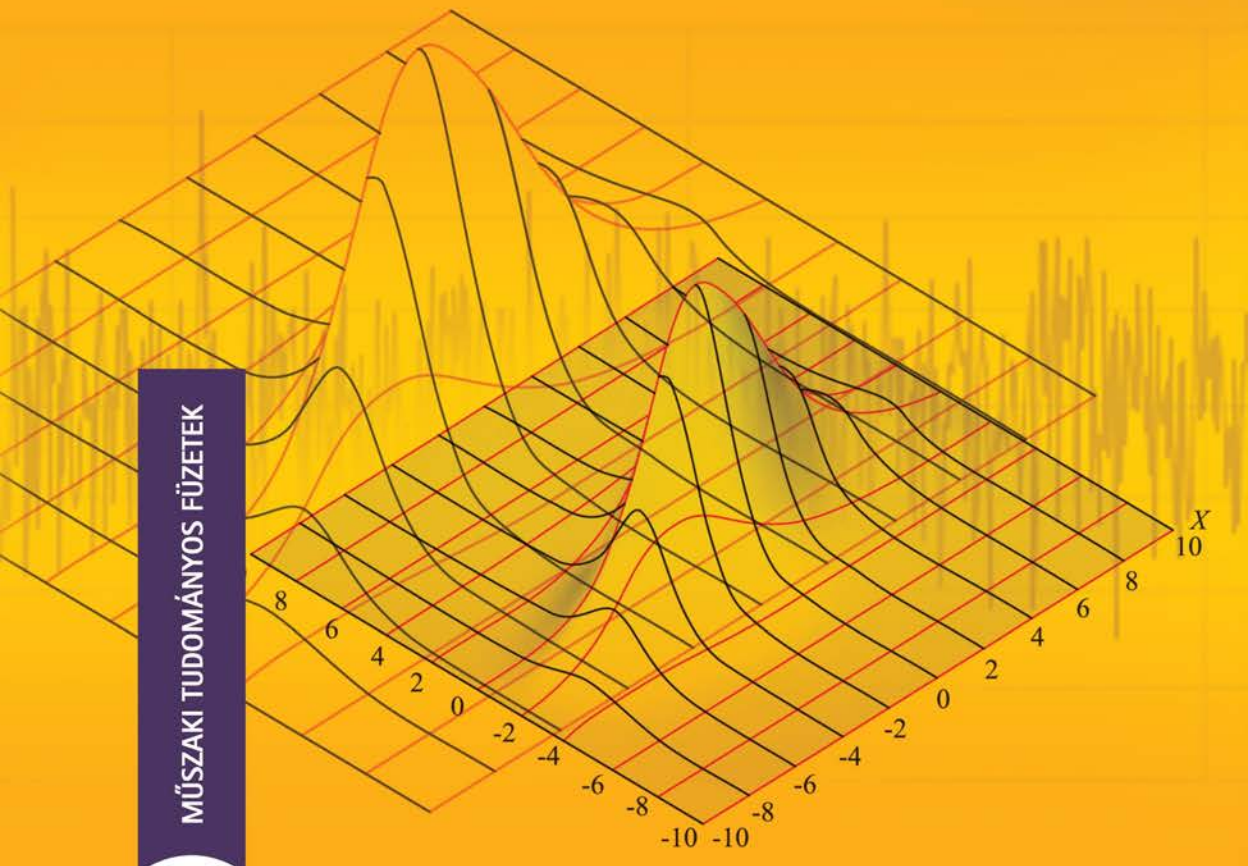
HARANGUS KATALIN – KAKUCS ANDRÁS

VALÓSZÍNŰSÉGSZÁMÍTÁS ÉS STATISZTIKA A MÉRNÖKI GYAKORLATBAN



MŰSZAKI TUDOMÁNYOS FÜZETEK

ERDÉLYI MÚZEUM-EGYESÜLET

EME

HARANGUS KATALIN – KAKUCS ANDRÁS

VALÓSZÍNŰSÉGSZÁMÍTÁS ÉS STATISZTIKA A MÉRNÖKI GYAKORLATBAN

EME

MŰSZAKI TUDOMÁNYOS FÜZETEK

15.

ISSN 2068 – 3081

EME

MŰSZAKI TUDOMÁNYOS FÜZETEK

15.

HARANGUS KATALIN – KAKUCS ANDRÁS

VALÓSZÍNŰSÉGSZÁMÍTÁS ÉS STATISZTIKA A MÉRNÖKI GYAKORLATBAN

ERDÉLYI MÚZEUM-EGYESÜLET

Kolozsvár

2021

EME

A kötet megjelenését támogatta a Nemzeti Kulturális Alap,
a Bethlen Gábor Alapkezelő Zrt., a Magyar Tudományos Akadémia és
az EME Műszaki Tudományok Szakosztálya.

Szaklektor: Máté Márton, Pokorádi László
Kiadja: az Erdélyi Múzeum-Egyesület
Felelős kiadó: Biró Annamária
Sorozatszerkesztő: Bitay Enikő
Olvasószerkesztő: András Zselyke
Borítóterv: Könczey Elemér
Műszaki szerkesztő: Kakucs András

Nyomdai munkálatok:
F&F International Kft. Kiadó és Nyomda, Gyergyószentmiklós
Ügyvezető igazgató: Ambrus Enikő
Tel./Fax: +40-266-364171

Descrierea CIP a Bibliotecii Naţionale a României
HARANGUS, KATALIN
Valószínűségszámítás és statisztika a mérnöki gyakorlatban / Harangus Katalin, Kakucs András. - Cluj-Napoca : Erdélyi Múzeum-Egyesület, 2021
Conţine bibliografie
ISBN 978-606-739-181-7

I. Kakucs, András

51

DOI: 10.36242/mtf-15

TARTALOMJEGYZÉK

ELŐSZÓ

Könyvünkben a valószínűségszámítás és a statisztika elemeivel ismerkedhetik meg az Olvasó. Az elsődleges cél a témakör alapvető fogalmainak a tisztázása, amelyet az alkalmazásuk szemszögéből és nem annyira a matematikai oldalról közelítünk meg. Ekképpen tehát a könyvben bemutatottak az alkalmazásszintű ismeretek megalapozását segítik elő, de ugyanakkor kiindulási pontot jelenthetnek a mélyebb elméleti fejtegetésekhez is.

A könyv elsősorban, de nem csak a mérnöki szakokat hallgató diákok számára készült, és feltételezi bizonyos matematikai alapismeretek meglétét.

Az alkalmazásokat az elterjedten használt Microsoft Office-hoz tartozó Excel-táblázat-kezelő programmal oldottuk meg. Az ismertetett példák az Excel 14.0-es vagy annál újabb verzióival kompatibilisek (tehát legalább 2010-es verziót kell használni).

Az alkalmazás csak a beépített statisztikai függvények egy részére korlátozódik, az olvasóra marad a többi függvény és az „Analysis ToolPak”-bővítmény felfedezése, mely utóbbi néhány esetben pl. a statisztikai tesztek könnyebb elvégzésében nyújthat segítséget. Az Excelben még így sem lelhető fel mindig az éppen szükséges eszköz vagy függvény: ha az előttünk álló feladatot készen alkalmazható eszközökkel szeretnénk megoldani, akkor ezeket külső kiegészítőkkel, pl. a „Real Statistic Resource Pack” (amely szabadon hozzáférhető) vagy az „XLSTAT” (amelyhez évi bérletet kell fizetni) telepítésével remélhetjük megoldani.

A könyvben az angol nyelvű Excel függvényeire hivatkozunk, de a könyv végén megtalálhatjuk a magyar nyelvű megfelelőket is.

Mivel a Microsoft Office használata jogokhoz kötött, használhatjuk az OpenOffice táblázatkezelő programját is, a „Calc”-ot. A két program között léteznek bizonyos különbségek, a statisztikai függvényeket azonban aránylag könnyen megfeleltethetjük.

1. ALEA IACTA EST…

Kísérletek, események

Adott jelenség előre meghatározott körülmények közötti elindítását és végrehajtását *kísérletnek* nevezzük. A kísérlet *körülményeit*, vagyis a jelenség lefolyását meghatározó tényezők minőségi és mennyiségi definícióját *kísérleti beállításnak* nevezzük. A kísérlet lehetséges eredményeit a kísérlet *elemi esemény*einek vagy kimeneteinek nevezzük. Az összes lehetséges eredmény, azaz elemi esemény halmaza az *eseménytér*.

A lehetséges eredmények száma lehet véges vagy akár végtelen. Ha például a kísérlet tárgya a címben is szereplő játékkockadobás, akkor a lehetséges eredmények (elemi események) száma véges: a kockának az 1–6 ponttal ellátott oldalainak egyike kerül felül, az eseményteret hat elemi esemény alkotja (kizárjuk az olyan események bekövetkezését, hogy a kocka valamelyik élén vagy sarkán álljon meg). Ha viszont puskával lövöldözünk egy kellőképpen nagy céltáblára, akkor a becsapódó golyó helye a táblán bárhol lehet. Az elemi esemény az, hogy a céltáblát egy adott pontban találat éri. A tábla (még ha véges kiterjedésű is) a kiterjedés nélküli pontok végtelen halmaza, tehát ez esetben az eseménytér végtelen sok elemi eseményből áll.

Véletlen és determinisztikus események

Bizonyos esetekben a kísérlet eredménye, vagyis hogy a lehetséges elemi események közül melyik következik be, a tanulmányozott jelenségre vonatkozó ismereteink alapján előre megmondható; ezeket *determinisztikus* jelenségeknek nevezzük. Más esetben a jelenségre vonatkozó ismereteink hiányosak, vagy pedig a jelenség lezajlását követhetetlenül sok paraméter befolyásolja, és ilyenkor a kísérlet eredményét nem lehet előre pontosan megmondani: ezeket az eredményeket *véletlen* elemi eseményeknek hívjuk.

Elemi események és összetett események

Az eseményeket vizsgálódásainknak megfelelően többféleképpen is megadhatjuk. Legyen a kísérlet a játékkocka dobása, az elemi eredmények halmaza pedig valamelyik szám megjelenése: ilyenkor hat lehetséges eredményünk van. Ugyanebben a kísérletben más szempontból is figyelhetjük az eredményeket. Van, amikor nemcsak az elemi események bekövetkezését vizsgáljuk, hanem több kiválasztott elemi esemény bekövetkezésének

valamelyikét, azaz bizonyos *összetett események*ét. Példánkban például azt, hogy páros vagy páratlan számot dobunk-e. Ekkor a kísérlet elvégzése után továbbra is az 1–6 számok jelennek meg, de azoknak csak a páros-páratlan minőségét vizsgáljuk, és így két összetett esemény valamelyikének (azt, hogy 2, 4 vagy 6, illetve 1, 3 vagy 5 a kapott eredmény) bekövetkezését várhatjuk.

A biztos esemény és a lehetetlen esemény

Vannak olyan eredmények, események, amelyek a kísérlet bármely megismétlésénél megjelennek, ezeket *biztos* eseményeknek nevezzük. A kockadobásnál biztos, hogy az eredmény 1 és 6 között lesz (beleértve az egyest és hatost is). A céltáblára való lövöldözésnél az eseményteret a céltábla felülete jelképezi: ha az kellőképpen nagy, és megfelelőképpen pontosan célzunk, akkor a találat biztosan annak valamely pontjában következik be.

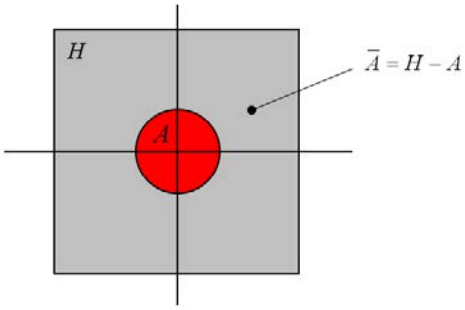
Vannak olyan eredmények, amelyek a kísérlet akárhány megismétlése után sem jelenhetnek meg, azokat *lehetetlen* eseményeknek nevezzük. A kockadobásnál sohasem jelenhet meg hetes vagy nyolcas (ha szokványos kockáról van szó), és a céllövésnél a fenti feltételek mellett sohasem megy el a golyó a tábla mellett.

Az ellentétes esemény

Valamely esemény bekövetkezése egyéb események megjelenését kizárhatja vagy sem. Például a kockadobás esetében, ha az események a kettővel osztható és a hárommal osztható számok megjelenésére vonatkoznak, akkor hatos dobás esetén mindkét említett esemény egyszerre következik be. Más esetben valamely esemény bekövetkezése kizárja egy másik esemény bekövetkezését, például ha páros számot dobtunk, akkor az biztosan nem páratlan. Az esemény be nem következésének az eseménye az előbbi *ellentétes esemény*e, másképpen *komplementer esemény*e vagy *ellentett*je. Ha a várt esemény hatos dobása, akkor annak ellentétes eseménye az egyes, kettes, ..., ötös dobásának elemi eseményeiből összetett esemény lesz. Ha a céllövésnél az esemény a céltábla közepét jelölő körrel elhatárolt terület eltalálása, akkor az azzal ellentétes eseményt a lövedék bármely más pontba történő becsapódása jelenti (tehát az, hogy nem találtuk el a tábla közepét).

Az eseményeket különböző módon, például betűkkel jelölhetjük (A, B, C stb.). Az ellentétes (*komplementer*) esemény jelölése felülvonással, a tagadás jelével történik: A ellentétes eseménye $\overline{A}$. A céllövés eseményeit grafikusan is szemléltethetjük: legyen H az eseménytér, vagyis a céltábla felülete. Jelölje A a

céltábla közepét jelentő kör pontjainak halmazát, ekkor $\bar{A}$ a körön kívüli pontok halmaza lesz (1.1. ábra).

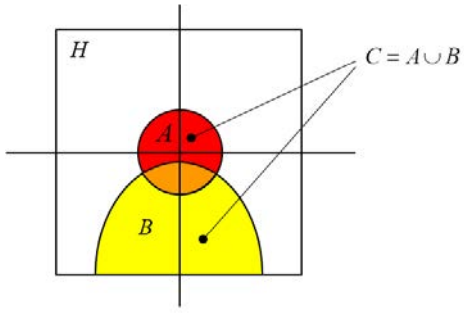


1.1. ábra. *Eseménytér, esemény és ellentétes (komplementer) esemény: ha nem találtuk el a kört, akkor mellélőttünk*

Egyesített esemény

Ha két esemény (pl. A és B) valamelyikének bekövetkezésére várunk, akkor a várt eredményt az két esemény *egyesített esemény*ének vagy *unió*jának nevezzük. Például ha a várt esemény „egyes vagy hatos dobása", ahol az egyes dobását A-val, a hatos dobást pedig B-vel jelöljük, akkor az „egyes vagy hatos dobás" eseménye $C = A \cup B$, ami A és B egyesített eseménye. Az „A vagy B" eseményt a kettő *összeg*ének is nevezik és akképpen is jelölik: $C = A + B$.

A céllövésnél jelölje B az alsó, a céltábla közepét jelentő kört részben takaró idomot alkotó pontok halmazát. Tegyük fel, hogy a lövés jónak számít, ha e két idom valamelyikébe beletalálunk. Ennek bekövetkeztét az jelenti, amikor a két idom pontjainak $C = A \cup B$ egyesített halmazában van a találat (1.2. ábra).

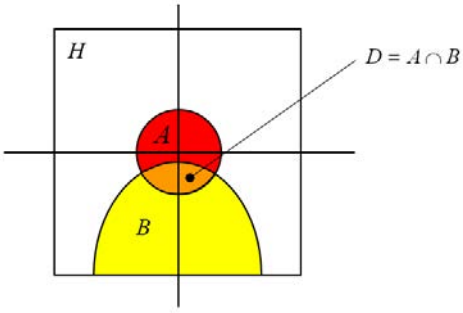


1.2. ábra. *Egyesített események (események uniója): a lövés jónak számít, ha eltaláltuk a kört vagy az alatta levő, azt részben átfedő idomot*

Események keresztmetszete

Ha két esemény együttes bekövetkezését várjuk, akkor a várt eredményt a két esemény *keresztmetszeté*nek (vagy *metszeté*nek) nevezzük. Ilyen a kockadobási kísérlet vizsgálatakor már említett kettővel (az A esemény) és hárommal osztható számok (a B esemény) figyelésének esetében a hatos megjelenése: $C = A \cap B$, ami a kettő keresztmetszete. Az „A és B” eseményt a kettő *szorzatá*nak is nevezik, másik jelölése: $C = A \cdot B$.

A céllövésnél tegyük fel, hogy a kitűnő eredményt az jelenti, amikor az A és B idomok egymást fedő részébe, a $D = A \cap B$ keresztmetszetébe esik a találat (1.3. ábra).

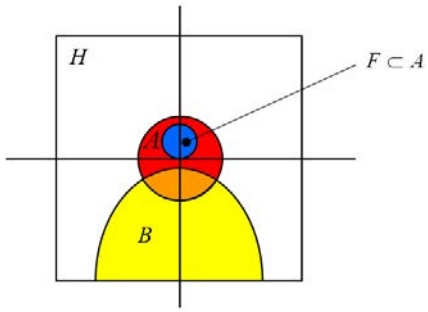


1.3. ábra. *Események keresztmetszete:*
kitűnő eredménynek számít, ha a két idom egymást fedő részét találjuk el

Események implikációja

Ha A egy olyan esemény, amelynek bekövetkezése egyúttal a B esemény bekövetkezését is jelenti, akkor azt mondjuk, hogy A bekövetkezése *maga után vonja* (implikálja) B bekövetkezését, és ezt a tényt $A \subset B$ módon jelöljük. Az $A \subseteq B$ jelölés a két esemény esetleges egyenlőségét is jelenti. Tegyük fel, hogy a kockadobálásnál az A esemény például a megjelenő 2-es szám, B pedig az, hogy a megjelenő szám páros. Ez azt jelenti, hogy A bekövetkezése (az, hogy kettest dobtunk) maga után vonja B bekövetkezését is (mivel páros számot dobtunk).

Céltáblás példánknál a célkör belsejében jelöljünk ki egy szűkebb kört, amelyet jelöljünk F -fel. Ha ebbe beletalálunk, akkor biztosan beletaláltunk az A-val jelölt nagyobb körbe is, tehát $F \subset A$. Geometriai interpretálása szerint az F kör pontjainak halmaza az A kör pontjainak egy részhalmaza (1.4. ábra).

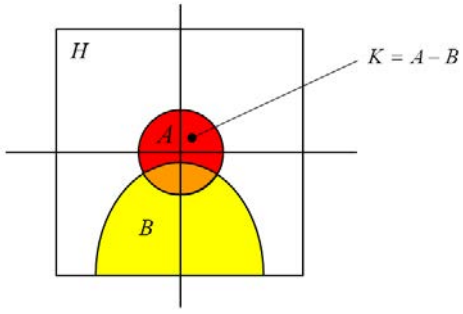


1.4. ábra. *Események implikációja:*
ha eltaláltuk a kisebb kört, akkor a találat biztosan a nagyobb kör belsejére esik

Ugyanígy $A \subset H$, $B \subset H$, $D \subset A$ és $D \subset B$, és így tovább. Ugyanitt észrevehetjük, hogy ha $F \subset A$, akkor az $F \supset A$ esemény abban és csak abban az esetben lehet igaz, ha a két halmaz egyenlő: $A = F$.

Események különbsége

Két esemény, A és B *különbségén* azt az eseményt értjük, amikor az A esemény bekövetkezik, de B nem: $G = A - B$. Ha a kockadobásnál az A esemény páros szám dobása, B pedig hatos dobása, akkor a kettő különbsége kettes vagy négyes dobása lesz: ezek ugyan páros számok, akárcsak a hatos (tehát A bekövetkezik, ami hatos dobásnál is bekövetkezne), de mivel nem hatost dobtunk, a B nem következett be. A különbség idegen eredetű szóval történő megnevezése: „*differencia*”.



1.5. ábra. *Események különbsége:*
a kör felső részét – és nem az alsót – találtuk el, ha a találat nem a nagyobb idom belsejére esik

Céltáblás példánknál az események különbségét már az első ábrán használtuk, ugyanis az A esemény bekövetkezésének $\overline{A}$ ellentettjét a H halmaz (az

eseménytért jelentő pontok) és az A halmaz különbsége jelenti. Egyébként e különbség az A halmaz H-ra vonatkoztatott komplementer halmazának definíciója. Továbbá, a második ábrán A és B különbségét a felső félhold alakú terület jelenti (vagyis hogy a találat oda esik, 1.5. ábra).

Az események keresztmetszetének (szorzatának) segítségével két esemény különbsége két esemény szorzataként is megadható: $A-B=A\cap\bar{B}$, tehát a különbségképzés szigorúan véve nem elemi művelet.

Összetett események

Az elemi műveletekkel további összetett eseményeket is definiálhatunk. Egy tétel szerint minden olyan esemény, amely nem lehetetlen és nem elemi esemény, egyértelműen felírható meghatározott elemi események összegeként.

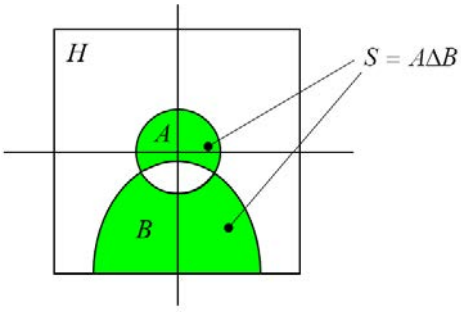
Ilyen összetett eseményt takar például a magyar nyelvű szakirodalomban ritkábban előforduló *szimmetrikus különbség* (*szimmetrikus differencia*) fogalma: két esemény szimmetrikus különbségén azt az eseményt értjük, amelyik nem tartozik a két esemény keresztmetszetébe:

$$A\Delta B=(A-B)\cup(B-A)=(A\cup B)-(A\cap B)\,. \qquad (1.1)$$

A logikában ez a „kizáró vagy” műveletnek felel meg, és az operátora „$\oplus$”.

Ha a kockadobásnál A a páros szám megjelenésének eseménye, B pedig a hárommal oszthatóé, akkor a két esemény szimmetrikus különbsége a kettővel vagy hárommal osztható, de nem a kettővel is és hárommal is osztható számok megjelenése (vagyis az lehet 2, 3 és 4).

A céllövésnél az 1.6. ábrán az $A\Delta B$ eseményt a színes idomok egymást át nem fedő részére eső találat jelenti.



1.6. ábra. *Események szimmetrikus különbsége:*
a találat a kör vagy a nagyobb idom belsejére esik, de nem a kettő által átfedett területre

Az eseményekkel végzett műveletek tulajdonságai

Fontos ismernünk az említett műveletek tulajdonságait, így az események összeadása és szorzása, akárcsak a számok összeadása és szorzása, kommutatív:

$$A \cup B = B \cup A, \text{ illetve } A \cap B = B \cap A; \qquad (1.2)$$

és asszociatív:

$$A \cup (B \cup C) = (A \cup B) \cup C, \text{ illetve } A \cap (B \cap C) = (A \cap B) \cap C; \qquad (1.3)$$

a szorzás pedig disztributív az összeadásra nézve:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C); \qquad (1.4)$$

a $\varnothing$ lehetetlen esemény az összeadásra vonatkozólag semleges elem:

$$A \cup \varnothing = A; \qquad (1.5)$$

és a szorzásra nézve zéruselem:

$$A \cap \varnothing = \varnothing; \qquad (1.6)$$

az E biztos esemény pedig a szorzásra vonatkozólag egységelem.

$$A \cap E = A. \qquad (1.7)$$

A számokkal végzett algebrai műveletek tulajdonságain túlmenően az eseményekkel végzett műveleteknek van egy pár sajátságos tulajdonságuk, így az események összeadása és szorzása idempotens művelet:

$$A \cup A = A, \text{ illetve } A \cap A = A; \qquad (1.8)$$

az összeadás pedig disztributív a szorzásra nézve:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \qquad (1.9)$$

valamint

$$A \cup E = E. \qquad (1.10)$$

A komplementer esemény meghatározásából következően:

$$\bar{\bar{A}} = A, \; A \cup \bar{A} = E, \; A \cap \bar{A} = \varnothing. \qquad (1.11)$$

A fentiek alapján egyéb tulajdonságokat is előállíthatunk. Nevezetesek a De Morgan-féle egyenletek:

$$\overline{A \cup B} = \bar{A} \cap \bar{B} \text{ és } \overline{A \cap B} = \bar{A} \cup \bar{B}; \qquad (1.12)$$

gyakran használt képletek pedig a műveletek disztributivitásából eredő

$$A \cup (A \cap B) = A \text{ és } A \cap (A \cup B) = A \qquad (1.13)$$

egyenlőségek, valamint az összeadandókra bontás:

$$A = (A \cap B) \cup (A \cap \bar{B}). \qquad (1.14)$$

Az A és a B eseményeket *egymást kizáró események*nek nevezzük, ha $A \cap B = \varnothing$. Észrevehetjük, hogy az előbbi összeadandókra bontásnál az összeg két tagja egymást kizáró események.

A kivonásra és az implikációra vonatkozóan könnyen beláthatjuk, hogy azok nem kommutatív műveletek. Az is belátható, hogy mivel $A \cup \bar{A} = E$,

$$E - A = \bar{A}. \qquad (1.15)$$

Mint mondtuk, az események kivonását szorzattá is alakíthatjuk: $A - B = A \cap \bar{B}$. Ha itt A az E biztos esemény és B a $\varnothing$ lehetetlen esemény, akkor e szorzattá alakításból következik, hogy

$$E - \varnothing = E = \bar{\varnothing}. \qquad (1.16)$$

Hasonlóképpen, ha A is, és B is biztos esemény, akkor

$$E - E = \varnothing = \bar{E}. \qquad (1.17)$$

Következésképpen a biztos esemény és a lehetetlen esemény egymás komplementerei, fordítottjai.

A halmazokra vonatkozó ismereteink alapján felírható még:

$$A \cap B \subset A \text{ és } A \cap B \subset B \qquad (1.18)$$

(ezek egyidejűleg érvényesek).

Teljes eseményrendszer

Az $A_1, A_2, \ldots A_n$ események *teljes eseményrendszer*t alkotnak, ha azok páronként egymást kizáró események (azaz bármely $i \neq j$-re $A_i \cap A_j = \varnothing$), és egyesítésük a biztos eseményt eredményezi (azaz $A_1 \cup A_2 \cup \ldots \cup A_n = \bigcup_{i=1}^{n} A = E$). A teljes eseményrendszert alkotó események száma végtelen nagy is lehet.

Ezek alapján az A esemény és annak ellentétese $\bar{A}$, egymást kiegészítve, egy teljes eseményrendszert alkotnak (ezért nevezik $\bar{A}$-t komplementer eseménynek).

I. Excel-szimuláció

A kockadobás kísérletét számítógéppel, a Microsoft Excel táblázatkezelő programmal könnyen szimulálhatjuk:

1. Jelöljünk ki két cellát, amelyek a bemeneti adatokat fogják tartalmazni. Ezek:

– az előforduló legkisebb érték (kockadobás esetében ez 1), az *A2*-es cellában;

– az előforduló legnagyobb érték (kockadobás esetében ez 6), az *A5*-ös cellában.

2. A *B2* cellában hozzunk létre egy x véletlenszámot, a megadott legkisebb és legnagyobb értékek között. Ez lesz a kockadobás eredménye.

A véletlenszámokat a RAND() függvénnyel lehet generálni. Az Excel a véletlenszámokat egy bizonyos numerikus eljárás szerint hozza létre. Az így nyert valós számok a $[0,1)$ intervallumon egyenletesen oszlanak meg, tehát elméletileg azonos valószínűséggel előforduló számok. Ezek a számok nem teljesen véletlenek, csak véletlenszerű, „pszeudovéletlen" számok sorozatából valók, azonban ez a sorozat annyira hosszú, hogy gyakorlatilag nem fogunk ismétlődő értékeket kapni.

Ha a megadott legkisebb és legnagyobb értékek közé eső véletlen egész számokat akarunk létrehozni, a *B2*-es cellába a következő függvényt kell beírnunk:

$$=\$A\$2+INT((1+\$A\$5-\$A\$2)*RAND()), \qquad (1.19)$$

miszerint oda az

$$x = min + \mathrm{Int}\left[(1+max-min)\cdot rnd\right] \qquad (1.20)$$

egész szám kerül. Mivel az rnd pszeudo-véletlen szám a $[0,1)$ intervallumon vehet fel értékeket, az egészrész-függvény argumentuma a $[0,1+max-min)$ intervallumra korlátozódik. Így az $\mathrm{Int}()$ függvény által visszatérített érték egy 0 és $max-min$ között levő egész szám lesz, amelyhez ha hozzáadjuk a min értéket, akkor egy min és max közé eső egész véletlenszámot kapunk, beleértve a két határértéket is.

3. Definiáljunk két eseményt:

– a *D1* cellában „ *A* " – a kapott véletlen szám páros:

$$=IF(MOD(\$B\$2,2)=0,TRUE,FALSE); \qquad (1.21)$$

– a *D2* cellában „ *B* " – a kapott véletlen szám hárommal osztható;

$$=IF(MOD(\$B\$2,3)=0,TRUE,FALSE). \qquad (1.22)$$

4. E két eseménnyel adjunk meg továbbiakat, például:

– a *D4* cellában „ *A* " komplementer eseményét (a véletlen szám páratlan):

=NOT(D1); (*1.23*)

– a *D5* cellában „*A*” és „*B*” egyesített eseményét (a véletlen szám páros vagy hárommal osztható):

=OR(D1,D2); (*1.24*)

– a *D6* cellában „*A*” és „*B*” keresztmetszetét (a véletlen szám páros és hárommal osztható):

=AND(D1,D2); (*1.25*)

– a *D7* cellában „*A*” és „*B*” különbségét (a véletlen szám páros, de nem osztható hárommal):

=AND(D1,NOT(D2)); (*1.26*)

és így tovább folytathatjuk tetszés szerint (1.7. ábra).

5. Ha új számot akarunk „dobni”, a számítógép billentyűzetén nyomjuk meg az F9 gombot: ekkor az Excel frissíti a cellák tartalmát, és minden frissítéskor a RAND() függvény egy újabb értéket ad vissza.

	A	B	C	D
1	**Legkisebb érték**	***x***	"A" esemény: páros szám	FALSE
2	1	1	"B" esemény: osztható hárommal	FALSE
3				
4	**Legnagyobb érték**		"A" komplementer eseménye: páratlan szám	TRUE
5	6		"A" és "B" egyesített eseménye: páros vagy osztható hárommal	FALSE
6			"A" és "B" keresztmetszete: páros és osztható hárommal	FALSE
7			"A" és "B" különbsége: páros, de nem osztható hárommal	FALSE
8				
9				

1.7. ábra. *Kockadobás szimulálása Excelben*

Kombinatorika

A valószínűségszámításhoz szükséges lehet események, elemek véges halmazának valamilyen szabály szerinti rendezésére, például olyan kérdések megválaszolására is, hogy n egymástól különböző elem hányféle egymástól eltérő módon rendezhető sorba. Ezek a számítások a kombinatorika tárgyköréhez tartoznak. Dióhéjban:

n egymástól különböző elem sorba állításakor a lehetséges egymástól különböző elrendezések, *ismétlés nélküli permutációk* száma

$$P_n = n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1, \quad (1.27)$$

amit általában csak permutáció néven említünk (érdemes megemlíteni, hogy nulla elem permutációinak száma $0!=1$). Például az ábécé első három betűjét hat különböző módon lehet sorrendbe állítani ($P_3 = 3! = 6$):

abc–acb–bac–bca–cab–cba.

Ha a sorba állítandó n elem között k egyforma, akkor az egyforma elemek felcserélésével nem jutunk újabb elrendezéshez. Az egymástól különböző elrendezések, *ismétléses permutációk* számát a

$$P_n^k = \frac{n!}{k!} \qquad (1.28)$$

hányados adja. Például három betűt, amelyből kettő azonos, háromféle módon lehet elrendezni ($P_3^2 = 3!/2! = 3$):

abb–bab–bba.

Ha az előbbi, ismétlés nélküli permutációk sorozatában *c*-t *b*-vel helyettesítettük volna, akkor a hat elrendezés között három ismétlődést láthatnánk.

Ha az n elem mind egyforma, akkor azokat csak egyféle módon állíthatjuk sorba.

Ha az n elem között egymástól eltérő elemek ismétlődését tapasztaljuk (például az egyik fajtából k_1, egy másikból pedig k_2 darab van a halmazban, és így tovább, $n = k_1 + k_2 + \ldots$), akkor az ismétlődéses permutációk képletét a következő módon általánosíthatjuk:

$$P_n^{k_1,k_2\ldots} = \frac{n!}{k_1! \cdot k_2! \cdot \ldots}. \qquad (1.29)$$

Amikor n egymástól különböző elem halmazából kiragadunk k elemet, az egymástól eltérő lehetőségek száma az *ismétlés nélküli kombinációk* száma adja:

$$C_n^k = \frac{n!}{k! \cdot (n-k)!}, \qquad (1.30)$$

amit általában csak kombináció néven említünk. A k elem sorrendje itt nem számít. Amennyiben $k = n$, csak egy lehetőségünk van, mert a kiragadott részhalmaz a forráshalmazzal azonos. Ha a kiragadott részhalmaz kisebb, akkor a lehetőségek száma növekedik. Például az ábécé első három betűjéből három, egymástól különböző, kétbetűs kombinációt tudunk alkotni ($C_3^2 = 3!/[2! \cdot (3-2)!] = 3$):

ab–ac–bc.

Amikor az n egymástól különböző elem halmazából úgy ragadunk ki k elemet, hogy megengedjük ugyanannak a kiragadott elemnek a többszöri előfordulását is, a lehetőségeket az *ismétléses kombinációk* száma adja:

$$C_n^{k,i} = \frac{(n+k-1)!}{k!\cdot(n-1)!}. \qquad (1.31)$$

Ez az eset a visszatevéses mintavételezésnél fordul elő: az egymástól különböző minták halmazából a kombináció képzése során ugyanazt az elemet többször is kivehetjük (és a kivétel után vissza is tesszük az eredeti halmazba). Például az ábécé első három betűjéből hat egymástól különböző, de ismétléseket is tartalmazó, kétbetűs kombinációt tudunk alkotni ($C_3^{2,i} = (3+2-1)!/[2!\cdot(3-1)!] = 6$):

aa–ab–ac–bb–bc–cc.

Amikor n egymástól különböző elem halmazából kiragadunk k elemet, a kiragadott elemek sorrendjét is tekintetbe vevő, egymástól eltérő lehetőségek száma az *ismétlés nélküli variációk* száma adja:

$$V_n^k = \frac{n!}{(n-k)!}, \qquad (1.32)$$

amit általában csak variáció néven említünk. Itt tehát a kombinációtól eltérően a k elem sorrendje is számít. Amint az 1.20. és az 1.29. képletek összevetéséből is kiderül, a variációk száma a lehetséges kombinációk permutációinak összes számával egyenlő, tehát

$$V_n^k = C_n^k \cdot P_n. \qquad (1.33)$$

Az ábécé első három betűjéből alkotott kétbetűs variációk száma hat (a képlet alkalmazásával $V_3^2 = 3!/(3-2)! = 6$):

ab–ba–ac–ca–bc–cb.

Amikor az n egymástól különböző elem halmazából úgy ragadunk ki k elemet, hogy megengedjük ugyanannak a kiragadott elemnek a többszöri előfordulását is, az *ismétléses variációk* száma:

$$V_n^{k,i} = n^k. \qquad (1.34)$$

Ez az eset is a visszatevéses mintavételezésnél fordul elő, az ismétléses kombinációtól eltérően a kiragadott elemek ismétléses permutációit is figyelembe

veszi. Az ábécé első három betűjéből kilenc egymástól különböző, de ismétléseket is tartalmazó kétbetűs variációt tudunk alkotni:

aa–ab–ba–ac–ca–bb–bc–cb–cc.

Amikor az n elem halmazából k elemet kiragadva ismétlés nélküli kombinációkat vagy variációkat képezünk, és az n elem között ismétlődők is vannak, akkor a lehetőségek száma csökken. A lehetséges kombinációk és variációk számát olyan módon tudjuk kiszámítani, hogy az n elem közül eltávolítjuk az ismétlődéseket. Ilyenkor a nem ismétlődő elemek száma nem lehet kevesebb a kombinációt, illetve a variációt alkotó elemek k számánál. Például ha az ismétlődést tartalmazó *abbc* sorból szeretnénk ismétlődés nélküli kombinációkat vagy variációkat alkotni, akkor először el kell távolítanunk a megismételt *b* betűt.

Ha dobókockákkal akarjuk példázni mindezt, akkor vegyünk hat kockát:

– ha minden kockán más szám van, akkor a kockákat az asztalon

$$P_6 = 6! = 720 \qquad (1.35)$$

-féleképpen tudjuk sorba rendezni;

– ha két kockán ugyanaz a szám fordul elő, de a többi egymástól különböző, akkor a lehetséges elrendezések száma csak

$$P_6^2 = \frac{6!}{2!} = 320\,; \qquad (1.36)$$

– ha csak három kockán előforduló különböző számokat figyelünk, összesen

$$C_6^3 = \frac{6!}{3!\cdot(6-3)!} = 20 \qquad (1.37)$$

társításban jelenhetnek meg;

– ha az iménti esetben a sorrendet is számításba vesszük, akkor összesen

$$V_6^3 = \frac{6!}{(6-3)!} = 120 \qquad (1.38)$$

elrendezést állíthatunk össze;

– ha a megfigyelt három kockán a számok többszöri megjelenését is megengedjük, akkor a lehetséges társítások száma

$$C_6^{3,i} = \frac{(6+3-1)!}{3!\cdot(6-1)!} = 56 \qquad (1.39)$$

(három kocka dobásának 56 lehetséges kimenete van);

– az utóbbiakat pedig

$$V_6^{3,i} = 6^3 = 216 \qquad (1.40)$$

-féle módon tudjuk sorba állítani (ha a három kockát egymás után gurítjuk el, akkor 216 lehetséges elrendezésben jelennek meg a számok).

Binomiális együtthatók

Az $(1+x)$ binom n-edik hatványa egy n-ed-rendű polinom. E polinom x^k tagjának együtthatóját többféleképpen is fel lehet írni, az egyik lehetőség éppen az ismétlés nélküli kombinációk képletével azonos. A valószínűségszámításban gyakran fordul elő a kombinációnak binomiális együtthatóként történő felírása. A binomiális együtthatók egyezményes jelölése

$$\binom{n}{k}, \qquad (1.41)$$

amivel

$$C_n^k = \frac{n!}{k!\cdot(n-k)!} = \binom{n}{k}. \qquad (1.42)$$

2. … DE MELYIK OLDALÁRA ESIK?

Véletlen események

Az előbbi fejezetben két olyan kísérlettel példálóztunk, amelynek kimenetét nem lehet pontosan megjósolni: a kocka elvetését megelőzően nem tudhatjuk, hogy az végül is melyik oldalára esik, és a céllövésnél sem tudjuk pontosan megjósolni a lövés leendő helyét (még akkor sem, ha a fegyvert egy merev állványra szerelve sütjük el). Ennek oka az, hogy a megfigyelés alatt álló jelenségek bonyolultsága miatt azokat egzakt matematikai modellel nem írhatjuk le, és emiatt a kísérlet eredményét sem tudjuk előre kiszámolni.

Az előbbi fejezetben definiáltuk a kísérlet kimeneteként kapott elemi eseményeket, valamint az elemi eseményekkel összetett eseményeket építettünk fel, és azok tulajdonságait is meghatároztuk. Mindezen tudás azonban nem elegendő annak megállapításához, hogy a kísérlet eredménye melyik elemi vagy összetett esemény lesz. A tapasztalat azt mutatja, hogy e kísérletekben, azok azonos körülmények közötti megismétlésénél, a lehetséges kimenetek közül véletlenszerűen bukkan elő egy-egy eredmény.

A kockadobáskor, ha elég nagy számú gurítást végzünk, az 1–6 számok mindegyike elő fog fordulni, méghozzá ha becsületesen játszunk, és nem befolyásoljuk valamilyen módon a várható eredményt, e számok nagy vonalakban azonos gyakorisággal fognak szerepelni.

II. Excel-szimuláció

Excelben, az előbbi fejezetben leírt szimuláció alapján a következőképpen modellezhetjük a kockadobás sorozatát:

1. Egy üres lapon hozzunk létre 1-től 10000-ig számozott cellákat, célszerűen a *B* oszlopban. Ez egy legtöbb 10000 dobásból álló sorozat szimulálásához lesz elegendő.

Gondoskodjunk arról, hogy az oszlop második cellája legyen kiválasztva (*B2*), az első sort ugyanis fejlécként fogjuk használni. Írjunk be ebbe a cellába egy „1”-est.

A „Home” menüszalagon válasszuk ki a „Fill” listából a „Series” tételt (2.1. ábra), majd töltsük ki a megjelenő táblázatot a 2.2. ábra szerint.

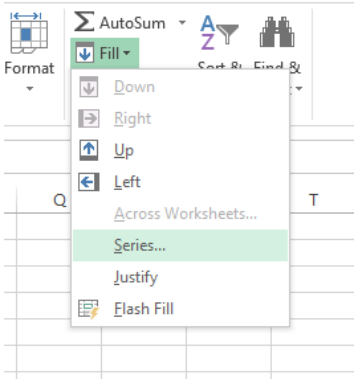
Az „OK” gomb megnyomása után a táblázatnak a kiválasztott cellát tartalmazó oszlopa számokkal telik fel 1-től 10000-ig, folyamatos számozást hozván így létre.

2. Jelöljünk ki három cellát, amelyek a bemeneti adatokat fogják tartalmazni. Ezek:

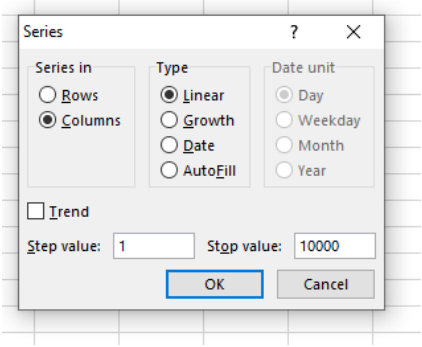
– a dobások N száma, az *A2*-es cellában;

– az előforduló legkisebb érték (kockadobás esetében ez 1), az *A5*-ös cellában;

– az előforduló legnagyobb érték (kockadobás esetében ez 6), az *A8*-as cellában.



2.1. ábra. *„Fill”, „Series”*

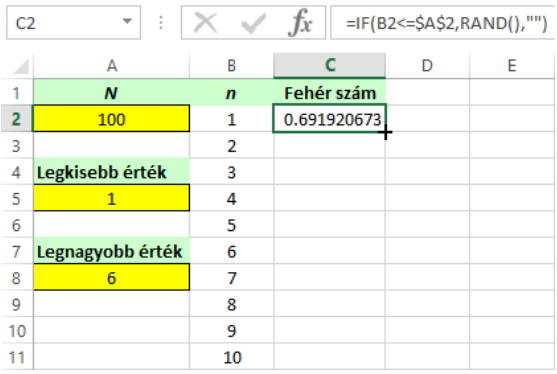


2.2. ábra. *Beállítások, adatok a kitöltéshez*

3. A C oszlopban hozzunk létre N véletlenszámot. Ehhez írjuk be a *C2*-es cellába a következő függvényt:

$$=\text{IF(B2<=\$A\$2,RAND(),""),} \qquad (2.1)$$

ahol az IF függvény által megvizsgált feltétel az, miszerint a cella sorszáma nem haladja meg az N paraméterrel megadott értéket (ellenben a cella üres marad). Másoljuk át ezt a képletet az oszlop többi megszámozott cellájába: ezt a legkönnyebben a cella jobb alsó sarkára való dupla kattintással tehetjük meg (a jobb alsó sarkon az egérkurzor egy fekete keresztté változik – 2.3. ábra).



2.3. ábra. *Képlet átmásolása*

Az így nyert valós véletlenszámok a $[0, 1)$ intervallumon egyenletesen oszlanak meg (a táblázatban „fehér” számok név alatt – ez a megnevezés a későbbiekben ismertetett „fehér zaj” fogalmából származik).

4. A C oszlop fehér számaiból hozzunk létre a megadott legkisebb és legnagyobb értékek (1 és 6) közé eső véletlen egész számokat. Ehhez írjuk be a *D2*-es cellába a következő függvényt:

$$=\text{IF(C2=""},\text{""},\$A\$5+\text{INT((1+}\$A\$8-\$A\$5)*C2)), \qquad (2.2)$$

miszerint ha a C oszlopban van adat (azaz a cella sorszáma nem haladja meg az N paraméterrel megadott értéket), akkor oda az

$$x = min + \text{Int}\left[(1 + max - min) \cdot rnd\right] \qquad (2.3)$$

egész szám kerül, ahol min az *A5*-ös cellában levő legkisebb, max az *A8*-as cellában levő legnagyobb érték, és rnd a C oszlopban szereplő fehér szám. A képletet másoljuk át az oszlop többi cellájába.

5. Az *E2* cellába írjuk be az első lehetséges előforduló értéket („=A5”, ez az *A5* cellában szereplő szám), majd az *F2* cellában határozzuk meg azt, hogy ez hányszor fordul elő a sorozatban. Ezt a következő képlettel tudjuk elvégezni:

=IF(E2="","",COUNTIF($D$2:$D$10001,E2)). (*2.4*)

6. Az *E3* cellától lefelé az oszlopot töltsük ki a többi lehetséges értékkel:

=IF(AND(E2<$A$8, E2<>""),E2+1,""), (*2.5*)

a képlet átmásolásával (csak az *A8* cellába beírt legnagyobb értéknek megfelelő számú cellában fog adat megjelenni), majd másoljuk át az *F2* cellában levő képletet is.

Az eredmény a 2.4. ábrán látható. Az F9 gombbal a véletlenszámokat frissíthetjük.

	A	B	C	D	E	F
1	***N***	***n***	**Fehér szám**	***x***	**Lehetséges értékek**	**Absz. gyakoriság**
2	100	1	0.117484522	1	1	12
3		2	0.214035551	2	2	19
4	**Legkisebb érték**	3	0.701354614	5	3	15
5	1	4	0.186501233	2	4	17
6		5	0.06768441	1	5	22
7	**Legnagyobb érték**	6	0.264276872	2	6	15
8	6	7	0.34766337	3		
9		8	0.200675195	2		
10		9	0.656334289	4		
11		10	0.605311968	4		

2.4. ábra. *Kockadobások sorozata*

A céllövésnél némileg eltérő dolgot tapasztalunk, az eredmények (a lövések) most is szóródnak, de azok valahol a tábla közepén tömörülnek. Ez utóbbi esetben már nehezebb számokkal kifejezni a dolgot, ilyenkor például a tábla középpontjából egymástól egyenlő távolságra koncentrikus köröket rajzolhatunk, és így azt kisebb területű részekre osztjuk. Megszámoljuk, hogy egy-egy ilyen gyűrű alakú rész, illetve a legbelső kör belsejébe hány találat esik, és így számszerűen kifejezhetjük a találatok gyakoriságát (másként: a végtelen sok pontból álló tartományt véges számú résztartományra osztjuk, és azt figyeljük, hogy a kísérlet kimenete melyik résztartományba esik).

III. Excel-szimuláció

Mivel a céltábla eseménytere kétdimenziós, egyszerűsítsük a feladatot akképpen, hogy a találatnak csak a középponttól számított x függőleges távolságát vesszük figyelembe (a céltábla középvonala alatt ezt negatívnak tekintjük). Így egy egydimenziós véletlen eseményhez jutunk.

A teljesebb körű szimuláció a találatok vízszintes irányú szórását is figyelembe kell vegye, valamint azt a tényt, hogy a lövések nem a céltábla közepe, hanem egy fegyvertől és lövésztől függő középpont körül tömörülnek.

1. Egy üres lapon hozzunk létre 1-től 10000-ig számozott cellákat a *B* oszlopban, ahogyan azt a kockadobás-sorozat szimulálása esetében tettük.

2. Az *A2*-es cella a lövések teljes N számát fogja tartalmazni, az *A5*-ös pedig azoknak a legtávolabbi találatok közötti, azonos hosszúságú intervallumoknak a k számát, amelyek szerint a távolság függvényében fogjuk megadni a találatok számát.

3. A *C* oszlopban hozzunk létre N véletlenszámot, a kockadobás-sorozat szimulálása során alkalmazott eljárással.

4. A *C* oszlopban szereplő számok egyenletesen oszlanak meg a $[0,1)$ intervallumon, a lövések pedig a céltábla közepe körül sűrűsödnek: emiatt az előbbiekre támaszkodva a 4.3. fejezet végén ismertetett Box–Müller-eljárás alapján, a valódi jelenséget jobban utánozó véletlen számokat hozunk létre:

$$x_i = \sqrt{-2 \cdot \ln rnd_i} + \cos(2 \cdot \pi \cdot rnd_{i+1}) \qquad (2.6)$$

és

$$x_{i+1} = \sqrt{-2 \cdot \ln rnd_i} + \sin(2 \cdot \pi \cdot rnd_{i+1}), \qquad (2.7)$$

amelyek felhasználásával a *C* oszlop két egymást követő cellájából kiolvasott rnd_i és rnd_{i+1} értékekkel számítjuk ki a *D* oszlop véletlen számait. Ehhez a *D2* cellába írjuk be a következő képletet, majd másoljuk azt át az oszlop többi cellájába is:

```
=IF(MOD(B2,2)=0,IF(C2="","",SQRT(-2*LOG(C1))*SIN(2*PI()*C2)),
    IF(C3="","",SQRT(-2*LOG(C2))*COS(2*PI()*C3))).
```
(*2.8*)

5. Az *A8*-as cellában határozzuk meg a *D* oszlopban előforduló véletlen számok minimumát:

```
=MIN(D2:D10001),
```
(*2.9*)

az *A11*-ben pedig azok maximumát:

```
=MAX(D2:D10001).
```
(*2.10*)

6. Az előbbi pontban megállapított szélsőértékekkel, valamint az intervallumoknak az *A5*-ös cellába beírt értékeivel állapítsuk meg azok határait. Ehhez az *E2*-es cellába írjuk be a legkisebb értéket (=A8), ami az első intervallum alsó határértéke. A felső határértékhez az *F2*-be írjuk be a következő képletet:

=IF(E2="","",E2+($A$11-$A$8)/$A$5). (*2.11*)

Ez utóbbi az alsó határértékhez (ami az *E2* cellában van) hozzáadja a maximum és a minimum különbségének és az intervallumok számának hányadosaként kiszámított lépést.

Ezután az *E3*-as cellába írjuk be az

=IF(AND(F2<$A$11,E2<>""),E2+($A$11-$A$8)/$A$5,"") (*2.12*)

képletet, amely lépteti az intervallum alsó határértékét, amennyiben az előbbi sorban szereplő intervallum felső határértéke még nem érte el a maximumot.

Az *E3* és az *F2* cellák képleteit másoljuk át az oszlop többi cellájába.

7. Az intervallumok határértékeivel meghatározzuk azok közepét, a *G* oszlopban. A *G2*-es cellában:

=IF(E2="","",(E2+F2)/2), (*2.13*)

majd másoljuk át a képletet az oszlop többi cellájába.

8. A *H* oszlopban meghatározzuk a *D* oszlop véletlen számainak intervallumonkénti gyakoriságát. A *H1* cellába írjuk be a következő függvényt:

=IF(E2="","",COUNTIF($D$2:$D$10001,"<="&F2)), (*2.14*)

(ez megszámolja az első intervallum felső határértékénél nem nagyobb elemek számát), az alatta levőbe pedig a következőt:

=IF(E3="","",COUNTIF($D$2:$D$10001,"<="&F3)-
COUNTIF($D$2:$D$10001,"<="&F2)). (*2.15*)

Másoljuk át ezt a képletet is az oszlop többi cellájába.

Az eredményt a 2.5. ábrán láthatjuk. Ebben az esetben is az F9 gombbal frissíthetjük a véletlenszámokat.

Abszolút és relatív gyakoriság

A kísérletek sokszori megismétlése után bizonyos szabályszerűségekre figyelhetünk fel, amit az elemi események gyakoriságával írhatunk le. Tegyük fel, hogy m kísérletet végzünk, és az m kísérlet alatt valamelyik A esemény bekövetkezéseinek száma m_A – ezt egyébként az esemény *abszolút gyakoriság*ának nevezzük. Ez a szám 0 és m között lehet (az előbbi esetben az A esemény sohasem következett be a kísérletek során, az utóbbiban pedig minden kísérlet eredménye az A esemény volt).

	A	B	C	D	E	F	G	H
1	**N**	**n**	**Fehér szám**	**x**		**Intervallum**	**Intervallum közepe**	**Abs. gyakoriság**
2	1000	1	0.704386054	0.27889874	-2.018020993	-1.814698932	-1.916359962	5
3		2	0.165650079	0.476018903	-1.814698932	-1.611376871	-1.713037901	5
4	**k**	3	0.164286408	-0.655185655	-1.611376871	-1.40805481	-1.50971584	10
5	20	4	0.662388538	-1.067486997	-1.40805481	-1.204732748	-1.306393779	16
6		5	0.239007177	-0.015805868	-1.204732748	-1.001410687	-1.103071718	34
7	**Legkisebb érték**	6	0.747743754	-1.114866938	-1.001410687	-0.798088626	-0.899749657	34
8	-2.018020993	7	0.311563747	-0.954819752	-0.798088626	-0.594766565	-0.696427596	80
9		8	0.448808889	0.318159403	-0.594766565	-0.391444504	-0.493105535	103
10	**Legnagyobb érték**	9	0.100708053	-1.328264917	-0.391444504	-0.188122443	-0.289783474	95
11	2.048420228	10	0.555100148	-0.479149152	-0.188122443	0.015199618	-0.086461413	126
12		11	0.729150265	0.455764767	0.015199618	0.218521679	0.116860648	121

2.5. ábra. *Találatok távolsága*

Hogy ha kiszámítjuk az esemény bekövetkezéseinek az összes elvégzett kísérlet számához viszonyított

$$f_A = \frac{m_A}{m} \qquad (2.16)$$

arányát (hányadosát), akkor ez a mennyiség az A esemény bekövetkezésének *relatív gyakoriság*át alkotja. Beláthatjuk, hogy ez a szám 0 és 1 között lehet, az

előbb említett két lehetőségnek megfelelően. E mennyiséget gyakran százalékban fejezzük ki.

A nagy számok törvénye

Az abszolút gyakoriság értéke az elvégzett kísérletek számától függ, a relatív gyakoriság pedig ugyan függhet az elvégzett kísérletek tényleges számától, de az megfigyelhetően egy középérték körül ingadozik, stabilan. Minél nagyobb számú kísérletet végzünk, a relatív gyakoriság annál közelebb lesz ehhez a középértékhez: ez a tapasztalati tény a *nagy számok (tapasztalati) törvénye*.

A kockadobálásnál a tapasztalat szerint bármely szám relatív gyakorisága 1/6 körül van (tehát a játékokban a továbblépéshez szükséges hatos dobás egyáltalán nem ritkább például az ötös dobásoknál), viszont a céllövésnél, mint már kijelentettük, a lövések száma a céltábla közepe felé sűrűsödvén, a találatok megfelelő relatív gyakorisága is a belső körök felé haladván egyre növekszik.

Megjegyzendő, hogy ha a kocka nem teljesen szabályos, vagy ha a puska félrehord, akkor is megfigyelhető lesz bizonyos szabályszerűség az eredmények megjelenésében, azonban az nem lesz azonos az elvártakkal.

Az esemény valószínűsége

Az A eseményhez hozzárendelhetünk egy $P(A)$-val jelölt számot, amely azt a mennyiséget jelöli, amely körül A bekövetkezésének relatív gyakorisága ingadozik. Ezt a számot az A esemény valószínűségének nevezzük.

A relatív gyakoriságoknál tett észrevételek értelmében tehát:

– valamely esemény valószínűsége egy 0 és 1 (vagy 0% és 100%) közötti szám;

– 0 vagy 0% a lehetetlen esemény valószínűsége, 1 vagy 100% a biztos esemény valószínűsége;

– ha A és B *egymást kizáró események*, akkor

$$P(A \cup B) = P(A) + P(B)\,. \qquad (2.17)$$

Az utóbbi három kijelentést a valószínűség Kolmogorov-féle axiómáiként ismerjük.

A megfigyelés szerint a kockadobásnál bármely szám azonos (1/6) valószínűséggel jelenik meg, tehát a címben levő kérdésre a válasz az, hogy a kocka bármely oldalára azonos valószínűséggel eshet (viszont azt előre nem tudjuk megmondani, hogy melyikre). Ha az A esemény 6-os, a B pedig páratlan szám dobása, akkor az első valószínűsége $P(A) = 1/6$, a másodiké pedig

$P(B)=3/6$, mivel a hat lehetséges számból három páratlan (1, 3 és 5). Ezek egymást kizáró események, mivel a hatos nem páratlan, így annak a valószínűsége, hogy hatos vagy páratlan szám legyen a dobás eredménye,

$$P(A\cup B)=P(A)+P(B)=1/6+3/6=4/6. \qquad (2.18)$$

A valószínűségek tulajdonságai

Az események tulajdonságait és a valószínűségek axiómáit figyelembe véve, a valószínűségek tulajdonságaira is következtethetünk.

Így, ha ismerjük valamely A esemény valószínűségét, akkor az ellentétes (komplementer) esemény valószínűsége

$$P(\bar{A})=1-P(A). \qquad (2.19)$$

Például ha az A esemény 6-os dobása (aminek a valószínűsége $P(A)=1/6$), akkor annak a valószínűsége, hogy ne hatost dobjunk,

$$P(\bar{A})=1-P(A)=1-1/6=5/6. \qquad (2.20)$$

Amennyiben az A_i $(i=1,n)$ események egy teljes eseményrendszert alkotnak, akkor valószínűségeik összege a biztos esemény valószínűsége:

$$\sum_{i=1}^{n}P(A_i)=P(A_1)+P(A_2)+\ldots+P(A_n)=1. \qquad (2.21)$$

Kockadobás esetében A_i-vel az i szám megjelenésének eseményét jelöljük. A_i valószínűsége i-től függetlenül 1/6. Mivel a lehetőségek összessége a hat szám megjelenéséből áll,

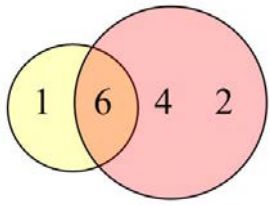
$$\sum_{i=1}^{6}P(A_i)=6\cdot(1/6)=1. \qquad (2.22)$$

Kolmogorov harmadik axiómája, miszerint $P(A\cup B)=P(A)+P(B)$, csak egymást kizáró eseményekre igaz. Példaként, ha az A esemény 1-es vagy 6-os, a B pedig páros szám dobása, akkor az első valószínűsége $P(A)=2/6$, a másodiké pedig $P(B)=3/6$, mivel a hat lehetséges számból három páros (2, 4 és 6). Ezek viszont nem egymást kizáró események, mivel a hatos páros, így a megjelenése egyszerre jelenti az A és B események bekövetkezését. Így a Kolmogorov harmadik axiómája alapján kiszámított összeggel az egyesített esemény bekövetkezésének valószínűsége nagyobb lenne a valóságosnál, s ekképpen abból ki kell vonni az A és B események együttes bekövetkezésének a valószínűségét

(2.6. ábra). Így születik Kolmogorov harmadik axiómájának általánosításaként a Poincaré-féle képlet:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \qquad (2.23)$$

ahol ezúttal A és B nem zárják ki egymást.



2.6. ábra. *Egymást nem kizáró események egyesítése*

Az A és B *egymástól független* események együttes bekövetkezésének eseménye,

$$P(A \cap B) = P(A) \cdot P(B), \qquad (2.24)$$

tehát példánkban $P(A \cap B) = (2/6) \cdot (3/6) = 1/6$, mivel csak a hatos páros, az egyes nem. Így annak a valószínűsége, hogy egyes, hatos vagy páros szám legyen a dobás eredménye,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 2/6 + 3/6 - 1/6 = 4/6, \qquad (2.25)$$

mivel az 1, 2, 4 és 6-os számok megjelenését várjuk az összesen hat lehetőség közül.

Ugyancsak két tetszőleges, A és B eseményre igaz:

$$P(A - B) = P(A) - P(A \cap B). \qquad (2.26)$$

Ha az A esemény páros szám dobása ($P(A) = 3/6$), B pedig hárommal osztható szám dobása (ez 3 vagy 6 lehet, $P(B) = 2/6$), akkor $P(A \cap B) = (3/6) \cdot (2/6) = 1/6$ annak a valószínűsége, hogy a szám kettővel is és hárommal is osztható legyen: egy ilyen lehetőségünk van, a 6-os szám. Így annak a valószínűsége, hogy páros, de hárommal nem osztható számot gurítsunk:

$$P(A - B) = P(A) - P(A \cap B) = 3/6 - 1/6 = 2/6; \qquad (2.27)$$

a két lehetőség a 2-es és a 4-es.

Sajátságos esetben, ha $B \subset A$, akkor az előbbi összefüggés a következőképpen módosul:

$$P(A - B) = P(A) - P(B). \qquad (2.28)$$

Például, ha A páros szám dobása (amire $P(A)=3/6$), B pedig 6-os dobása (amire $P(B)=1/6$), akkor

$$P(A-B)=P(A)-P(B)=3/6-1/6=2/6 \qquad (2.29)$$

annak a valószínűsége, hogy páros számot, de ne hatost gurítsunk (vagyis 2-est vagy 4-est).

Tekintsünk egy nevezetes egyenlőtlenséget is:

$$\text{ha } A\subseteq B\text{, akkor } P(A)\le P(B). \qquad (2.30)$$

Például, ha A a hatos dobásának az eseménye, B pedig a páros szám megjelenéséé, akkor ez utóbbit több kimenetel teljesíti, mint az elsőt. A bekövetkezése tehát csak az egyike B bekövetkezésének három lehetséges módozata közül. Emiatt a kísérletek ismételt elvégzése során a B esemény nagyobb gyakorisággal következik be az A-hoz viszonyítva, a nagy számok törvényének alapján pedig $P(B)$ háromszor nagyobb $P(A)$-nál.

Feltételes valószínűség

A továbbiakban vizsgáljuk meg két esemény együttes bekövetkezésének valószínűségét. Tegyük fel, hogy egy kísérlet során valamely B esemény abszolút gyakorisága m_B és ugyanakkor a B eseménnyel *egy időben* az A esemény $m_{A\cap B}$ számú kísérletben fordul elő. Amennyiben $m_B\neq 0$, definiálhatunk egy $m_{A\cap B}/m_B$ hányadost, amely az A esemény B-re vonatkoztatott *feltételes relatív gyakorisága*. Azt a számértéket, amely körül e hányados stabilan ingadozik az A esemény B-re vonatkoztatott *feltételes valószínűségé*nek nevezzük, és azt $P(A\,|\,B)$-vel jelöljük.

Ennek mintájára $m_{A\cap B}$ az A esemény B-re vonatkoztatott feltételes abszolút gyakoriságaként értelmezhető.

Ugyancsak a fentiek mintájára a $m_{A\cap B}/m_A$ hányadost a B esemény A-ra vonatkoztatott feltételes relatív gyakoriságaként értelmezhetjük, és bevezethetjük a $P(B\,|\,A)$ feltételes valószínűséget is.

Belátható, hogy ha az A esemény m_A abszolút gyakorisága nem egyenlő a B esemény m_B abszolút gyakoriságával, akkor a B esemény A-ra vonatkoztatott feltételes relatív gyakorisága és feltételes valószínűsége sem lesz azonos az A esemény B-re vonatkoztatott feltételes relatív gyakoriságával, illetve feltételes valószínűségével, vagyis $P(A\,|\,B)\neq P(B\,|\,A)$.

Bebizonyítható, hogy

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0, \tag{2.31}$$

illetve

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) \neq 0. \tag{2.32}$$

Megjegyzendő, hogy ha az A és B események egymástól nem függetlenek, akkor a számlálóban szereplő valószínűség

$$P(A \cap B) \neq P(A) \cdot P(B). \tag{2.33}$$

Ellenben, ha a két esemény független, akkor $P(A \cap B) = P(A) \cdot P(B)$ és a feltételes valószínűségek az A, illetve a B esemény valószínűségével lennének egyenlők: $P(A \mid B) = P(A)$, illetve $P(B \mid A) = P(B)$.

E két feltételes valószínűség értéke gyakran kísérleti megfigyelésből származik, azonban ha valamely elvek mentén meg tudjuk határozni a tört számlálóját és nevezőjét jelentő valószínűségeket, akkor ki is tudjuk számítani azokat.

Például ha $P(A \cap B)$ annak a valószínűsége, hogy a dobott páros szám éppen hatos ($P(A \cap B) = P(A) = 1/6$, mivel a hat lehetőségből csak a hatos dobása számít), B pedig a páros szám dobásának az esélye ($P(B) = 3/6$), akkor annak a valószínűsége, hogy páros szám dobása esetében az hatos legyen,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = 1/3: \tag{2.34}$$

a kísérletnek három egyforma valószínűséggel előforduló olyan lehetséges kimenetele van, hogy páros számot gurítsunk, ezek közül a hatos dobása az egyik a három közül.

A Bayes-tétel és a teljes valószínűség tétele

Ha a két feltételes valószínűség 13. és 14. kifejezéséből külön-külön kifejezzük a két esemény együttes megjelenésének $P(A \cap B)$ valószínűségét, akkor a következőkhöz jutunk:

$$\begin{aligned} P(A \cap B) &= P(B) \cdot P(A \mid B), \\ P(A \cap B) &= P(A) \cdot P(B \mid A). \end{aligned} \tag{2.35}$$

Ezek ekvivalenciájából következik, hogy

$$P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A), \qquad (2.36)$$

ami a feltételes valószínűség és a fordítottja közötti kapcsolatot adó *Bayes-tétel*.

Ha $\overline{B}$ a B esemény ellentettje (komplementer eseménye), akkor e tétel alapján

$$P(A \mid \overline{B}) \cdot P(\overline{B}) = P(\overline{B} \mid A) \cdot P(A). \qquad (2.37)$$

Az esemény és a komplementer esemény együttes valószínűsége 1 (a biztos esemény valószínűsége). Ez a kijelentés a feltételes eseményekre is vonatkozik, így kijelenthetjük, hogy $P(B \mid A) + P(\overline{B} \mid A) = 1$, és az előbbi két képlet összegzésével:

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B}). \qquad (2.38)$$

Ez utóbbi *a teljes valószínűség tétele*. Ebből adódóan, ha $B_1, B_2, \ldots B_n$ egy teljes eseményrendszert alkotnak (együttes valószínűségük 1), akkor e tételt egy n tagú összegként lehet általánosítani, az A eseménynek az eseményrendszer B_i eseményére vonatkoztatott feltételes valószínűségét figyelembe véve:

$$P(A) = \sum_{i=1,n} P(A \mid B_i) \cdot P(B_i). \qquad (2.39)$$

Független és nem független események

Az egymástól nem független események esetében – a feltételes valószínűség ismeretében – az események szorzatának valószínűségét a 16. képletek adják.

Amennyiben a feltétel nélküli és a feltételes valószínűségek egymással egyenlők:

$$P(A \mid B) = P(A), \text{ illetve } P(B \mid A) = P(B), \qquad (2.40)$$

akkor azt mondjuk hogy az események egymástól *sztochasztikusan* függetlenek: az A és B események egymástól függetlenül következnek be. Ez azt jelenti, hogy a kísérletileg megállapítható m_A / m hányados azonos (legalábbis nagyon közel áll) az $m_{A \cap B} / m_B$ hányadoshoz, ugyanakkor az m_B / m hányados megfelelőképpen közelít az $m_{A \cap B} / m_A$ hányadoshoz (anélkül, hogy az előbbiek szükségszerűen azonosak legyenek az utóbbiakkal). Ilyenkor az események szorzatának (együttes bekövetkezésének) valószínűsége az események valószínűségének szorzataként számítandó:

$$P(A \cap B) = P(A) \cdot P(B). \qquad (2.41)$$

Az egymást követő kockadobások egymástól függetlenek, tehát a következő megjelenő szám valószínűsége nem függ az addig megjelentektől. Ha például hatost dobtunk, akkor a következő gurításnál is 1/6 valószínűséggel fog a hatos megjelenni. A nagy számok törvénye szerint, ha nagyon sokszor megismételjük a kockadobást, a számok kb. azonos gyakorisággal fordulnak elő. Gyakori az a tévedés, miszerint ha egy szám már régóta nem jelent meg, akkor a nagy számok törvénye miatt egyre valószínűbb lesz az illető szám megjelenése: ez nem így van, a következő gurításnál az illető szám megjelenésének valószínűsége továbbra is 1/6 marad, még akkor is, ha az már nagyon rég nem jelent meg.

3. A KÍSÉRLETI EREDMÉNYEK ELEMEZÉSE

Valószínűségi változó

Az előbbi fejezet szerint, ha egy kísérletet változatlan körülmények között nagyon sokszor megismétlünk, akkor a kísérlet eredményeként megjelenő valamely véletlen esemény relatív gyakoriságát az illető esemény valószínűsége adja. Itt különböző kérdések vetődhetnek fel, mint például az, hogy melyik a legvalószínűbb kísérleti eredmény, és hogy hogyan számoljuk ki azt akkor, amikor a kísérletet nem tudjuk nagyon sokszor megismételni.

Tegyük fel, hogy egy bizonyos kísérletet végzünk el, például megmérjük egy automata szerszámgéppel előállított alkatrész valamilyen jellemző méretét, mondjuk egy csavar hosszát. Azt fogjuk tapasztalni, hogy ez a méret véletlenszerűen fog változni, bizonyos határértékek között.

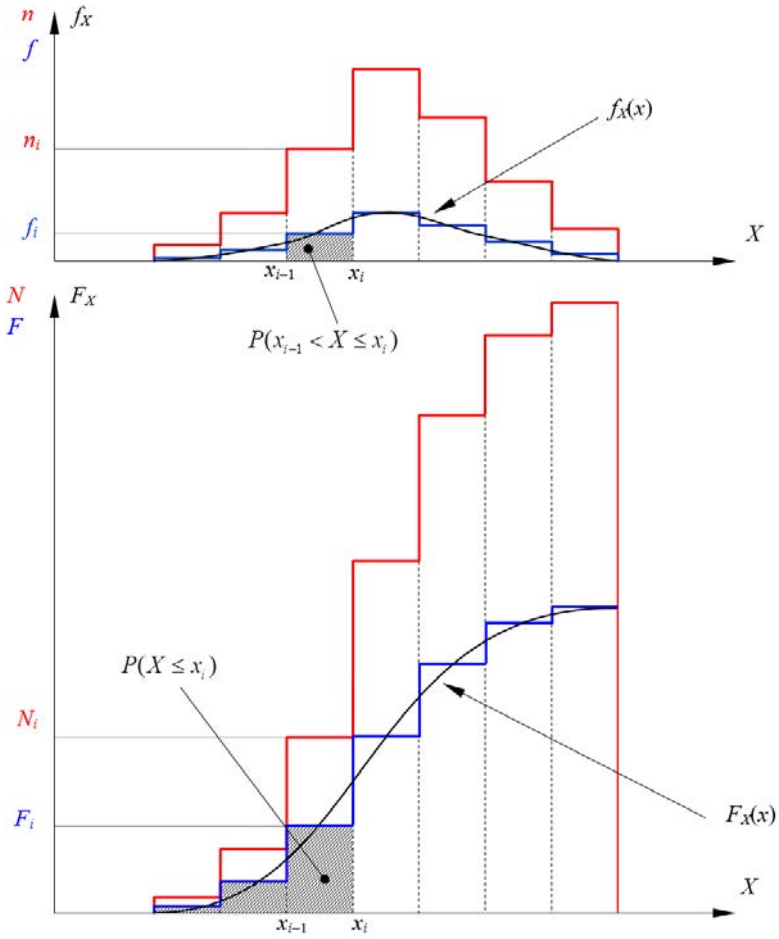
A kísérlet eredményét tehát egy véletlen értékeket felvevő változó írja le, valamely érték megjelenésének várható relatív gyakoriságát az adott érték megjelenésének (annak az eseménynek) a valószínűsége jelenti. Az ilyen véletlen értékeket felvevő mennyiségeket *valószínűségi változó*nak nevezzük.

A kísérlet kimeneteleit leíró információ nem feltétlenül számszerű (pl. az urnából kihúzott golyó színe nem az), azonban a jelenséget leíró összefüggések tanulmányozásának érdekében egy számszerű mennyiséget kell ahhoz hozzárendelnünk. Matematikai megfogalmazásban a valószínűségi változó egy függvény, amelynek az értelmezési tartománya a véletlen jelenség lehetséges kimeneteleit tartalmazó eseménytér, amelyet a valós számok halmazára képez le.

Van olyan eset, amikor a valószínűségi változó csak bizonyos értékeket vesz vagy vehet fel, ilyen például a kockadobásnál megjelenő szám. Az ilyen valószínűségi változót *diszkrét*nek nevezzük. Más esetben a valószínűségi változó, legalább egy intervallumon, bármilyen értéket felvehet, az ilyent *folytonos*nak nevezzük. Ilyen volt a céllövésnél a tábla közepe és a becsapódási pont közötti távolság.

A gyakoriságok hisztogramja

Nos, ha ábrázolni szeretnénk a kísérleti eredmények gyakoriságát, akkor rajzoljunk egy sík koordinátarendszert. A vízszintes X tengely fogja képviselni a diszkrét vagy folytonos valószínűségi változó értékeit, a függőleges pedig a valószínűségi változó értékeinek előfordulási gyakoriságát.



3.1. ábra. *Az abszolút és a relatív gyakoriságok hisztogramja (fent) és a kumulatív gyakoriságok hisztogramja (lent)*

Ha a tanulmányozott jelenséget leíró valószínűségi változó folytonos, akkor annak – még egy szűk intervallumon belül is – végtelen sok lehetséges értéke van. Azonban véges számú kísérletben csak véges számú érték fordulhat elő, a lehetőségek számát a mérőeszköz pontossága is korlátozza (a megmért mennyiséget csak véges számú tizedessel tudjuk megadni). A meg nem jelent értékek végtelenjének gyakorisága zérus lesz, a megjelent értékek pedig nagy valószínűséggel csak ritkán ismétlődnek meg. Ábránkat ez esetben nehezebb lenne kiértékelni. Ahhoz, hogy szemléletesebb ábrát kapjunk, a folytonos valószínűségi változó lehetséges értékeinek tartományát véges számú, gyakran

egyenlő hosszúságú intervallumra osztjuk fel, és azt ábrázoljuk, hogy az egy-egy intervallumon (osztályon) belüli értékek hányszor fordulnak elő. Ez tekinthető a folytonos valószínűségi változó egyfajta diszkretizálásának is.

Az ábrázolás lépcsős függvény formájában, téglalapokkal történik, mint ahogyan azt a 3.1. ábrán, annak felső grafikonján láthatjuk. Az ábrán X jelöli a tanulmányozott valószínűségi változót, n_i pedig annak az eseménynek az abszolút gyakoriságát, hogy X az x_{i-1} és az x_i határértékekkel kijelölt intervallumba esik. Ez az intervallum (osztály) $(x_i, x_{i+1}]$. Az első intervallumhoz hozzá kell számítanunk az alsó határát is, mert egyébként elvesztődne ez az érték.

Ugyanígy rajzolhatunk egy másik diagramot (lépcsős függvényt) is, ahol N_i azon események számát adja meg, midőn X értéke nem haladja meg az x_i határértéket. Ez a függvény az ábra alsó grafikonján látható. Az ilyen formájú függvényeket *hisztogram*oknak nevezik.

Ezek szerint a felső függvény az abszolút gyakoriságok hisztogramja, míg az alsó a kumulatív abszolút gyakoriságok hisztogramja. Az abszolút gyakoriságok hisztogramja azt mutatja, hogy egy adott $(x_{i-1}, x_i]$ intervallumra hány észlelt érték esett, és jelöli e számot n_i. A kumulatív abszolút gyakoriságok hisztogramja pedig azt, hogy hány érték van, amelyik nem haladja meg x_i-t (ha X a valós számok halmazának egészén vehet fel értékeket, akkor a $(-\infty, x_i]$ intervallumra eső észlelések számáról van szó), ezt a számot N_i-vel jelöljük. Belátható, hogy $N_i = \sum_{k=1}^{i} n_k$, az utolsó N érték pedig az elvégzett kísérletek M számát fogja adni, a kumulatív abszolút gyakoriságok pedig egy lépcsőzetesen növekvő függvényt definiálnak.

Definiálhatnánk egy harmadik függvényt is, amely az adott x_i határértéknél nagyobb (tehát például az $(x_i, +\infty)$ intervallumra eső) értékek megjelenésének abszolút N_i^* gyakoriságát adná. Ez a mennyiség azonban redundáns (és azt bizonyos esetekben mégis használjuk), mivel bármely i-re $N_i^* = M - N_i$.

Az előbbi fejezetben elmondtuk, hogy az abszolút gyakoriságok helyett hasznosabb és kézenfekvőbb a relatív gyakoriságokkal dolgozni, mivel használatuk esetén a különböző eredmények könnyebben összevethetők. A relatív gyakoriságok hisztogramját az abszolút gyakoriságok hisztogramjából könnyen meghatározhatjuk, minden n_i, illetve N_i értéket az elvégzett kísérletek számával,

tehát M-mel kell elosztanunk. A kétfajta hisztogram között tehát csak a függőleges tengelyen felvett értékek eltérő skálái jelentik az egyedüli különbséget.

A relatív gyakoriságok hisztogramján az f_i értékek annak az eseménynek a relatív gyakoriságát jelentik, hogy X értéke az $(x_{i-1}, x_i]$ intervallumra esik. A kumulatív relatív gyakoriságok hisztogramján az F_i értékek az x_i-nél kisebb értékek megjelenésének kumulatív relatív gyakoriságát adják. Könnyen belátható, hogy az f_i és az F_i értékek 0 és 1 közötti számok lehetnek, a kumulatív relatív gyakoriság legnagyobb (legutolsó) értéke pedig egységnyi, és az lépcsőzetesen növekszik 0 és 1 között.

Az előbbi fejezetben azt is elmondtuk, hogy ha kellőképpen nagy számú kísérletet végzünk el, akkor valamely esemény relatív gyakorisága az illető esemény valószínűségének jó közelítését fogja adni. Hisztogramjainkat úgy is skálázhatjuk, hogy a függvényszakaszok alatti terület éppen a megfelelő valószínűségeket adja. Ezeket a diagramokat a relatív gyakoriságok, illetve a kumulatív relatív gyakoriságok *normalizált* hisztogramjának nevezzük. A relatív gyakoriságok normalizált diagramjának az x_{i-1} és x_i közötti téglalapjának területe éppen az f_i relatív gyakorisággal kell, hogy egyenlő legyen, mert az a $P(x_{i-1} < X \le x_i)$ valószínűség közelítő értékét adja (ahol a közelítés pontossága a kísérletek M számával együtt növekedik). A kumulatív relatív gyakoriságok normalizált diagramján a hisztogram alatti terület, amelyet az x_i abszcisszáig számítunk, a $P(X \le x_i)$ valószínűség közelítését jelenti. A relatív gyakoriságok teljes normalizált hisztogramja alatti terület nagysága egységnyi.

A hisztogram normalizálása egyszerű: a relatív gyakoriságokat el kell osztani az intervallumok $\Delta x_i = x_i - x_{i-1}$ hosszával (ezek az intervallumok nem feltétlenül egyenlő hosszúságúak, a kapott érték jelenti a normalizált ordinátát.

A II. Excel-szimuláció folytatása

A 2. fejezetben a kockadobás-sorozat szimulálása során megállapítottuk a lehetséges eredmények előfordulásának abszolút gyakoriságát. Egészítsük ki a táblázatot a következőkkel:

7. A G oszlopban számítsuk ki a kumulatív abszolút gyakoriságokat. Az első, legkisebb lehetséges érték esetében ez egyenlő annak abszolút gyakoriságával

(=F2), a második értéktől kezdve pedig azt a következő függvénnyel állapítjuk meg:

=IF(F3="","",G2+F3); (*3.1*)

vagyis az illető érték abszolút gyakoriságához hozzáadjuk az azelőtti érték kumulált abszolút gyakoriságát.

8. A *H* oszlopban számoljuk ki a relatív gyakoriságokat (ehhez az abszolút gyakoriságokat a dobások N számával osztjuk, amit az *A2* cella tárol):

=IF(G2="","",F2/$A$2); (*3.2*)

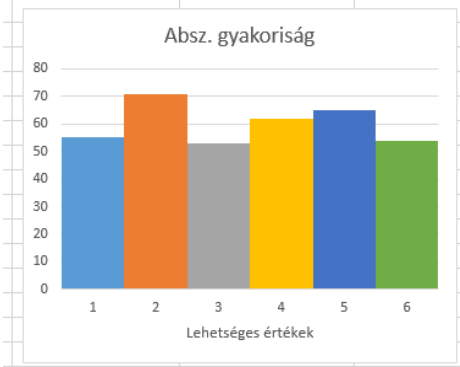
az *I* oszlopban pedig a kumulatív relatív gyakoriságokat (ehhez a kumulatív abszolút gyakoriságokat osztjuk a dobások N számával):

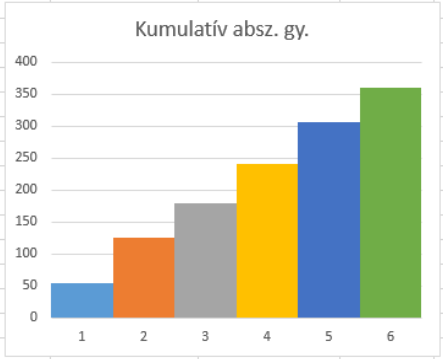
=IF(G2="","",G2/$A$2). (*3.3*)

Mindkét oszlopban a cellák formátumozását három tizedessel adjuk meg.

Ha a gyakoriságokat hisztogram formájában ábrázoljuk („Clustered column" opció a grafikonok ábrázolásánál), akkor az eredmény a 3.2. ábrán láthatóhoz hasonló lesz. Az ábrán a minták N száma 360.

E	F	G	H	I	J
Lehetséges értékek	**Absz. gyakoriság**	**Kumulatív absz. gy.**	**Rel. gyakoriság**	**Kumulatív rel. gy.**	
1	55	55	0.153	0.153	
2	71	126	0.197	0.350	
3	53	179	0.147	0.497	
4	62	241	0.172	0.669	
5	65	306	0.181	0.850	
6	54	360	0.150	1.000	

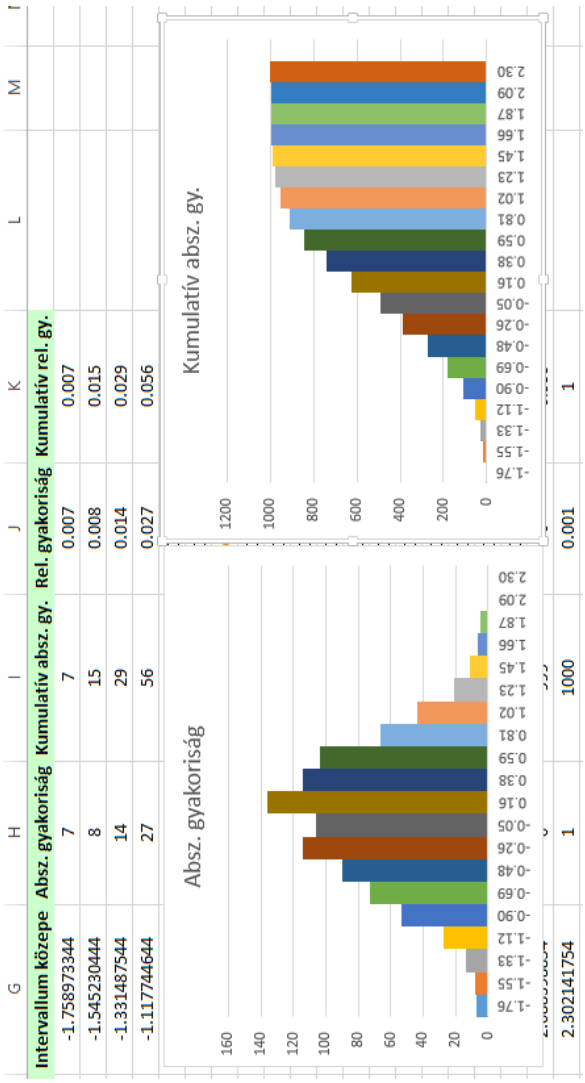




3.2. ábra. *Abszolút gyakoriságok kockadobás-sorozat esetén*

A III. Excel-szimuláció folytatása

A 2. fejezetben leírt céllövés-szimulálást is hasonlóképpen folytathatjuk, a táblázat kiegészítésével és a hisztogramok ábrázolásával. Az ábrákon a vízszintes tengelyen megjelenő értékek az intervallumok középértékei (3.3. ábra).



Intervallum közepe	Absz. gyakoriság	Kumulatív absz. gy.	Rel. gyakoriság	Kumulatív rel. gy.
-1.758973344	7	7	0.007	0.007
-1.545230444	8	15	0.008	0.015
-1.331487544	14	29	0.014	0.029
-1.117744644	27	56	0.027	0.056
	1	1000	0.001	1
2.302141754				

3.3. ábra. *Abszolút gyakoriságok céllövés esetén*

Ha a két szimuláció eredményét, grafikonjait összehasonlítjuk, akkor rögtön észrevehető, hogy a kétfajta kísérlet véletlen eredményei nem azonos módon

oszlanak el: míg a kockadobás esetében a lehetséges kimenetek kb. azonos valószínűséggel fordulnak elő, addig a céltáblát ért találatok helyzete valamilyen másfajta törvény szerint alakul, mert a középétől távolodva egyre kevesebb találatot számolhatunk meg.

A sűrűségfüggvény és az eloszlásfüggvény

Mire jó az előbbiekben említett normalizálás? A válasz egyszerű: folytonos valószínűségi változók esetében, ha $x_i \to x_{i-1}$, akkor $\Delta x_i = \mathrm{d}x \to 0$, a relatív gyakoriságok normalizált diagramján $f_i \to P(X = x_i)$ (tehát az ordináta az x_t abszcissza valószínűségét adja), míg a kumulatív relatív gyakoriságok normalizált diagramján $F_i \to P(X \le x_i)$ (tehát az ordináta azt a valószínűséget adja, hogy az X valószínűségi változó ne haladja meg az x_t értéket).

Ekkor, tehát ha $\Delta x_i \to 0$, a lépcsős függvények helyét folytonos függvények veszik át. Ezek grafikonját a 3.1. ábrán levő görbék jelentik. A felsőt, amely a relatív gyakoriságok normalizált hisztogramjából ered, az X valószínűségi változó *sűrűségfüggvény*ének nevezzük, és $f_X(x)$-szel jelöljük, míg az alsót, amely a kumulatív relatív gyakoriságok normalizált hisztogramjából származtatható, az X valószínűségi változó *eloszlásfüggvény*ének nevezzük, és $F_X(x)$-szel jelöljük.

A sűrűségfüggvény és az eloszlásfüggvény tulajdonságai

A két hisztogram és a két függvény definíciójából eredendően bizonyos összefüggéseket állapíthatunk meg:

– annak a valószínűsége, hogy az X valószínűségi változó értéke ne haladja meg x-et:

$$P(X \le x) = F_X(x), \qquad (3.4)$$

ami az $F_X(x)$ eloszlásfüggvény definíciója;

– a $P(X \le x)$ valószínűséget a sűrűségfüggvény x koordinátáig terjedő határozott integráljaként számíthatjuk ki:

$$F_X(x) = \int_{-\infty}^{x} f_x(t)\,\mathrm{d}t\,; \qquad (3.5)$$

– ennek alapján, a műveletet megfordítva, a sűrűségfüggvény az eloszlásfüggvény deriváltja:

$$f_X(x) = F_X'(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}, \tag{3.6}$$

ahonnan következik, hogy

$$P(x < X \le x + \mathrm{d}x) = f_X(x) \cdot \mathrm{d}x = \mathrm{d}F_X(x); \tag{3.7}$$

– innen az következik, hogy ha a véges δ kellőképpen kicsi, akkor a valószínűségi változó bármely x értékére a sűrűségfüggvény behelyettesítési értékének és δ-nek a szorzata annak a valószínűsége felé közelít, hogy a változó az $x < X \le x + \delta$ tartományra essen:

$$P(x < X \le x + \delta) \approx f_x(x) \cdot \delta; \tag{3.8}$$

– mivel az eloszlásfüggvény valószínűségeket ad vissza, az értéktartománya a valós számok $[0, 1]$ intervalluma:

$$0 \le F_X(x) \le 1, \quad \forall\, x \in (-\infty, +\infty); \tag{3.9}$$

– értelmezéséből fakadóan az eloszlásfüggvény monoton növekvő:

$$F_X(x_2) > F_X(x_1) \quad \text{ha} \quad x_2 > x_1, \tag{3.10}$$

ugyanis az $x \le x_1$ esemény implikálja az $x \le x_2$ eseményt. Az eloszlásfüggvény a bal oldalon a lehetetlen esemény valószínűségével indul:

$$F_x(x) = 0 \quad \text{ha} \quad x \to -\infty, \tag{3.11}$$

a teljes értelmezési tartományon (az eseménytér egészén) pedig a biztos esemény valószínűségét adja (ez a sűrűségfüggvény egésze alatti terület):

$$F_x(x) = 1 \quad \text{ha} \quad x \to +\infty, \tag{3.12}$$

avagy másképpen:

$$P(X \in \mathbb{R}) = \int_{-\infty}^{+\infty} f_X(x)\, \mathrm{d}x = 1 \cdot \tag{3.13}$$

Más szavakkal: az eloszlásfüggvény a bal oldalon nulla, a jobb oldalon pedig egységnyi.

– mivel az eloszlásfüggvény monoton növekvő, a sűrűségfüggvény (mint az előbbi deriváltja) sehol sem lehet negatív:

$$f_X(x) \ge 0, \quad \forall\, x \in (-\infty, +\infty); \tag{3.14}$$

– a valószínűségi változó végtelen nagy értékei nem fordulnak elő (a végtelen nagy értékek megjelenése a lehetetlen esemény):

$$P(X \to -\infty) = f_X(X \to -\infty) = 0, \tag{3.15}$$

$$P(X \to +\infty) = f_X(X \to +\infty) = 0\,. \tag{3.16}$$

Más szavakkal: a sűrűségfüggvény mindkét végletben nulla.

– mivel a sűrűségfüggvény és az eloszlásfüggvény közötti kapcsolatból eredően

$$P(X \le b) = \int_{-\infty}^{b} f_X(x)\,\mathrm{d}x = F_X(b) \tag{3.17}$$

és

$$P(X > a) = \int_{a}^{+\infty} f_X(x)\,\mathrm{d}x = 1 - F_X(a)\,, \tag{3.18}$$

annak a valószínűsége, hogy a valószínűségi változó az $(a, b]$ intervallumon vegyen fel értékeket,

$$P(a < X \le b) = \int_{a}^{b} f_X(x)\,\mathrm{d}x = F_X(b) - F_X(a)\,. \tag{3.19}$$

Megjegyzendő, hogy a fenti képletek bizonyos esetekben értelmezésre szorulnak:

– Így például ha az X valószínűségi változó csak egy bizonyos tartományon vehet fel értékeket, és nem a valós számok teljes $\mathbb{R} = (-\infty, +\infty)$ halmazán, akkor azt mondjuk, hogy ezen a tartományon kívül a sűrűségfüggvény értéke nulla. Például a lognormális eloszlású valószínűségi változó csak pozitív értékeket vehet fel, a valós számok teljes halmazán értelmezett sűrűségfüggvényének meghatározását pedig a következőképpen terjesztjük ki a valós számok teljes halmazára:

$$f_X(x) = \begin{cases} 0 & \text{ha } x \le 0, \\ \dfrac{1}{x} \cdot \dfrac{1}{\sqrt{2 \cdot \pi} \cdot \sigma_{\ln X}} \cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln x - m_{\ln X}}{\sigma_{\ln X}}\right)^2} & \text{ha } x > 0. \end{cases} \tag{3.20}$$

– Hasonlóképpen lehet kiterjeszteni az eloszlásfüggvény értelmezési tartományát is, azzal a megjegyzéssel, hogy az intervallumok közös határán az eloszlásfüggvény nem mindig differenciálható.

– Amikor X tulajdonképpen csak egy szűkebb, pl. $[A, B]$ intervallumon vehet fel értékeket, a sűrűségfüggvény értéke az intervallum szélein nullától különböző lehet (ld. az exponenciális eloszlás esetét).

– $f_x(x)\cdot\delta$ annak a valószínűségnek a közelítését adja, hogy az X valószínűségi változó érteke egy δ hosszúságú intervallumra essen. Ha az x értéket az intervallum közepén vesszük fel, akkor ezt a valószínűséget a következőképpen írjuk fel:

$$P(x-\delta/2<X\le x+\delta/2)\approx f_X(x)\cdot\delta, \qquad (3.21)$$

a közelítés pedig annál jobb, minél szűkebb az intervallum, tehát a δ értéket megfelelőképpen kicsinek vesszük (a „megfelelőképpen”-ben az is benne van, hogy a numerikus instabilitások miatt az nem lehet tetszőlegesen kicsi).

– Beláthatjuk, hogy amennyiben a valószínűségi változó folytonos, akkor annak sűrűségfüggvénye és eloszlásfüggvénye is folytonos lesz. Ellenben, ha a valószínűségi változó diszkrét, akkor a kétfajta függvény is diszkrét lesz. Mivel az eloszlásfüggvény nem folytonos, az nem differenciálható, és emiatt nem érvényesek a fenti $\mathrm{d}F_X(x)$-et tartalmazó összefüggések. Ebből kifolyólag a sűrűségfüggvény fogalmát a *valószínűségi függvény* helyettesíti, amelynek a behelyettesítési értéke ezúttal ténylegesen a $P(X=x)$ valószínűséget adja vissza. Diszkrét változójú függvények esetében az integrálást összegzés helyettesíti:

$$F_X(x_k)=\sum_{i=1}^{k} f_x(x_i)\,. \qquad (3.22)$$

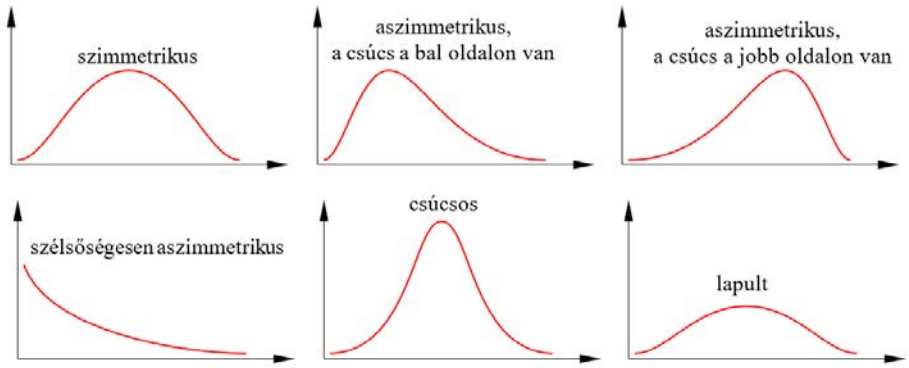
A diszkrét függvények ábrázolása értelemszerűen nem folytonos vonallal, hanem az $(x_k, f_X(x_k))$, illetve $(x_k, F_X(x_k))$ koordinátájú pontok halmazával történik, ahol $k=1,2,\ldots,n$.

A sűrűségfüggvény grafikonjának jellegzetességei

A sűrűségfüggvény ábrázolásakor különböző lefutású görbékhez juthatunk. A sűrűségfüggvény grafikonja lehet szimmetrikus, mérsékelten vagy szélsőségesen aszimmetrikus, az aszimmetrikus sűrűségfüggvény nagyobb értékei pedig megjelenhetnek annak jobb vagy bal oldalán. A sűrűségfüggvénynek lehet egy magasan kiemelkedő csúcsa, de annak lehet az ellaposodott formára jellemző tompa maximuma is. A 3.4. alábbi ábra a gyakrabban előforduló sűrűségfüggvény-típusokat példázza.

Megtörténhet, hogy a sűrűségfüggvénynek két vagy több csúcsa (módusza) is van. Az is megtörténhet (mint a kockavetés esetén), hogy a valószínűségi változó lehetséges értékei egyforma valószínűséggel jelennek meg: ebben az esetben a sűrűségfüggvény pontjai egy vízszintes, az eloszlásfüggvény pontjai pedig egy

ferde egyenesen helyezkednek el. Itt felvethetünk egy kérdést: mi van akkor, ha a lehetséges értékek száma végtelen? Ebben az esetben is a sűrűségfüggvény alatti terület (diszkrét változók esetén az ordináták összege) egységnyi kell, hogy legyen, tehát a függvényt jelentő vízszintes vonal infinitezimálisan közel, az abszcissza fölött helyezkedik el, az eloszlásfüggvényt jelentő vonal pedig egy olyan egyenes, amely a $-\infty$-ben zérus, a $+\infty$-ben pedig egységnyi értéket képvisel.



3.4. ábra. *Sűrűségfüggvények grafikonjának jellegzetességei*

A várható (átlagos) érték

Látván a lehetőségek nagy változatosságát, felvetődik hát az a kérdés, hogy miként jellemezhetjük a valószínűségi változók eloszlását?

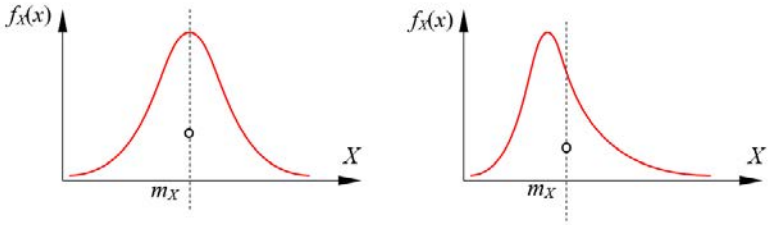
E kérdésre válaszolván az idők folyamán különböző mennyiségeket definiáltak és vezettek be, amelyek közül a fontosabbakat a következőkben soroljuk fel.

Először is definiáljuk az X valószínűségi változó átlagos, más néven várható értékét:

$$m_X = \sum_{i=1}^{Z} f_i \cdot x_i \quad \text{vagy} \quad m_X = \frac{\sum_{k=1}^{M} x_k}{M} = \frac{\sum_{i=1}^{Z} n_i \cdot x_i}{\sum_{i=1}^{Z} n_i}, \tag{3.23}$$

amit még $E(X)$, $M(X)$ vagy $\overline{X}$ formában is fellelhetünk (az „$E(\ldots)$” jelölés a várható érték kiszámításának operátora). E képletekben k az elvégzett kísérletek számára, i pedig a lehetséges kimenetek számára vonatkozik. Vegyük észre, hogy az x_i érték n_i alkalommal jelenik meg.

Matematikailag ez a mennyiség az előforduló x_i értékek súlyozott átlaga, a súlyok pedig a megfelelő gyakoriságok. Az átlagos értéket adó formulát geometriai értelemmel is felruházhatjuk: az nem más, mint a sűrűségfüggvény alatti terület geometriai középpontjának X mentén mért koordinátája. Ezen idomnak tehát az $x = m_X$ egyenlettel megadott tengelyre vonatkoztatott elsőrendű (sztatikai) nyomatéka zérus. Észrevehetjük, hogy az átlagos érték nem jelenti szükségszerűen a leggyakrabban előforduló (legvalószínűbb) értéket, ez az egyezés csak a szimmetrikus eloszlások esetén áll fenn.



3.5. ábra. *Az átlagos (várható) érték helyzete szimmetrikus és aszimmetrikus eloszlások esetében*

Amennyiben egy folytonos eloszlású valószínűségi változó sűrűségfüggvényét ismerjük, akkor az előbbi észrevételekkel, a geometriai középpont koordinátájának meghatározásával

$$m_X = \frac{\int_{-\infty}^{+\infty} x \cdot f_X(x)\,\mathrm{d}x}{\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x} = \int_{-\infty}^{+\infty} x \cdot f_X(x)\,\mathrm{d}x\,, \qquad (3.24)$$

ahol a számláló a görbe alatti terület függőleges tengelyre vonatkoztatott elsőrendű nyomatéka, a nevező pedig a sűrűségfüggvény görbéje alatti terület, ami definíció szerint egységnyi.

A szórásnégyzet

Az átlagos érték önmagában túl sokat nem árul el a valószínűségi változó eloszlásáról, így további jellemzőkre is szükségünk van. Fontos tudni például azt, hogy az előforduló értékek mennyire szorosan csoportosulnak ezen átlag körül. Ha például a céllövés kísérletében azt látjuk, hogy egy egyén esetében a találatok sokasága a céltábla középpontja körül csoportosul, akkor az illetőt remek lövészként olimpiai megmérettetésre lehet küldeni. Ha pedig egy gép által

előállított csapágygolyók mérete erősen szóródik a nominális körül, akkor annak a gépnek a gazdaságossága nem a legjobb, mivel túl sok selejtet állít elő.

Egy kísérleti eredmény eltérésének mértéke a meghatározott x_i érték és az átlag különbsége, tehát $x_i - m_X$ lenne. Ha ezeknek az eltéréseknek a közönséges számtani átlagát számítanánk, akkor az nulla lenne:

$$\sum_{i=1}^{M}(x_i - m_X)/M = \left(\sum_{i=1}^{M} x_i\right)/M - M \cdot m_X / M = m_X - m_X = 0 \cdot \qquad (3.25)$$

Ilyen esetben, amikor a pozitív és negatív irányú eltérések egymás hatását kioltják, nem közönséges számtani átlagot, hanem az eltérések négyzetének (e négyzetek ugyanis pozitív, esetleg nulla mennyiségek lesznek) vagy moduluszainak átlagát számoljuk. Az utóbbi lehetőség a következő mennyiséghez vezet:

$$\Delta_X = \frac{\sum_{k=1}^{M} | x_k - m_X |}{M} = \frac{\sum_{i=1}^{Z} n_i \cdot | x_i - m_X |}{\sum_{i=1}^{Z} n_i}, \qquad (3.26)$$

avagy a sűrűségfüggvény ismeretében

$$\Delta_X = \int_{-\infty}^{+\infty} f_X(x) \cdot | x - m_X | \, \mathrm{d}x, \qquad (3.27)$$

s ezt az átlagos abszolút eltérésnek nevezik. Ezzel a jellemzővel azonban csak ritkán találkozunk, sokkal elterjedtebb a négyzetek átlagának használata:

$$\sigma_X^2 = \frac{\sum_{k=1}^{M} (x_k - m_X)^2}{M} = \frac{\sum_{i=1}^{Z} n_i \cdot (x_i - m_X)^2}{\sum_{i=1}^{Z} n_i}, \qquad (3.28)$$

vagy pedig a sűrűségfüggvény ismeretében:

$$\sigma_X^2 = \int_{-\infty}^{+\infty} f_X(x) \cdot (x - m_X)^2 \, \mathrm{d}x \cdot \qquad (3.29)$$

Ez a mennyiség a szórásnégyzet vagy négyzetes szórás, melynek négyzetgyöke a σ szórás. E mennyiségeket is elláthatjuk geometriai értelemmel: a szórásnégyzet a sűrűségfüggvény alatti terület $x = m_X$ tengelyhez viszonyított másodrendű (tehetetlenségi) nyomatéka, ugyanis az elemi $f_X(x) \cdot \mathrm{d}x$ területet a távolság $(x_k - m_X)^2$ négyzetével szorozzuk meg. A szórás pedig

$$\sigma_X = \sqrt{\frac{\sigma_X^2}{\int_{-\infty}^{-\infty} f_X(x)\,\mathrm{d}x}}, \qquad (3.30)$$

tehát a négyzetgyök alatt az idom tehetetlenségi nyomatékának és területének hányadosa áll: ezek szerint a szórás a függvény alatti terület az $x = m_X$ tengelyhez viszonyított tehetetlenségi sugara. A szórás más jelölése: $D(X)$, ahol a „$D(\ldots)$" jelölés a szórás kiszámításának operátora. A szórás angol megnevezése „standard deviation". A szórásnégyzetet az angol szakirodalomban varianciának nevezik és $\mathrm{Var}(X)$-ként is jelölik.

Variációs együttható

Az átlagos értékkel és a szórással egy újabb jellemzőt, a

$$V_X = \frac{\sigma_X}{m_X} \qquad (3.31)$$

variációs együtthatót lehet definiálni, amely egy dimenzió nélküli szám. Ha az átlagos érték zérus, akkor a variációs együttható végtelenül nagy. Ezt a különböző fizikai természetű valószínűségi változók eloszlásának összehasonlításánál szokták használni.

Ferdeségi együttható

Valamely síkidom esetében magasabb rendű nyomatékokat (momentumokat) is lehet definiálni. Az m_X-en keresztül húzott függőleges tengelyre vonatkoztatva azokat az $\int_{-\infty}^{+\infty} (x - m_X)^n\,\mathrm{d}x$ integrál adja meg. Műszaki körökben ezeket (ha $n > 2$) nem igazán használják, azonban a valószínűségi változók vizsgálatának terén ezek is szerephez jutnak.

Így a másodrendű és a harmadrendű nyomatékok segítségével megadhatunk egy ferdeségi együtthatót:

$$\beta_{1,X} = \frac{\sum_{i=1}^{Z} n_i \cdot (x_i - m_X)^3}{\left[\sum_{i=1}^{Z} n_i \cdot (x_i - m_X)^2\right]^{3/2}} = \frac{\sum_{k=1}^{M} (x_k - m_X)^3}{\left[\sum_{k=1}^{M} (x_k - m_X)^2\right]^{3/2}}, \qquad (3.32)$$

vagy, folytonos esetben:

$$\beta_{1,X} = \frac{\int_{-\infty}^{+\infty} f_X(x)\cdot(x-m_X)^3 \mathrm{d}x}{\left[\int_{-\infty}^{+\infty} f_X(x)\cdot(x-m_X)^2 \mathrm{d}x\right]^{3/2}}. \qquad (3.33)$$

A szakirodalomban ezt néha γ-val jelölik.

Amennyiben $\beta_{1,X} = 0$, úgy a sűrűségfüggvény szimmetrikus az $x = m_X$ tengelyre nézve. Ha $\beta_{1,X} > 0$, akkor a sűrűségfüggvény maximuma az átlagos értéktől balra esik, és azt pozitív ferdeségűnek nevezzük, ellenben az attól jobbra van, és a sűrűségfüggvény negatív ferdeségű.

Lapultsági tényező

A negyedrendű és a másodrendű momentumokkal egy másik mennyiség, a lapultsági együttható:

$$\beta_{2,X} = \frac{\sum_{i=1}^{Z} n_i \cdot \sum_{i=1}^{Z} n_i \cdot (x_i - m_X)^4}{\left[\sum_{i=1}^{Z} n_i \cdot (x_i - m_X)^2\right]^2} = \frac{M \cdot \sum_{k=1}^{M} (x_k - m_X)^4}{\left[\sum_{k=1}^{M} (x_k - m_X)^2\right]^2}, \qquad (3.34)$$

folytonos esetben

$$\beta_{2,X} = \frac{\int_{-\infty}^{+\infty} f_X(x)\cdot(x-m_X)^4 \mathrm{d}x}{\left[\int_{-\infty}^{+\infty} f_X(x)\cdot(x-m_X)^2 \mathrm{d}x\right]^2}. \qquad (3.35)$$

Értéke a sűrűségfüggvény csúcsosságát adja (kurtózis, curtosis), amennyiben az nagyobb, mint 3, akkor az eloszlás csúcsos, magas (leptokurtózis), ha pedig az kisebb, mint 3, akkor az eloszlás lapos (platikurtózis). A 3-as érték nyilván egy referenciaérték, éspedig az egy nevezetes eloszlás, a normál eloszlás lapultsági tényezőjének értéke. A 3-as értéktől való eltérés a *fölöslegtényező*.

A ferdeségi és a lapultsági tényezőt a szakirodalom nem jelöli egységesen (a magyar sem).

Kvantilisek

A valószínűségi változók eloszlásához még néhány, a műszaki szempontból fontos fogalom kapcsolódik, ilyen a *kvantilisek* fogalma. A cél k darab, változó

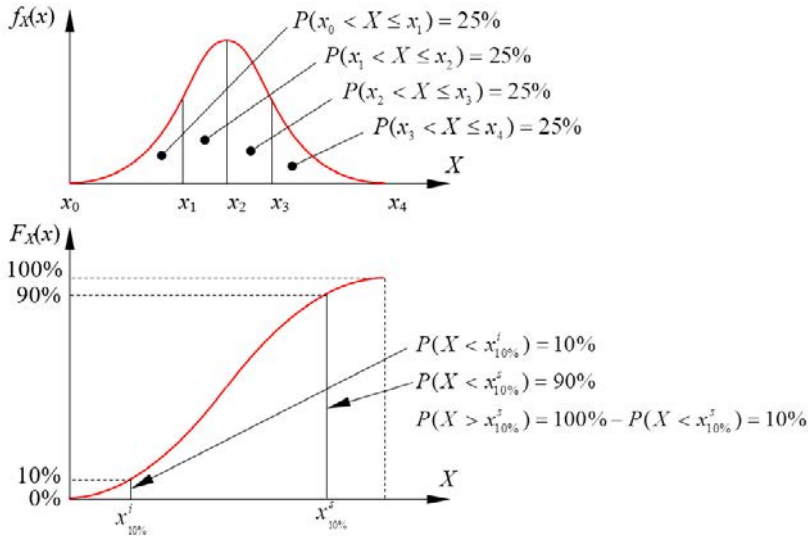
hosszúságú Δx_j, $j=1,k$ intervallum elkülönítése, amelyek egységesen M/k mért értéket tartalmaznak. Folytonos esetben a sűrűségfüggvénynek a Δx_j intervallumokra eső részterületei egyenlő nagyságúak (mindegyik terület $1/k$ nagyságú), az intervallumokat elhatároló pontokban pedig az eloszlásfüggvény $1/k$ lépcsőkben növekedő értékeket mutat. Annak a valószínűsége, hogy az X változó valamely Δx_j intervallumból vegyen értékeket, minden intervallumon azonos, vagyis $P(x_j-\Delta x_j<X\le x_j)=1/k$ bármely lehetséges j-re. A k számú intervallum elhatárolásához $k+1$ pontra van szükségünk, ezek x_j koordinátái a k-ad rendű kvantilisek.

Egyes kvantiliseket saját névvel jelölnek, így amikor k értéke 2, akkor mediánról, ha 4, akkor kvartilisról, ha 5, akkor kvintilisről, ha 10, akkor decilisről, ha pedig 100, akkor centilisről beszélünk. Legnevezetesebb ezek közül a *medián*, amely az eloszlásfüggvény 0.5 (50%) értékének felel meg. Jelölhetjük ezt az mennyiséget $\breve{m}_X$-szel, X ennél kisebb, illetve nagyobb értékek ugyanazzal a gyakorisággal fordulnak elő.

Fraktilisek

Némileg hasonló a *fraktilis*ek fogalma, ekkor azonban a cél nem az egyenlő valószínűséget mutató intervallumok meghatározása, hanem X értéktartományának két részre való elkülönítése: a fraktilis egy olyan érték, amely alatt vagy felett a valószínűségi változó egy meghatározott P valószínűséggel fordul elő (és egyértelmű, hogy a másik oldalon $1-P$ valószínűséggel vesz fel értékeket). Ezt a P valószínűséget a műszaki gyakorlatban százalékban szokták megadni, ekkor beszélünk „P százalékos alsó”, illetve „felső” fraktilisról. Értelmet nyer mindez, ha arra gondolunk, hogy a méretezésnél általában mind a terhelést, mind az anyag szilárdságát valószínűségi változók írják le. A kerékpár terhelését a kerékpáros személyének súlya jelenti. Nyilvánvaló, hogy a kerékpárt nem méretezhetjük az emberek átlagos testsúlyára, mert mi lesz akkor, ha az illető történetesen egy túlsúlyos személy? Az is nyilvánvaló, hogy az eddig élt emberek legnagyobb testsúlyára történő méretezés sem praktikus, mivel kevés ilyen ember él a Földön, és nem is biztos, hogy használnák a kerékpárunkat. Ilyen esetben a műszaki gyakorlat a terhelés valószínűségi eloszlásával számol, és bizonyos biztonsági szempontokat figyelembe véve kitűz egy P valószínűséget, amellyel a tényleges terhelés meghaladhatja a tervezésnél figyelembe vettet. Nyilvánvaló,

hogy minél kisebb ez a P valószínűség, annál nagyobb az a terhelés, amelyre méretezünk. Ez a P valószínűség egy felső fraktilist ad. A tervezésnél a másik figyelembe vett mennyiség a szerkezet szilárdsága, amely többek között az anyagparaméterek és a méretek változatossága miatt szintén valószínűségi változó. Itt a legbiztonságosabb dolog a lehető legkisebb érték elfogadása lenne, de hát ennek előfordulási valószínűsége igen kicsi, és még kisebb annak a valószínűsége, hogy pont ezt a leggyengébb szerkezetű kerékpárt a tervezésnél figyelembe vett legnagyobb súllyal rendelkező személy fogja használni. Itt is elfogadunk egy, a tervezett szerkezet fontosságával, kért megbízhatóságával összhangban levő P valószínűséget, amellyel a szerkezet szilárdsága a tervezett alá csökkenhet: ez egy alsó fraktilist határoz meg.



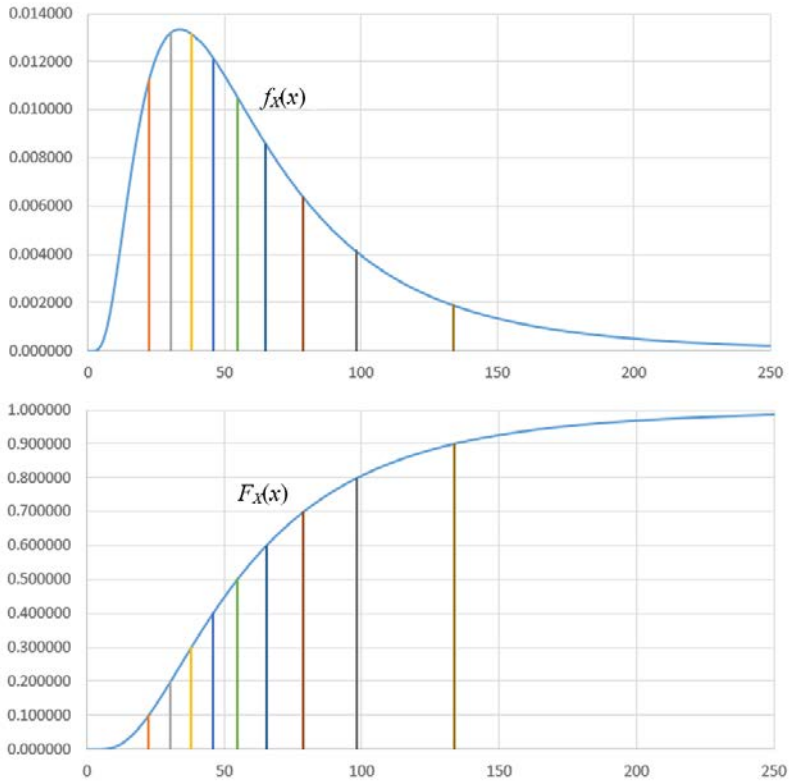
3.6. ábra. *Kvantilisek és fraktilisek*

A felső fraktilis értelmezése tehát az, hogy a változó értéke P százalékkal lehet nagyobb a tervezésnél figyelembe vett legnagyobb értéknél, az alsóé pedig az, hogy a változó P százalékkal lehet kisebb a figyelembe vett legkisebb értéknél.

A kvantilisek és a fraktilisek értelmezését az alábbi (3.6) ábrán látjuk. A sűrűségfüggvény alatti területet az $x_0, \ldots x_4$ kvantilisek négy azonos nagyságú részre osztják, míg az eloszlásfüggvény i és s indexű fraktilisei a 10% biztonsággal megadott minimális és maximális értékeket jelentik.

Példaként a 3.7. ábrán egy $m_{\ln X} = 4$ átlagú és $\sigma_{\ln X} = 0.7$ szórású lognormális eloszlású változó deciliseit (10%-os lépéssel felvett kvantiliseit) láthatjuk. Ezek:

P	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x	22.26	30.29	37.82	45.73	54.60	65.19	78.81	98.41	133.90



3.7. ábra. *Lognormális eloszlású változó decilisei*

A sűrűségfüggvény alatti területet a táblázatban szereplő x koordináták 10 egyenlő részre osztják.

Az eloszlás mediánja a $P = 0.5$ valószínűségnek megfelelő $x = 54.60$ érték: az x mennyiség ennél kisebb és ennél nagyobb értékei egyaránt 50%-os valószínűséggel fordulnak elő.

Az $x = 22.26$ értékhez tartozó 10%-os valószínűség azt jelenti, hogy ennél kisebb értékek 10%-os, az ennél nagyobb értékek pedig 90%-os valószínűséggel fordulnak elő. Ez az érték tehát a 10%-os valószínűségnek megfelelő alsó fraktilis.

Az $x = 133.90$ értéknél kisebb értékeket 90%-os, annál nagyobb értékeket pedig 10%-os valószínűséggel várhatunk, így ez az érték a 10%-os valószínűségnek megfelelő felső fraktilis.

A legvalószínűbb érték: a módusz

Azt láthattuk, hogy aszimmetrikus eloszlásoknál az átlagos érték nem a sűrűségfüggvény legnagyobb értékénél van. Ezt a legnagyobb értéket az X változó legvalószínűbb értéke jelenti, ott a sűrűségfüggvénynek maximuma, az eloszlásfüggvénynek inflexiós pontja van. Ezt az értéket az eloszlás (a valószínűségi változó) *módusz*ának nevezik, jelölhetjük azt $\tilde{m}_X$ -szel.

Az előbbi példában szereplő lognormális eloszlás esetén $\tilde{m}_X = 33.45$. Ennek az eloszlásnak a mediánja, mint láttuk, $\breve{m}_X = 54.60$, a következő fejezetben szereplő képlettel kiszámolt átlaga pedig $m_X = 69.76$.

Az utóbbi mennyiségek (a kvantilesek – közötte a módusz –, a fraktilisek és a medián) képletileg csak a sűrűségfüggvény és az eloszlásfüggvény ismeretében határozhatók meg. Megjegyzendő, hogy szimmetrikus eloszlásnál $m_X = \breve{m}_X = \tilde{m}_X$.

Bizonyos eloszlásoknál (például a könyvben nem szereplő bétaeloszlás a paraméterezésének függvényében bimodális lehet) a sűrűségfüggvénynek több csúcsa is lehet: ezek lokális maximumokat jelentenek. Az egyenletes eloszlás sűrűségfüggvényének minden pontja módusz.

4. FONTOSABB ELOSZLÁSOK

4.1. Az egyenletes eloszlás

A valószínűségi változó egyenletes eloszlásáról akkor beszélünk, amikor a lehetséges értékei egyenlő gyakorisággal, valószínűséggel jelennek meg. Egyenletes eloszlást mutat például a kockadobás eredménye, ahol a hat szám – ha kellő számú kísérletet végzünk el – közel azonos gyakorisággal jelentkezik. E példában az eloszlás diszkrét, mert a valószínűségi változónak csak néhány lehetséges értéke van.

Léteznek azonban legalább egy intervallumon folytonos egyenletes eloszlású valószínűségi változók is. Ha a folytonossági intervallum a valós számok teljes $\mathbb{R}$ halmazával azonos lenne, akkor a tanulmányozott változó bármely értékének valószínűsége nulla $(1/\infty)$ lenne. Ilyen eset a gyakorlatban nem fordul elő, éppen ezért az egyenletes eloszlású valószínűségi változó lehetséges értékeinek tartományát általában leszűkítik valamely $[x_{min}, x_{max}]$ intervallumra, és ekkor a sűrűségfüggvényét a következőképpen lehet definiálni:

$$f_X(x) = \begin{cases} \dfrac{1}{x_{max} - x_{min}} & \text{ha } x_{min} \le x \le x_{max},\ x_{min} \ne x_{max}, \\ 0 & \text{ha } x < x_{min} \text{ vagy } x > x_{max}. \end{cases} \qquad (4.1)$$

Ha $(x_{max} - x_{min}) < 1$, akkor a leszűkített intervallumon $f_X(x) > 1$.

Az ennek megfelelő eloszlásfüggvény a következő:

$$F_X(x) = \begin{cases} 0 & \text{ha } x < x_{min}, \\ \dfrac{x - x_{min}}{x_{max} - x_{min}} & \text{ha } x_{min} \le x \le x_{min},\ x_{min} \ne x_{max}, \\ 1 & \text{ha } x > x_{max}. \end{cases} \qquad (4.2)$$

E két függvény grafikonja a 4.1. ábrán látható.

Az egyenletes eloszlást mutató valószínűségi változó átlagos (várható) értéke a lehetséges értékek intervallumának közepe lesz:

$$m_X = \frac{x_{min} + x_{max}}{2}, \qquad (4.3)$$

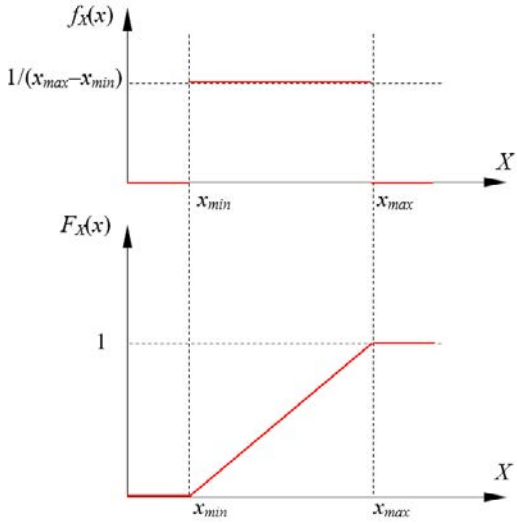
a szórásnégyzetet pedig

$$\sigma_X^2 = \frac{(x_{max} - x_{min})^2}{12} \qquad (4.4)$$

formában határozhatjuk meg. Ezekkel a variációs tényező

$$V_X = \frac{\sigma_X}{m_X} = \frac{x_{max} - x_{min}}{x_{max} + x_{min}} \cdot \frac{1}{\sqrt{3}}. \qquad (4.5)$$

Észrevehetjük, hogy az eloszlás szimmetrikus, ferdeségi tényezője $\beta_{1,X} = 0$, lapultsági tényezőjét ha kiszámítjuk, akkor a $\beta_{2,X} = \frac{9}{5} = 1.8$ értékhez jutunk (tehát az eloszlás lapos, platikurtikus). A szimmetria miatt az $\tilde{m}_X$ medián és az $\breve{m}_X$ módusz a középértékkel azonos.



4.1. ábra. *Egyenletes eloszlású folytonos valószínűségi változó sűrűségfüggvénye és eloszlásfüggvénye*

Van az egyenletes eloszlásnak egy sajátságos esete, amikor $x_{min} = 0$ és $x_{max} = 1$: ezt standard egyenletes eloszlásnak nevezik. Ekkor $m_X = 0.5$ és $\sigma_X^2 = 1/12 \approx 0.083$.

Egyenletes eloszlású számok létrehozása

A programozási nyelvek általában rendelkeznek egy olyan utasítással, amellyel standard egyenletes eloszlást mutató, legalábbis azt megközelítő számokat lehet előállítani. Az így előállított értékek lehetséges tartománya a $[0,1)$ intervallum, amely nem tartalmazza a felső határértékként megadott 1-est. Ezek a számok nem véletlen számok, mert valamilyen műveletsorral (tehát képletileg leírható, determinisztikus módon) állítják azokat elő, azonban az alkalmazott algoritmustól

függően jobb vagy rosszabb közelítését adják egy standard eloszlású véletlen számsornak. E számokat *pszeudovéletlen*nek (pseudo-random) nevezik. E számsor véges hosszúságú, azonban rendszerint eléggé hosszú ahhoz, hogy a programfutás alatt ne ismétlődjön meg az elkezdett szekvencia. Mi több, hogy a számsor két egymást követő előállításakor ne ismétlődjön meg, mert akkor a szimulált véletlen folyamat mindig azonos séma szerint zajlana le, általában szabadon megadható a belépési pont helye (hogy honnan kezdjük a sort). Hogy az ismétlődéseket itt is elkerüljük, pl. a számítógép óráját (az éjfél óta eltelt időt) veszik e paraméterként.

Amennyiben egyenletes, de nem standard eloszlású számokra van szükségünk, akkor a következő két módon járhatunk el:

– megadjuk a kívánt átlagos értéket és a kívánt szórásnégyzetet, ekkor a véletlen számsor intervallumát lehatároló értékeket e két mennyiség definíciójából határozhatjuk meg:

$$m_X = \frac{x_{min} + x_{max}}{2} \rightarrow x_{min} + x_{max} = 2 \cdot m_X \,, \qquad (4.6)$$

$$\sigma_X^2 = \frac{(x_{max} - x_{min})}{12} \rightarrow x_{max} - x_{min} = 2 \cdot \sqrt{3 \cdot \sigma_X} \,, \qquad (4.7)$$

ahonnan, az egyenletrendszert megoldva a határértékek

$$x_{min} = m_X - \sqrt{3} \cdot \sigma_X \,,\; x_{max} = m_X + \sqrt{3} \cdot \sigma_X \,. \qquad (4.8)$$

E határértékekkel, ha rnd a számítógép által előállított standard egyenletes eloszlású szám, akkor

$$x = x_{min} + (x_{max} - x_{min}) \cdot rnd \,. \qquad (4.9)$$

– ha az intervallumot meghatározó x_{min} és x_{max} értékeket adjuk meg, akkor csak az előbbi műveletet kell elvégeznünk, az átlag és a szórásnégyzet kiszámítása (ha egyáltalán szükségesek) az ismert képletekkel történik.

Érdekességként megjegyezhető, hogy léteznek „valódi” véletlenszám-generátorok, amelyek működési elve valamilyen véletlen fizikai jelenségen alapul (például félvezetők termikus zaján). Ha a kockadobás kísérletét valódi véletlen számokkal szeretnénk tanulmányozni, akkor vegyük kézbe a kockákat.

4.2. A binomiális eloszlás

Tegyük fel, hogy összesen n független kísérletet végzünk, egy bizonyos A esemény bekövetkezését figyelve. Mivel kísérletsorozatunkban azt figyeljük, hogy

az A esemény bekövetkezik-e, vagy sem, eseményterünk A-ra és $\bar{A}$-ra szűkül be. Tapasztalatunk szerint ez az esemény egy bizonyos p valószínűséggel következik be. Ha a bekövetkezések száma az X diszkrét valószínűségi változó, akkor annak a lehetséges értékei 0 és n közötti egész számok. Annak a valószínűsége, hogy X értéke k legyen, tehát az A esemény k-szor következzen be, bebizonyíthatóan

$$f_X(k,n,p) = P(X=k) = \frac{n!}{k!\cdot(n-k)!}\cdot p^k\cdot(1-p)^{n-k}, \qquad (4.10)$$

ami nem más, mint a diszkrét X valószínűségi változó három paraméterű valószínűségi függvénye. Az eloszlásfüggvény az A esemény legtöbb k alkalommal való bekövetkezésének valószínűségét adja:

$$F_X(k,n,p) = P(X\le k) = \sum_{i=0,k}\left(\frac{n!}{i!\cdot(n-i)!}\cdot p^i\cdot(1-p)^{n-i}\right). \qquad (4.11)$$

Ezt az eloszlást binomiálisnak nevezik, és a visszatevéses mintavételezést írja le.

A binomiális eloszlást követő valószínűségi változó várható értéke

$$m_X = n\cdot p, \qquad (4.12)$$

amit, ha figyelembe vesszük, hogy a diszkrét valószínűségi változó csak egész értékeket vehet fel, kerekítenünk kell. A változó szórása:

$$\sigma_X = \sqrt{n\cdot p\cdot(1-p)}. \qquad (4.13)$$

Az eloszlás ferdesége

$$\beta_{1,X} = (1-2\cdot p)\cdot\sqrt{n\cdot p\cdot(1-p)}. \qquad (4.14)$$

Ha $p\neq 0.5$, a valószínűségi függvény ferde, és a legvalószínűbb érték (a módusz) nem esik egybe a várható értékkel. A móduszt az

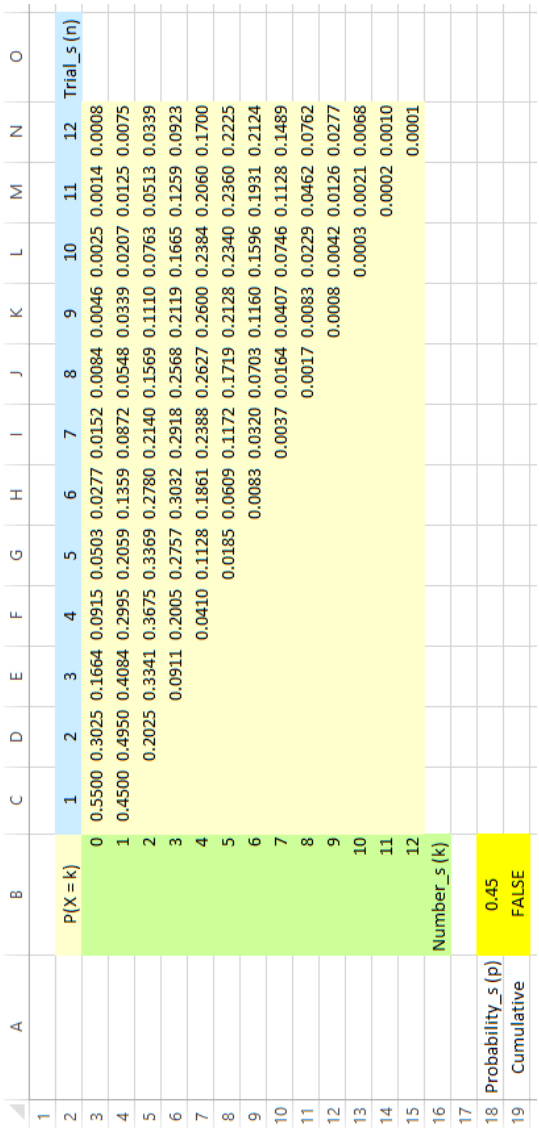
$$\tilde{m}_x = p\cdot(n+1) \qquad (4.15)$$

mennyiség lefelé kerekített értéke adja. Ha ez a mennyiség eleve egy egész szám, akkor két legvalószínűbb érték van, a második 1-gyel kisebb, mint a kiszámolt.

Az eloszlás lapultsága:

$$\beta_{2,X} = \frac{1-6\cdot p\cdot(1-p)}{n\cdot p\cdot(1-p)}. \qquad (4.16)$$

Binomiális eloszlás Excelben



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2		P(X = k)	1	2	3	4	5	6	7	8	9	10	11	12	Trial_s (n)
3		0	0.5500	0.3025	0.1664	0.0915	0.0503	0.0277	0.0152	0.0084	0.0046	0.0025	0.0014	0.0008	
4		1	0.4500	0.4950	0.4084	0.2995	0.2059	0.1359	0.0872	0.0548	0.0339	0.0207	0.0125	0.0075	
5		2		0.2025	0.3341	0.3675	0.3369	0.2780	0.2140	0.1569	0.1110	0.0763	0.0513	0.0339	
6		3			0.0911	0.2005	0.2757	0.3032	0.2918	0.2568	0.2119	0.1665	0.1259	0.0923	
7		4				0.0410	0.1128	0.1861	0.2388	0.2627	0.2600	0.2384	0.2060	0.1700	
8		5					0.0185	0.0609	0.1172	0.1719	0.2128	0.2340	0.2360	0.2225	
9		6						0.0083	0.0320	0.0703	0.1160	0.1596	0.1931	0.2124	
10		7							0.0037	0.0164	0.0407	0.0746	0.1128	0.1489	
11		8								0.0017	0.0083	0.0229	0.0462	0.0762	
12		9									0.0008	0.0042	0.0126	0.0277	
13		10										0.0003	0.0021	0.0068	
14		11											0.0002	0.0010	
15		12												0.0001	
16		Number_s (k)													
17															
18	Probability_s (p)	0.45													
19	Cumulative	FALSE													

4.2. ábra. *Binomiális eloszlás Excelben*

Excelben a statisztikai függvények között szereplő BINOM.DIST() segítségével lehet meghatározni az eloszlás valószínűségi függvényének és eloszlásfüggvényének értékét. A függvénynek négy paramétere van:

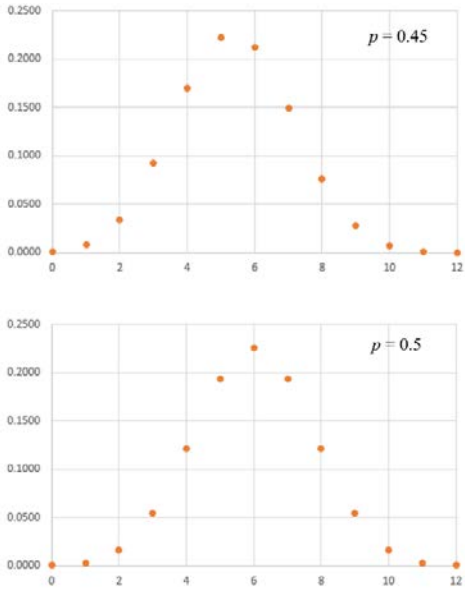
– *Number_s*: ez a k paraméter, amely a bekövetkező események számát rögzíti;

– *Trials*: ez az n paraméter, ami a kísérletek számát jelenti;

– *Probability_s*: ez a p paraméter, ami a megfigyelt esemény bekövetkezésének a valószínűsége;

– *Cumulative*: ennek *TRUE* (vagy 1-es) értéke az eloszlásfüggvény, *FALSE* (vagy 0) a valószínűségi függvény értékeinek kiszámítását állítja be.

Alkalmazásként hozzunk létre egy 13 sor × 12 oszlop méretű táblázatot, ahol vízszintesen a kísérletek n száma (1-től 12-ig a *C2* cellával kezdődően), függőlegesen pedig a bekövetkező események k száma áll (ez utóbbi 0-tól 12-ig, a *B3* cellától lefelé). Jelöljünk ki egy cellát a követett esemény p valószínűségének megadásához (*B18*), és egyet az eloszlásfüggvény vagy a valószínűségi függvény kiválasztásához (*B19*, 4.2. ábra).



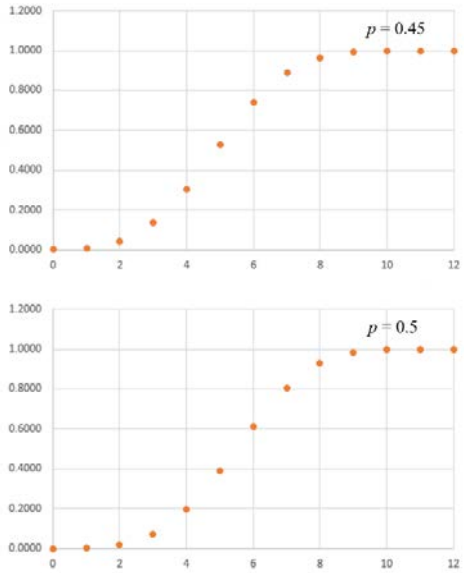
4.3. ábra. *A binomiális eloszlás valószínűségi függvénye*

A 13 × 12-es táblázat bal felső sarkában levő cellájába (*C3*) írjuk be a következő függvényt:

=IF($B3>C$2,"",BINOM.DIST($B3,C$2,$B$18,$B$19)) (*4.17*)

(az IF függvény a $k \le n$ feltétel biztosításához kell), majd másoljuk át a képletet a táblázat többi cellájába. Ekkor a táblázat adatokkal telik meg.

Ha a *Cumulative* értéke *FALSE,* oszloponként annak a valószínűségét kapjuk, hogy n próbálkozásból k esetben forduljon elő az esemény. Ha az utolsó oszlopban levő adatokat ábrázoljuk, akkor a valószínűségi függvény grafikonjához jutunk. p értékét változtatva a grafikon különböző formáihoz jutunk. A 4.3. ábra felső részén például az esemény bekövetkezésének valószínűsége $p = 0.45$, ami egy aszimmetrikus grafikonhoz vezet, míg az alsó, szimmetrikus görbe $p = 0.5$-nek felel meg.



4.4. ábra. *A binomiális eloszlás eloszlásfüggvénye*

Hasonlóképpen kapjuk az eloszlásfüggvényeket a *Cumulative* értéke *TRUE*-ra állításával (ha a táblázat tartalmát jelző *B2* cellába az

=IF(B19,"P(X <= k)","P(X = k)") (*4.18*)

függvényt írjuk be szöveg helyett, akkor a cím automatikusan igazodik a tartalomhoz). A 4.4. ábrán az előbbi két eset eloszlásfüggvénye látható.

A $p = 0.5$ érték az egyenlő valószínűséggel bekövetkező vagy be nem következő eseménynek felel meg, mint például annak, hogy páros számot dobunk egy kockával. A hatos szám megjelenésének tanulmányozásához ezt a valószínűséget 1/6-ra kell, hogy állítsuk. A binomiális eloszlással páros szám, illetve hatos szám dobásának valószínűségét tanulmányozhatjuk a dobások számának (vagy az egyszerre eldobott kockák számának) függvényében. Tizenkét

próbálkozásból a legnagyobb valószínűséggel hat páros szám dobása fordul elő (kb. 22.56%-os valószínűséggel), míg ugyanennyi próbálkozásból a hatos szám kétszeri megjelenése a legvalószínűbb (29.61%-os eséllyel).

Gyakran olyan kérdésre kell válaszolni, hogy egy adott valószínűséggel hányszor (vagy maximum hányszor) fog bekövetkezni a tanulmányozott esemény. Ennek a megválaszolására meg kellene oldani k-ban a valószínűségi függvény (illetve az eloszlásfüggvény) képletével megadott egyenletet. Tekintve e képletek bonyolultságát, ezt leginkább próbálkozással lehetne megoldani.

Az Excelben van egy BINOM.INV() függvény, amelynek a paraméterei

– *Trials*: ez az n paraméter, amely a kísérletek számát jelenti;

– *Probability_s*: ez a p paraméter, amely a megfigyelt esemény bekövetkezésének a valószínűsége;

– *Alpha*: az Excelben „kritérium”-nak nevezett valószínűség, amely felülről határolja a bekövetkező események számát.

Ez a függvény az eloszlásfüggvény alapján azt a legkisebb k számot adja vissza, amelyre $P(X \le k) \ge \alpha$. Az egyenlőtlenség az eloszlás diszkrét mivoltából ered.

A Bernoulli-eloszlás

A Bernoulli-eloszlás a binomiálisnak egy speciális esete, ahol $n=1$ – ebben az esetben k lehetséges értékei 0 és 1.

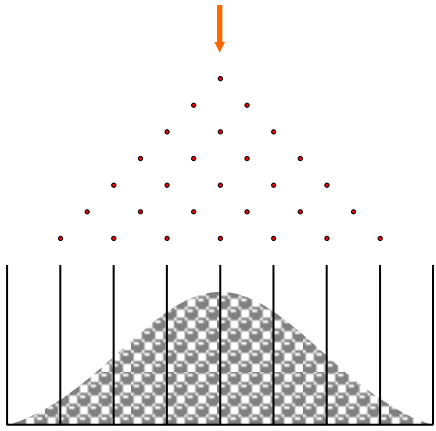
4.3. A normál eloszlás

A normál eloszlást nevezik ugyan még Gauss-félének is, de azt tulajdonképpen Carl Friedrich Gausst megelőzően, a trigonometriai alakban felírt komplex szám k-adik hatványát megadó képlet megalkotójaként is ismert Abraham de Moivre francia matematikus vezette be elsőként.

A legtöbb véletlen jelenség normál eloszlású ugyan, de ez nem jelenti azt, hogy ez kizárólagosságot élvezne: számos véletlen jelenséget leíró valószínűségi változó másfajta eloszlást követ.

Intuitív szemléltetésére Francis Galton, angol polihisztor, egy szerkezetet épített, amelyet Galton-deszkaként ismerünk. Ez tulajdonképpen egy ferdén elhelyezett sík lap, amelybe sakkmintaszerűen szegeket verünk be, egymástól pontos távolságra. A lap alján tartórekeszeket képezünk ki, amelyekben a felülről, ugyanabból a pontból elinduló korongokat vagy golyókat gyűjtjük össze. A korong lefelé csúszása közben beleütközik az első szegbe, ami megakadályozza a további

szabad lecsúszását. Itt tehát az pattan egyet, és véletlenszerűen, egyenlő valószínűséggel a szeg jobb vagy bal oldalán fog tovább csúszni lefelé, a következő szegig. Itt a jobbra vagy balra történő kitérés megismétlődik, és egészen addig folytatódik, míg a korong le nem ér a lap alján levő tartóba. A tartóban levő korongok körvonala egy harangszerű görbét ad, ez pedig nem más, mint a normál eloszlás sűrűségfüggvényének valamilyen megközelítése. Minél több sor szegünk van, és minél több korongot használunk, annál jobb lesz ez a megközelítés. (4.5. ábra).



4.5. ábra. *A Galton-deszka*

Ezen intuitív demonstráció célja a természetben előforduló véletlen események magyarázata volt: minden egyes véletlen esemény (az, hogy melyik tartóba esik a korong) tulajdonképpen sok-sok véletlen kimenetű, legtöbbször megfigyelhetetlen elemi eseményből áll össze, amelyek egy bizonyos valószínűséggel bekövetkeznek vagy sem (tehát hogy a szegre eső korong annak jobb oldalán esik-e tovább vagy sem). Ahogyan a szegeken pattogó korongoknak az ugyanabba a tartóba vezető lehetséges útjainak a száma a szélektől a deszka közepe felé növekedik, úgy a természetben előforduló véletlen események kimenete sem lesz azonos valószínűségű, még akkor sem, ha az elemi „igen-nem” kimenetű események egyenlő valószínűséggel következnének is be.

A Galton-deszka természetesen egy diszkrét valószínűségi eloszlást illusztrál, a címben szereplő normál eloszlás pedig folytonos: ilyenkor egy olyan deszkára kell gondolnunk, amely nagyon magas és megszámlálhatatlanul sok szeg van benne.

A Galton-deszka valójában azt a binomiális eloszlást szemlélteti, ahol a korong egyenlő $p = 0.5$ valószínűséggel kerül a szeg bal, illetve jobb oldalára. Ekkor a 4.2.

ábrán látható táblázatból, balról jobbra haladva annak a valószínűségét olvashatjuk ki, hogy milyen valószínűséggel halad tovább egy korong a szegek között: az első oszlopban megjelenő 1/2, 1/2 értékek azt jelentik, hogy a felső sorban levő szeg a korongokat kb. fele-fele arányban téríti el balra, illetve jobbra. Innen a korongok a második sorban levő szegekre esnek, amelyek szintén fele-fele arányban térítik el azokat balra és jobbra. Ekkor a bal oldali szegre eső korongok fele, tehát az összes korong kb. 1/4-e térül balra, és ugyanannyi jobbra. A jobb oldali szegre eső korongok esetében ugyanez történik, így a két szeg között a korongok kb. 1/4 + 1/4 = 1/2-e fog elhaladni. Ennek megfelelően a táblázat második oszlopában az 1/4, 1/2, 1/4 valószínűségek jelennek meg. Ez sorról sorra lefelé megismétlődik és így alakul ki a haranggörbéhez közelítő eloszlás.

Az elmélet részletesebb taglalása nélkül kijelenthetjük, hogy a Galton-deszka igen-nem elemi eseményeinek valószínűségéből ki lehet számítani a korongok beesésének helyét adó diszkrét valószínűségi változó eloszlását, és ha az egymást követő események száma $n \to \infty$, akkor ez az eloszlás a normál eloszlás felé tart. A folytonos normál eloszlás sűrűségfüggvénye így

$$f_X(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma_X} \cdot \mathrm{e}^{-\frac{1}{2} \cdot \left(\frac{x - m_X}{\sigma_X} \right)^2}, \qquad (4.19)$$

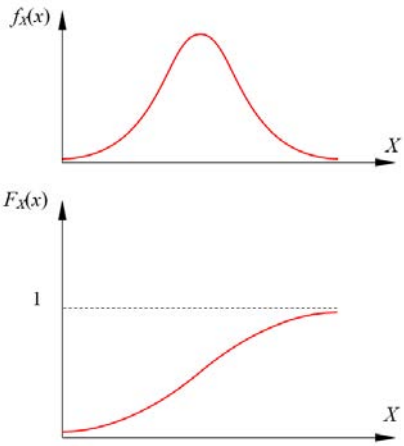
s ennek integrálásával az eloszlásfüggvénye

$$F_X(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma_X} \cdot \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2} \cdot \left(\frac{x - m_X}{\sigma_X} \right)^2} \mathrm{d}x \cdot \qquad (4.20)$$

Ezt az integrált analitikus formában kiszámítani nem lehet.

A normál eloszlást teljesen leírja a középértéke és a szórása. A sűrűségfüggvény ábrázolása a szimmetrikus Gauss-haranggörbét eredményezi (lehet, hogy azt Moivre-haranggörbének kellene nevezni...), amely mindkét végén aszimptotikusan tart a zéró felé. Az eloszlásfüggvény grafikonja egy növekvő, két vízszintes aszimptotával és egy inflexiós ponttal rendelkező görbe, amely az inflexiós pontra nézve szimmetrikus. A sűrűségfüggvény maximuma (a harang legmagasabb pontja) ugyanott van, ahol az eloszlásfüggvénynek az inflexiós pontja van (4.6. ábra).

A szimmetria miatt az átlagos érték, a medián és a módusz egybeesik, és azok a haranggörbe csúcsának, tehát az eloszlásfüggvény inflexiós pontjának felelnek meg. Ugyancsak a szimmetria miatt a ferdeség is nulla, a lapultság pedig $\beta_{2,X} = 3$.



4.6. ábra. *A normál eloszlás sűrűségfüggvénye és eloszlásfüggvénye*

Standard normál eloszlás

A normál eloszlásnak van egy igen fontos sajátságos esete, amikor $m_X = 0$ és $\sigma_X = 1$, ez pedig a standard normál eloszlás. A sűrűségfüggvényét és az eloszlásfüggvényét a normál eloszlás megfelelő függvényeivel, a nulla átlag és az egységnyi szórás behelyettesítésével kapjuk:

$$f_X(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \mathrm{e}^{-\frac{x^2}{2}}, \tag{4.21}$$

s ennek integrálásával az eloszlásfüggvénye

$$F_X(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^{x} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x \cdot \tag{4.22}$$

E függvények értékeit régebb táblázatosan vagy grafikonokon adták meg (akkor még nem voltak vagy nem terjedtek el a számítógépek, és ezeknek a mennyiségeknek a kiszámítása nehézkes volt). A standard normál eloszlás és a közönséges normál eloszlás egy változócserével összeköthető: ha Y standard normál eloszlású, akkor

$$X = m_X + Y \cdot \sigma_X \tag{4.23}$$

a megadott paraméterekkel rendelkező normál eloszlást követi. Avagy, ha X középértéke m_X, és szórása σ_X, akkor

$$Y = \frac{X - m_X}{\sigma_X} \tag{4.24}$$

standard normál eloszlású lesz. A táblázatokból tehát az $x \leftrightarrow y$ megfeleltetéssel ki lehetett olvasni valamely x érték megjelenésének valószínűségét, vagy pedig azt, hogy annál kisebb vagy nagyobb értékek milyen valószínűséggel jelennek meg. Ez sokkal gyorsabb módszer volt, mintsem hogy kézzel számolták volna ki a két függvény értékét.

A normál eloszlás Excelben

Excelben a statisztikai függvények között szereplő NORM.DIST() az eloszlás sűrűségfüggvényének és eloszlásfüggvényének értékét adja vissza. E függvénynek négy paramétere van, ezek:

– *X*: ez a valószínűségi változó azon x értéke, amire a sűrűségfüggvény vagy az eloszlásfüggvény értékét szeretnénk megkapni;

– *Mean*: ez az X valószínűségi változó m_X várható értéke (átlaga);

– *Standard_dev*: ez az X valószínűségi változó σ_X szórása;

– *Cumulative*: *TRUE* (vagy 1-es) értéke az eloszlásfüggvény, *FALSE* (vagy 0) a sűrűségfüggvény értékeinek kiszámítását állítja be.

Alkalmazásként egy oszlopban hozzunk létre egy lefelé növekvő számsort, amely a valószínűségi változó x értékeit tartalmazza. Ha több lehetőséget szeretnénk kipróbálni, akkor ezeket a számokat egy tényezővel skálázhatjuk, amihez az *sk* tényezőt a *B8*-as cellában adjuk meg (4.7. ábra). Ekkor a számsor első tagja a *D3*-as cellában =-5*$B$8, a második az alatta levőben =D3+0.5*$B$8. Ez utóbbi képletet terjesszük ki az oszlop többi cellájára is.

A valószínűségi változó átlagát a *B9*, a szórását a *B10* cellában adjuk meg, míg az eloszlásfüggvény és a sűrűségfüggvény közötti választásához a *B11* cellát használjuk.

Az *E* oszlop felső cellájába (*E3*) írjuk be a következő függvényt:

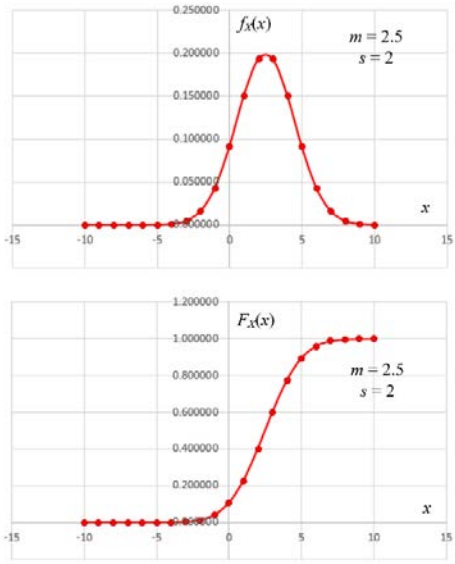
=NORM.DIST(D3,$B$9,$B$10,$B$11), (*4.25*)

majd másoljuk át a képletet az oszlop többi cellájába. Ekkor ez az oszlop adatokkal telik meg.

A 4.8. ábrán az $m_X = 2.5$ átlagú és $\sigma_X = 2$ szórású normál eloszlás sűrűségfüggvénye és eloszlásfüggvénye látható.

	A	B	C	D	E
1					
2				x	P(X = x)
3				-10	0.000000
4				-9	0.000000
5				-8	0.000000
6				-7	0.000003
7				-6	0.000024
8	sk	2		-5	0.000176
9	mean (m)	2.5		-4	0.001015
10	standard_dev (s)	2		-3	0.004547
11	cumulative	FALSE		-2	0.015870
12				-1	0.043139
13				0	0.091325
14				1	0.150569
15				2	0.193334
16				3	0.193334
17				4	0.150569
18				5	0.091325
19				6	0.043139
20				7	0.015870
21				8	0.004547
22				9	0.001015
23				10	0.000176
24					

4.7. ábra. *Normál eloszlás Excelben*



4.8. ábra. *A normál eloszlás sűrűségfüggvénye és eloszlásfüggvénye Excelben*

Az X valószínűségi változónak az eloszlásfüggvény egy bizonyos értékéhez tartozó értékét a NORM.INV() függvénnyel lehet kiszámítani, amelynek a paraméterei a következők:

– *Probability*: ez a $P*$ valószínűség, amely felülről határolja a valószínűségi változó értékét;
– *Mean*: az X valószínűségi változó átlaga;
– *Standard_dev*: X szórása.

Ez a függvény azt az x értéket adja, amelyre $P(X \le x) = P*$.

Az Excelben a NORM.S.DIST() és a NORM.S.INV() függvények hasonlóképpen működnek, a standard normál eloszlás esetében használhatók. Mivel ez esetben az átlag és a szórás definíció szerint adott (0, illetve 1), e két mennyiség nem szerepel e függvények paraméterei között.

Normál eloszlású számsor számítógépes létrehozása

Normál eloszlású pszeudovéletlen számokat számítógéppel is elő lehet állítani. E téren gyakran használt algoritmus a Box–Müller-eljárás. Elméleti alapján bizonyos valószínűségi összefüggések állnak, a lényege pedig a következő két műveletből áll: ha rnd_1 és rnd_2 két standard egyenletes eloszlású (pszeudo)véletlen szám, akkor

$$y_1 = \sqrt{-2 \cdot \ln rnd_1} \cdot \cos(2 \cdot \pi \cdot rnd_2) \qquad (4.26)$$

és

$$y_2 = \sqrt{-2 \cdot \ln rnd_1} \cdot \sin(2 \cdot \pi \cdot rnd_2) \qquad (4.27)$$

standard normál eloszlású (pszeudo)véletlen szám lesz. Amennyiben m_X átlaggal és σ_X szórással rendelkező normál eloszlású véletlen számokra van szükségünk, akkor az előbbi két lépést ki kell egészítenünk a következő módon:

$$x_1 = m_X + y_1 \cdot \sigma_X \text{ és } x_2 = m_X + y_2 \cdot \sigma_X . \qquad (4.28)$$

4.4. A lognormál eloszlás

A valószínűségi változó eloszlása lognormális, ha logaritmusának eloszlása normális. Más szavakkal: ha Y normál eloszlást követ, akkor $X = \mathrm{e}^Y$ lognormál eloszlású valószínűségi változó (ekkor $Y = \ln X$). Ilyen eloszlásra akkor számíthatunk, amikor a véletlen eseményünket normál eloszlást mutató elemi

események szorzatának tekinthetjük. A szakirodalom szerint ilyen eloszlást követnek például a hidrológiai adatok (havi, évi csapadékmaximum, folyók vízhozamának maximuma), a települések mérete, a lakosság jövedelme, a szerkezetek, szerkezeti elemek élettartama stb.

Definíciója szerint az eloszlásfüggvényét a következő módon írhatjuk fel:

$$F_X(x) = \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_{\ln X}} \cdot \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln x - m_{\ln X}}{\sigma_{\ln X}}\right)^2} \mathrm{d}(\ln x) = \\ = \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_{\ln X}} \cdot \int_{0}^{x} \frac{1}{x} \cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln x - m_{\ln X}}{\sigma_{\ln X}}\right)^2} \mathrm{d}x. \qquad (4.29)$$

A sűrűségfüggvény az előbbi deriváltja:

$$f_X(x) = \frac{1}{x} \cdot \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma_{\ln X}} \cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln x - m_{\ln X}}{\sigma_{\ln X}}\right)^2}. \qquad (4.30)$$

Ha ez utóbbi függvényt a normál eloszlás sűrűségfüggvényének közvetlen felhasználásával írtuk volna fel (a változó átírásával), akkor kimaradt volna a változó deriváltját jelentő $1/x$ tag (e derivált a normál eloszlás esetén egységnyi).

Észrevehetjük, hogy mivel a természetes logaritmusfüggvénynek csak pozitív argumentuma lehet, a lognormál eloszlást követő változó csak pozitív értékeket vehet fel.

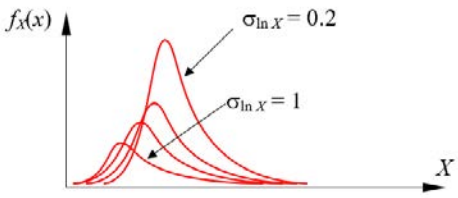
Azt is észrevehetjük, hogy az eloszlást az $m_{\ln X}$ átlag és a $\sigma_{\ln X}$ szórás (a valószínűségi változó természetes logaritmusának átlagáról és szórásáról van szó, és nem pl. az átlag logaritmusáról) egyértelműen meghatározza.

Bebizonyítható, hogy az X és a $\ln X$ valószínűségi változók eloszlásának jellemzői között a következő összefüggések léteznek:

$$m_X = \mathrm{e}^{m_{\ln X} + \sigma_{\ln X}^2/2} \quad \leftrightarrow \quad m_{\ln X} = \ln \frac{m_X}{\sqrt{1+V_X^2}} \approx \ln m_X, \qquad (4.31)$$

$$V_X = \frac{\sigma_X}{m_X} = \sqrt{\mathrm{e}^{\sigma_{\ln X}^2} - 1} \quad \leftrightarrow \quad \sigma_{\ln X} = \sqrt{\ln\left(1+V_X^2\right)} \approx V_X. \qquad (4.32)$$

A sűrűségfüggvény a normál eloszlástól eltérően ez esetben aszimmetrikus, a görbe csúcsa bal kéz felől esik (pozitív aszimmetria, 4.9. ábra). Az aszimmetria $\sigma_{\ln X}$ növekedésével fokozódik, ugyanakkor a sűrűségfüggvény egyre laposabb lesz (de az eloszlás leptokurtózist mutat, tehát csúcsosabb a standard normál eloszlásnál).



4.9. ábra. *Lognormális eloszlás sűrűségfüggvényei*

Az eloszlás módusza

$$\tilde{m}_X = \mathrm{e}^{m_{\ln X} - \sigma^2_{\ln X}}, \tag{4.33}$$

a mediánja pedig

$$\breve{m}_X = \mathrm{e}^{m_{\ln X}}. \tag{4.34}$$

Lognormális eloszlás létrehozása normál eloszlás alapján

Az

$$Y = \frac{\ln X - m_{\ln X}}{\sigma_{\ln X}} \tag{4.35}$$

átalakítással a lognormális eloszlású valószínűségi változó standard normál eloszlású változóvá transzformálható. Ebből következően, ha standard normál eloszlású y értékekből lognormális eloszlású x értékeket kell előállítanunk, akkor azt az

$$x = \mathrm{e}^{y \cdot \sigma_{\ln X} + m_{\ln X}} \tag{4.36}$$

művelettel tehetjük meg.

Lognormális eloszlás Excelben

A normál eloszláshoz hasonlóan, a lognormális eloszlású valószínűségi változókhoz Excelben a LOGNORM.DIST() és LOGNORM.INV() függvényeket használhatjuk, a normál eloszlás függvényeivel azonos módon. Megjegyzendő, hogy a függvény paraméterei között szereplő átlag és szórás a változó természetes logaritmusának $m_{\ln X}$ átlaga és $\sigma_{\ln X}$ szórása.

4.5. A hipergeometrikus eloszlás

A binomiális eloszlás a visszatevéses mintavételezést írja le, amikor a megfigyelt események egymástól függetlenül és azonos valószínűséggel következnek be. Példaképpen: amikor egy fehér és fekete golyókat tartalmazó urnából találomra kihúzunk egyet, és azt a következő húzás előtt visszahelyezzük az urnába, a fehér és a fekete golyók megjelenése ilyen eloszlást követ. Ha ugyanezt a kísérletet úgy hajtjuk végre, hogy a kihúzott golyókat nem helyezzük vissza, akkor minden egyes golyó kihúzása után a fehér és a fekete szín megjelenésének a valószínűsége megváltozik, a már kivett golyók színének függvényében. Az utolsó golyó színe biztosan meghatározható. A visszatevés nélküli mintavételezést a hipergeometrikus (avagy hipergeometriai) eloszlás írja le.

A visszatevés nélküli mintavételezés során jelölje N az elemek számát, amelyek közül n elemet választunk ki a mintavételezés során. Az N elem közül M rendelkezik azzal az adott tulajdonsággal, amelynek a megjelenését (az A esemény bekövetkezését) a mintavételezés során megfigyeljük. Annak a valószínűsége, hogy az A esemény k -szor következzen be,

$$f_X(k,N,n,M)=P(X=k)=\frac{C_M^k\cdot C_{N-M}^{n-k}}{C_N^n}=\frac{\binom{M}{k}\cdot\binom{N-M}{n-k}}{\binom{N}{n}}, \qquad (4.37)$$

ami a szakirodalomban rendszerint a binomiális együtthatókkal felírt formájában fordul elő. Ez a diszkrét X valószínűségi változó négy paraméterű valószínűségi függvénye. Az eloszlásfüggvény, amely az A esemény legtöbb k alkalommal való bekövetkezésének számát adja:

$$F_X(k,N,n,M)=P(X\le k)=\frac{\sum\limits_{i=0,k}\left(C_M^i\cdot C_{N-M}^{n-i}\right)}{C_N^n}. \qquad (4.38)$$

Az X változó várható értéke

$$m_X=\frac{n\cdot M}{N}, \qquad (4.39)$$

a szórása pedig

$$\sigma_X=\sqrt{\frac{n\cdot M}{N}\cdot\frac{N-n}{N-1}\cdot\frac{N-M}{N}}. \qquad (4.40)$$

A hipergeometrikus eloszlás ferdeségét és a lapultságát komplikált kifejezések írják le.

Amennyiben N nagyon nagy a mintához képest, a visszatevéses valószínűség nem sokkal különbözik a visszatevés nélkülihez: ilyen módon a binomiális eloszlás a hipergeometrikus határesetének tekinthető. Ugyanilyen módon, ha $N \to \infty$, akkor a diszkrét hipergeometrikus eloszlás a folytonos normál eloszlással mosódik egybe.

Hipergeometrikus eloszlás Excelben

A hipergeometrikus eloszlás tanulmányozására az Excelben a HYPGEOM.DIST() függvény alkalmas, amelynek öt paramétere van:

– *Sample_s*: a k paraméter, a bekövetkező események száma;

– *Number_sample*: az n paraméter, a kísérletek száma;

– *Population_s*: az M paraméter, a populáció azon elemeinek a száma, amelyek a megfigyelt tulajdonsággal rendelkeznek;

– *Number_pop*: az N paraméter, a populáció elemeinek teljes száma;

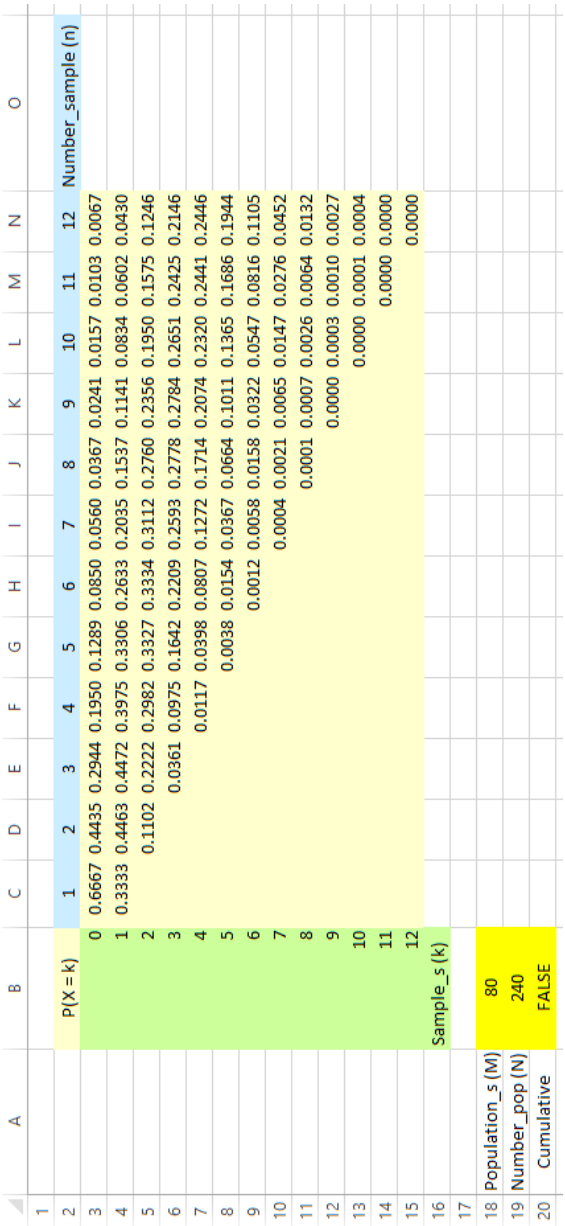
– *Cumulative*: ennek *TRUE* (vagy 1-es) értéke az eloszlásfüggvény, *FALSE* (vagy 0) a valószínűségi függvény értékeinek kiszámítását állítja be.

A kipróbáláshoz, a binomiális eloszlás esetéhez hasonló módon hozzunk létre egy 13 sor × 12 oszlop méretű táblázatot, ahol vízszintesen a kísérletek n száma, függőlegesen pedig a bekövetkező események k száma áll. Jelöljünk ki egy-egy cellát a követett esemény teljes populációbeli M számának (*B18*), a populáció N számának (*B19*) és egyet az eloszlásfüggvény vagy a valószínűségi függvény kiválasztásához (*B20*, 4.10. ábra). A táblázat bal felső sarkába írjuk be az

```
=IF($B3>C$2,"",HYPGEOM.DIST
($B3,C$2,$B$18,$B$19,$B$20))
```
(*4.41*)

függvényt, és azt terjesszük ki a teljes táblázatra.

Az ábrán látható eset egy olyan urnával példázható, amelyben összesen 240 golyó van, amelyek közül 80 fehér (az összes golyó 1/3-a), a többi pedig fekete. Az oszlopokban annak a valószínűsége jelenik meg, hogy n kivett golyó közül éppen k legyen fehér.

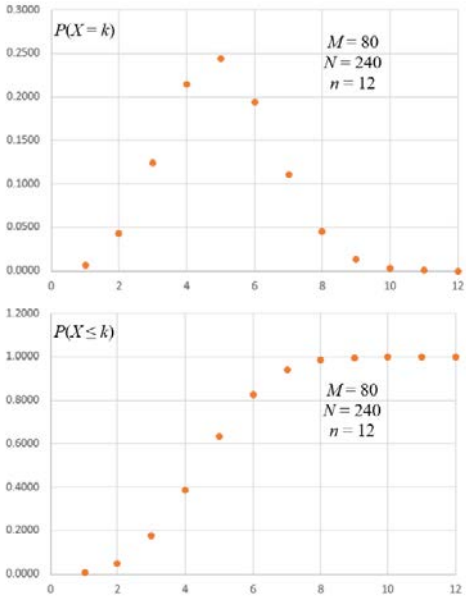


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2		P(X = k)	1	2	3	4	5	6	7	8	9	10	11	12	Number_sample (n)
3		0	0.6667	0.4435	0.2944	0.1950	0.1289	0.0850	0.0560	0.0367	0.0241	0.0157	0.0103	0.0067	
4		1	0.3333	0.4463	0.4472	0.3975	0.3306	0.2633	0.2035	0.1537	0.1141	0.0834	0.0602	0.0430	
5		2		0.1102	0.2222	0.2982	0.3327	0.3334	0.3112	0.2760	0.2356	0.1950	0.1575	0.1246	
6		3			0.0361	0.0975	0.1642	0.2209	0.2593	0.2778	0.2784	0.2651	0.2425	0.2146	
7		4				0.0117	0.0398	0.0807	0.1272	0.1714	0.2074	0.2320	0.2441	0.2446	
8		5					0.0038	0.0154	0.0367	0.0664	0.1011	0.1365	0.1686	0.1944	
9		6						0.0012	0.0058	0.0158	0.0322	0.0547	0.0816	0.1105	
10		7							0.0004	0.0021	0.0065	0.0147	0.0276	0.0452	
11		8								0.0001	0.0007	0.0026	0.0064	0.0132	
12		9									0.0000	0.0003	0.0010	0.0027	
13		10										0.0000	0.0001	0.0004	
14		11											0.0000	0.0000	
15		12												0.0000	
16		Sample_s (k)													
17															
18	Population_s (M)	80													
19	Number_pop (N)	240													
20	Cumulative	FALSE													

4.10. ábra. *Hipergeometrikus eloszlás Excelben*

Észrevehetjük, hogy az első oszlop annak a valószínűségét adja meg, hogy az urnából egyetlen golyót kihúzva, az milyen eséllyel lesz fekete ($k=0$), illetve fehér ($k=1$), mely két valószínűség éppen a fekete, illetve a fehér golyók arányát

adja. Ez az oszlop a binomiális eloszlással azonos értékeket tartalmaz, hiszen ha a kísérlet csak egyetlen minta vételére támaszkodik, akkor az egyaránt tekinthető visszatevésesnek vagy visszatevés nélkülinek.



4.11. ábra. *Hipergeometrikus eloszlás valószínűségi függvénye és eloszlásfüggvénye*

A 4.11. ábrán a táblázatban feltüntetett eset valószínűségi függvényét és eloszlásfüggvényét láthatjuk.

Megjegyzendő, hogy az Excelben nincs ~~HYPGEOM.INV()~~ függvény.

4.6. A Poisson-eloszlás

A Poisson-eloszlás az egy adott idő alatt egyenlő, de kis valószínűséggel bekövetkező, egymástól független események számát írja le. Az időtartam helyett halmazokat, geometriai fogalmakat, pl. távolságot, területet is használhatunk. Ilyen eloszlást követ a radioaktív elemek elbomló atomjainak a száma, egy szerverhez történő hozzáférések száma a nap egy adott szakaszában, vagy pedig csomagoláskor az egy doboz szeg közé bekerülő selejtes darabok száma.

Ez utóbbi példánál maradva, a k darab selejtes szeg dobozba kerülésének a valószínűségét a binomiális eloszlással határozhatnánk meg. Ha azonban feltételezzük, hogy a dobozba kerülő szegek (a minták) n száma nagyon nagy a

selejtekhez képest, akkor a Poisson-eloszlás mint a binomiálisnak az $n \to \infty$-ra számított határértéke a dobozba kerülő selejtek k számát a következőképpen adja:

$$f_X(k,\lambda) = P(X = k) = \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda}. \qquad (4.42)$$

ahol λ a selejtek számának a hányada, ami nem más, mint az X valószínűségi változó várható értéke:

$$m_X = \lambda. \qquad (4.43)$$

Másképpen, ha az esemény bekövetkezésének valószínűsége $p \to 0$, a minták száma pedig $n \to \infty$, akkor a $\lambda = n \cdot p$ véges szám az eloszlás paramétere. Az eloszlásfüggvény a következő:

$$F_X(k,\lambda) = P(X \le k) = \mathrm{e}^{-\lambda} \cdot \sum_{i=0,k} \left(\frac{\lambda^i}{i!} \right), \qquad (4.44)$$

a szórás pedig

$$\sigma_X = \sqrt{\lambda}. \qquad (4.45)$$

A valószínűségi függvény ferdesége $\beta_{1,X} = 1/\sqrt{\lambda}$, lapultsága pedig $\beta_{2.X} = 1/\lambda$.

	A	B	C	D	E
1					
2				X (k)	P(X <= k)
3				0	0.006738
4				1	0.040428
5				2	0.124652
6				3	0.265026
7				4	0.440493
8	Mean (lambda)	5		5	0.615961
9	cumulative	TRUE		6	0.762183
10				7	0.866628
11				8	0.931906
12				9	0.968172
13				10	0.986305
14				11	0.994547
15				12	0.997981
16				13	0.999302
17				14	0.999774
18				15	0.999931
19				16	0.999980
20				17	0.999995
21				18	0.999999
22				19	1.000000
23				20	1.000000

4.12. ábra. *Poisson-eloszlás Excelben*

Poisson-eloszlás Excelben

Ez esetben Excelben a POISSON.DIST() függvényt használjuk, aminek a paraméterei:

– *X*: a k paraméter, a bekövetkező események száma;

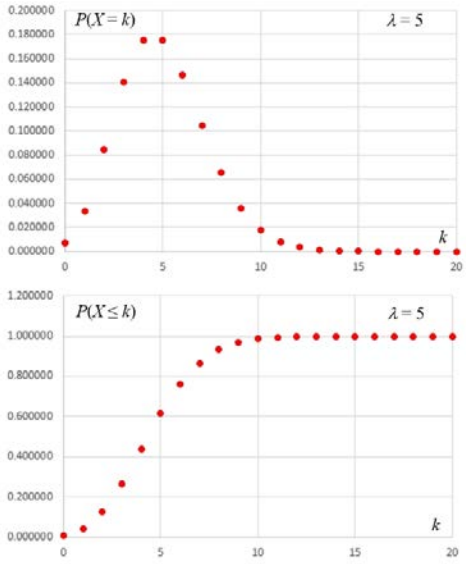
– *Mean*: a λ paraméter, a bekövetkező események számának várható értéke;

– *Cumulative*: ennek *TRUE* (vagy 1-es) értéke az eloszlásfüggvény, *FALSE* (vagy 0) a valószínűségi függvény értékeinek kiszámítását állítja be.

A kipróbáláshoz hozzuk létre a 4.12. ábrán látható táblázatot. A λ paraméterhez használjuk a *B8*-as cellát, a valószínűségi függvény és az eloszlásfüggvény közötti átváltáshoz pedig a *B9*-est. A k értékeket a *D* oszlopba írjuk be. Az *E* oszlop felső cellájába írjuk be a

=POISSON.DIST(D3,$B$8,$B$9) (*4.46*)

függvényt, majd másoljuk át az oszlop többi cellájába. Ha a táblázatban levő adatokat függvényként ábrázoljuk, akkor a valószínűségi függvény, illetve az eloszlásfüggvény grafikonjához jutunk (a 4.13. ábrán a paraméter $\lambda = 5$ értékének megfelelő adatok alapján).



4.13. ábra. *A Poisson-eloszlás valószínűségi függvénye és eloszlásfüggvénye*

4.7. Az exponenciális eloszlás

Tegyük fel, hogy egy megfigyelt jelenség tanulmányozása során egy véletlen ritka esemény bekövetkezéseit figyeljük. Ezeknek a számát a Poisson-eloszlás segítségével írhatjuk le. Az egymást követő események között eltelt időtartam szintén egy valószínűségi változó, amely exponenciális eloszlást mutat. Az ilyen fajta véletlen lefolyású jelenség egy sztochasztikus folyamat, amelyet Poisson-folyamatnak neveznek. Ilyen jelenség például a radioaktív anyagok bomlása, de ezt az eloszlást használjuk a különböző alkatrészek élettartamának meghatározásakor vagy pedig várakozási idővel kapcsolatos problémák megoldásakor is.

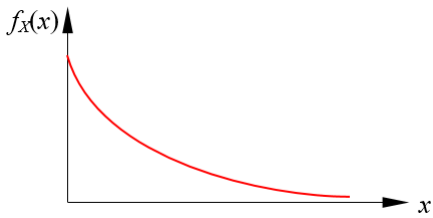
Az exponenciális eloszlás folytonos (a Poisson-eloszlás diszkrét), és ha azt a valós számok halmazán értelmezzük, akkor a sűrűségfüggvénye:

$$f_X(x) = \begin{cases} \lambda \cdot \mathrm{e}^{-\lambda \cdot x} & \text{ha } x \geq 0, \\ 0 & \text{ha } x < 0. \end{cases} \qquad (4.47)$$

a megfelelő eloszlásfüggvény pedig:

$$F_X(x) = \begin{cases} 1 - \mathrm{e}^{-\lambda \cdot x} & \text{ha } x \geq 0, \\ 0 & \text{ha } x < 0. \end{cases} \qquad (4.48)$$

Ez az eloszlás szélsőségesen aszimmetrikus, módusza (legnagyobb értéke) az origóban van, amely éppen λ-val egyenlő (4.14. ábra).



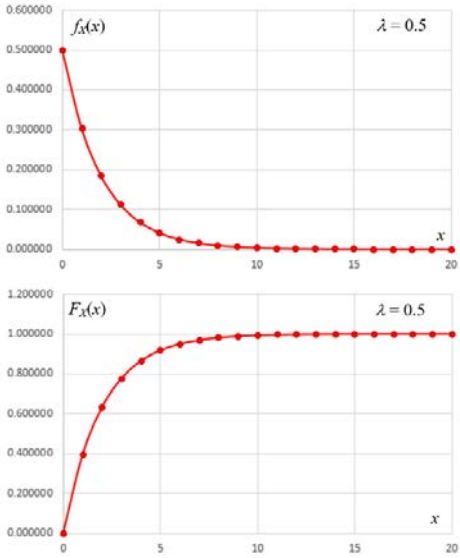
4.14. ábra. *Az exponenciális eloszlás sűrűségfüggvénye*

Az exponenciális eloszlást követő változó átlagos értéke $m_X = 1/\lambda$, szórása pedig szintén ugyanannyi: $\sigma_X = 1/\lambda$. Ezek szerint (és a sűrűségfüggvény definíciójának képlete szerint is) az eloszlás egyetlen paramétere a λ tényező.

Exponenciális eloszlás Excelben

Excelben az EXPON.DIST() függvényt használjuk, amelynek a paraméterei:

– *X*: az x változó értéke;
– *Lambda*: a λ paraméter, a várható érték inverze;
– *Cumulative*: ennek *TRUE* (vagy 1-es) értéke az eloszlásfüggvény, *FALSE* (vagy 0) a sűrűségfüggvény értékeinek kiszámítását állítja be.



4.15. ábra. *Az exponenciális eloszlás sűrűségfüggvénye és eloszlásfüggvénye Excelben*

A kipróbáláshoz hozzunk létre egy táblázatot, a Poisson-eloszlás esetéhez hasonló módon. Az *E* oszlop kitöltéséhez ezúttal használjuk az

=EXPON.DIST(D3,$B$8,$B$9) (*4.49*)

függvényt. A paraméter $\lambda = 0.5$ értékére a 4.15. ábrán látható sűrűségfüggvényhez és az eloszlásfüggvényhez jutunk.

4.8. A geometriai eloszlás

Az exponenciális eloszláshoz hasonlóan a geometriai eloszlás is a várt esemény bekövetkezései közötti történéseket írja le, azonban ez utóbbi a sikeres próbálkozások közötti sikertelen kísérletek számával kapcsolatos. Feltételezzük, hogy a kísérletek során az események egymástól függetlenek, és a várt esemény

bekövetkezésének valószínűsége nem változik meg. Így egy kísérlet során a várt esemény p valószínűséggel következik be, a sikertelenség valószínűsége pedig $q = 1 - p$. Az ilyen fajta kísérletekre a legegyszerűbb példa a pénzfeldobás, amikor pl. a fej megjelenését figyeljük (és itt p is és q is $1/2$, mert mindkét eset azonos eséllyel fordul elő), de ugyanilyen a kockavetés is, amikor a siker lehet a hatos dobása ($p = 1/6$), a sikertelenség pedig az összes többi eset ($q = 5/6$). Ezeket *Bernoulli-kísérlet*nek nevezzük, a kimenetük valószínűségét pedig a már említett Bernoulli-eloszlás mint a binomiális eloszlás legegyszerűbb esete írja le.

A geometriai eloszlást két verzióban is használjuk:

a). a siker eléréséhez szükséges próbálkozások (Bernoulli-kísérletek) X számának leírásához (a siker az X -edik kísérlet során következik be, aminek az értéke legalább 1);

b). a siker elérése előtti próbálkozások Y számának leírásához ($Y = X - 1$).

Mindkét valószínűségi változó diszkrét.

Annak a valószínűsége, hogy a várt esemény az n-edik kísérlet során bekövetkezzen, az X változó valószínűségi függvényét adja:

$$P(x = n) = f_X(n, p) = p \cdot q^{n-1} = p \cdot (1-p)^{n-1}, \qquad (4.50)$$

míg annak a valószínűsége, hogy az n-edik kísérletet követően (tehát az $n+1$-edik során) érjük el a sikert,

$$P(y = n) = f_Y(n, p) = p \cdot q^n = p \cdot (1-p)^n. \qquad (4.51)$$

Az első esetben n legalább 1, a másodikban legalább 0, így a megfelelő eloszlásfüggvények:

$$P(x \le n) = F_X(n, p) = \sum_{i=1,n} p \cdot (1-p)^{i-1}, \qquad (4.52)$$

illetve

$$P(y \le n) = F_Y(n, p) = \sum_{i=0,n} p \cdot (1-p)^i. \qquad (4.53)$$

A várható értékek a két változatban

$$m_X = \frac{1}{p}, \text{ illetve } m_Y = \frac{q}{p} = \frac{1-p}{p}, \qquad (4.54)$$

a szórás pedig mindkét verzióban

$$\sigma_X = \sigma_Y = \frac{\sqrt{q}}{p} = \frac{\sqrt{1-p}}{p}. \qquad (4.55)$$

Geometriai eloszlás Excelben

Az Excelben nincs olyan beépített függvény, ami rögtön alkalmas lenne a geometriai eloszlás tanulmányozására, ezért magunknak kell megoldanunk a függvényértékek kiszámítását.

Ehhez létrehozzuk a 4.16. ábrán látható táblázatot, ahol a *B8*-as cellába a siker bekövetkezésének p valószínűségét kell beírnunk. A *D* oszlopban a siker eléréséhez szükséges kísérletek n számát tartalmazza. Az *E* és *F* oszlopokban az *a*). verzióban értelmezett valószínűségeket számítjuk ki: az *E* oszlopban a valószínűségi függvény, az *F* oszlopban pedig az eloszlásfüggvény értékeit láthatjuk.

	A	B	C	D	E	F
1						
2				n	P(X = n)	P(X <=n)
3				1	0.166667	0.166667
4				2	0.138889	0.305556
5				3	0.115741	0.421296
6				4	0.096451	0.517747
7				5	0.080376	0.598122
8	p	0.166667		6	0.066980	0.665102
9				7	0.055816	0.720918
10				8	0.046514	0.767432
11				9	0.038761	0.806193
12				10	0.032301	0.838494
13				11	0.026918	0.865412
14				12	0.022431	0.887843
15				13	0.018693	0.906536
16				14	0.015577	0.922113
17				15	0.012981	0.935095
18				16	0.010818	0.945912
19				17	0.009015	0.954927
20				18	0.007512	0.962439
21				19	0.006260	0.968699
22				20	0.005217	0.973916

4.16. ábra. *Geometriai eloszlás Excelben*

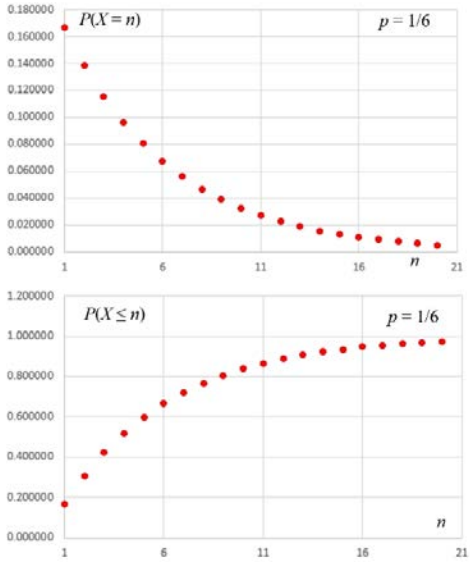
A valószínűségi függvényhez az *E* oszlopban használjuk a

$$=\$B\$8*(1-\$B\$8)\^{}(D3-1) \qquad (4.56)$$

függvényt. Az *F* oszlop felső cellájába másoljuk át az *E* oszlop felső cellájában található értéket (=E3), az alatta levőtől lefelé pedig alkalmazzuk a következő képletet:

=E4+F3. (*4.57*)

Az ábrán látható táblázat értékei a $p = 1/6$ valószínűségnek felelnek meg. Ezt értelmezhetjük például úgy, hogy milyen eséllyel dobunk az n-edik próbálkozás során hatost. A *D* oszlopban például $n = 6$-ra annak a valószínűsége áll (ami kb. 6.7%), hogy éppen a hatodik próbálkozásra sikerüljön a hatos dobása, az *E*-ben pedig az, hogy milyen valószínűséggel gurítunk hat egymás utáni próbálkozás alatt hatost (erre pedig 66.5%-os esélyünk van). A kiszámított értékek alapján megrajzolt valószínűségi függvény és eloszlásfüggvény a 4.17. ábrán látható.



4.17. ábra. *A geometriai eloszlás valószínűségi függvénye és eloszlásfüggvénye Excelben*

4.9. A szélsőértékek eloszlása

A szélsőértékek (maximumok, minimumok) eloszlását a Fisher–Tippet-eloszlással, annak különböző sajátságos eseteivel szokták leírni. A sajátságos esetek a Gumbel-, a Fréchet- és a Weibull-eloszlás, a Fisher–Tippet-eloszlást pedig ezek közös nevezőjeként a *szélsőértékek általánosított eloszlása*ként emlegetik. Az általánosítás az eloszlásfüggvényt célozza meg, amit három paraméter felhasználásával lehet a megfigyelésből származó eredményekhez igazítani:

$$F_X(x,\sigma,\mu,\xi)=\begin{cases} \mathrm{e}^{\mathrm{e}^{-s}} & \text{ha } \xi=0, \\ \mathrm{e}^{-(1+\xi\cdot s)^{-1/\xi}} & \text{ha } \xi\neq 0 \text{ és } \xi\cdot s>-1, \\ 0 & \text{ha } \xi>0 \text{ és } \xi\cdot s\leq -1, \\ 1 & \text{ha } \xi<0 \text{ és } \xi\cdot s\leq -1, \end{cases} \quad s=\frac{x-\mu}{\sigma}, \quad \sigma>0. \qquad (4.58)$$

A három paraméter az eloszlás (empirikus) lokalizációs paramétere (μ, az eloszlás móduszát, tehát a sűrűségfüggvény csúcsának helyét állítja be), skálatényezője (σ, az eloszlás aszimmetriáját befolyásoló paraméter), valamint az alaktényezője (ξ, ami a görbe alakját módosítja).

A megfelelő sűrűségfüggvények:

$$f_X(x,\sigma,\mu,\xi)=\begin{cases} \mathrm{e}^{-s}\cdot\mathrm{e}^{-\mathrm{e}^{-s}} & \text{ha } \xi=0, \\ (1+\xi\cdot s)^{-(1+1/\xi)}\cdot\mathrm{e}^{-(1+\xi\cdot s)^{-1/\xi}} & \text{ha } \xi\neq 0 \text{ és } \xi\cdot s>-1, \\ 0 & \text{egyébként.} \end{cases} \qquad (4.59)$$

A három paraméter megválasztásával az eloszlást a kísérletekből származó adatokhoz lehet igazítani.

A Gumbel-eloszlás

Az I. típusú Fisher–Tippet-eloszlás a Gumbel-eloszlás, ami a $\xi=0$ eset. Ennek sűrűségfüggvénye

$$f_X(x)=\frac{\mathrm{z}\cdot\mathrm{e}^{-z}}{\sigma}, \text{ ahol } \mathrm{z}=\mathrm{e}^{-\frac{x-\mu}{\sigma}},\ x\in(-\infty,+\infty), \qquad (4.60)$$

eloszlásfüggvénye pedig

$$F_X(x)=\mathrm{e}^{-z}. \qquad (4.61)$$

A Gumbel-eloszlás paraméterei tehát μ és σ, amiket a valószínűségi változó megfigyelt értékeinek segítségével határozhatunk meg. E paraméterekkel az átlagos érték $m_X=\mu+\sigma\cdot\gamma$, ahol $\gamma\approx 0.57721$ az Euler–Mascheroni-állandó, a szórás pedig $\sigma_X=\pi\cdot\sigma/\sqrt{6}$. A ferdeségi tényező a paraméterektől függetlenül $\sqrt{\beta_{1,X}}\simeq 1.139$, a sűrűségfüggvénynek a bal oldalon van a maximuma.

Megemlíthető, hogy amennyiben Y standard normál eloszlású valószínűségi változó, akkor $X=\mu-\sigma\cdot\ln(-\ln Y)$ az adott paraméterekkel rendelkező Gumbel-eloszlást fogja követni.

Negatív Gumbel-eloszlás

Az előbbi formában a Gumbel-eloszlást a maximumok eloszlására szokták használni. Amennyiben minimumokra kellene alkalmazni, akkor azt a megfordított előjelű $-X$ változóra kell felírni, a

$$F_X(x) = 1 - F_X(-x) \tag{4.62}$$

szimmetriatulajdonság felhasználásával. Az eloszlásfüggvény ekkor

$$F_X(x) = 1 - \mathrm{e}^{-z}, \tag{4.63}$$

a megfelelő sűrűségfüggvény pedig ez esetben is

$$f_X(x) = \frac{z \cdot \mathrm{e}^{-z}}{\sigma}, \tag{4.64}$$

azonban most

$$z = \mathrm{e}^{\frac{x-\mu}{\sigma}}. \tag{4.65}$$

A Gumbel-eloszlás Excelben

A Gumbel-eloszláshoz hozzuk létre a 4.18. ábrán látható táblázatot! Ebben a *B8*-as cella tartalmazza az eloszlás μ lokalizációs paraméterét, a *B9*-es pedig annak σ skálatényezőjét. A *D* oszlop tartalmazza a valószínűségi változó egymást követő értékeit (az adott példában –5 és +5 között, 0.5-ös lépéssel).

Az *E* és *H* oszlopokban a z transzformált változó értékeit látjuk, az előbbiben a maximumok eloszlásának modellezésére alkalmas pozitív eloszlásra az

=EXP(-(D3-$B$8)/$B$9), (*4.66*)

az utóbbiban pedig a minimumok esetében használható negatív eloszlásra az

=EXP((D3-$B$8)/$B$9) (*4.67*)

függvénnyel kiszámolva.

Az *F* és *I* oszlopokban a sűrűségfüggvények értékei vannak, az

=(E3/$B$9)*EXP(-E3), (*4.68*)

illetve a

=(H3/$B$9)*EXP(-H3) (*4.69*)

függvényeknek megfelelően.

Végül a *G* és a *J* oszlopokban az eloszlásfüggvény értékei kerültek kiszámításra, az

=EXP(-E3), (*4.70*)

illetve a

$$=1\text{-EXP}(-H3) \qquad (4.71)$$

függvény felhasználásával.

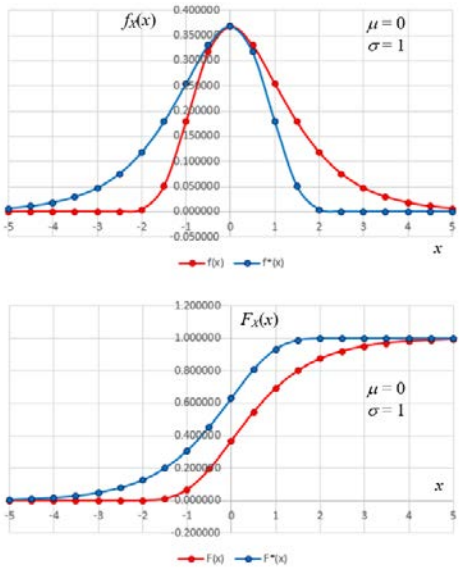
E képletek a felső celláknak felelnek meg, amiket ki kell terjeszteni az oszlop alsóbb celláira is.

	A	B	C	D	E	F	G	H	I	J
1					Maximumra			Minimumra		
2				x	z	f(x)	F(x)	z*	f*(x)	F*(x)
3				-5	148.4132	0.000000	0.000000	0.006738	0.006693	0.006715
4				-4.5	90.01713	0.000000	0.000000	0.011109	0.010986	0.011048
5				-4	54.59815	0.000000	0.000000	0.018316	0.017983	0.018149
6				-3.5	33.11545	0.000000	0.000000	0.030197	0.029299	0.029746
7				-3	20.08554	0.000000	0.000000	0.049787	0.047369	0.048568
8	μ	0		-2.5	12.18249	0.000062	0.000005	0.082085	0.075616	0.078806
9	σ	1		-2	7.389056	0.004566	0.000618	0.135335	0.118205	0.126577
10				-1.5	4.481689	0.050707	0.011314	0.22313	0.178507	0.199989
11				-1	2.718282	0.179374	0.065988	0.367879	0.254646	0.307799
12				-0.5	1.648721	0.317042	0.192296	0.606531	0.330704	0.454761
13				0	1	0.367879	0.367879	1	0.367879	0.632121
14				0.5	0.606531	0.330704	0.545239	1.648721	0.317042	0.807704
15				1	0.367879	0.254646	0.692201	2.718282	0.179374	0.934012
16				1.5	0.22313	0.178507	0.800011	4.481689	0.050707	0.988686
17				2	0.135335	0.118205	0.873423	7.389056	0.004566	0.999382
18				2.5	0.082085	0.075616	0.921194	12.18249	0.000062	0.999995
19				3	0.049787	0.047369	0.951432	20.08554	0.000000	1.000000
20				3.5	0.030197	0.029299	0.970254	33.11545	0.000000	1.000000
21				4	0.018316	0.017983	0.981851	54.59815	0.000000	1.000000
22				4.5	0.011109	0.010986	0.988952	90.01713	0.000000	1.000000
23				5	0.006738	0.006693	0.993285	148.4132	0.000000	1.000000

4.18. ábra. *A Gumbel-eloszlás Excelben*

Az így kapott függvényértékek ábrázolásával a 4.19. ábrán látható sűrűségfüggvényekhez és eloszlásfüggvényekhez jutunk. Ezek tanulmányozása során észrevehetjük, hogy a két eset sűrűségfüggvénye egymás tükörképe (a módusz függőlegesére nézve), az eloszlásfüggvények pedig egymáshoz viszonyítva 180º-kal vannak elforgatva a $(\mu, 0.5)$ pont körül.

A táblázatban és az ábrán a standard Gumbel-eloszlás esete látható; μ és σ értékének megváltoztatásával más esetek, s azokon keresztül a paraméterek hatásai is tanulmányozhatók.



4.19. ábra. *A Gumbel-eloszlás sűrűségfüggvénye és eloszlásfüggvénye (a *-gal jelöltek a negatív eloszlás függvényei)*

A Fréchet-eloszlás

A Fréchet-eloszlás a II. típusú Fisher–Tippet-eloszlás, sűrűségfüggvényét és eloszlásfüggvényét általános formájában az

$$f_X(x) = a \cdot b \cdot (x \cdot b)^{-a-1} \cdot \mathrm{e}^{-b \cdot x^{-a}}, \tag{4.72}$$

$$F_X(x) = \mathrm{e}^{-(b \cdot x)^{-a}} \tag{4.73}$$

kifejezésekkel szokták megadni, az a (vagy ξ) alaktényező és a b $(1/\sigma)$ skálatényező segítségével. Néha egy lokalizációs tényezőt is beiktatnak, akkor a fenti képletekbe „x” helyett „$x-\mu$”-t kell írni (ez a görbét az átlagával, móduszával és mediánjával együtt μ-vel jobbra tolja el). Ebben a formában II. típusú Gumbel-eloszlásnak is mondják azzal a megjegyzéssel, hogy a tulajdonképpeni Fréchet-eloszlás a $b=1$, $\mu=0$ esetnek felel meg. Ezt az eloszlást a maximumok statisztikai leírásához használják.

Esetenként a két paramétert a valószínűségi változó várható értékével és variációs tényezőjével lehet meghatározni, a következő egyenletek megoldásával:

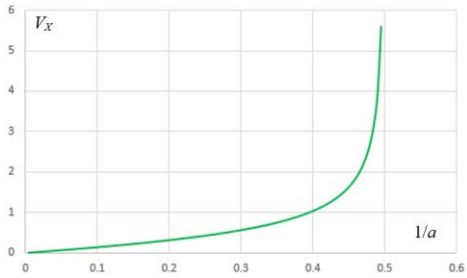
$$\begin{cases} \sqrt{\dfrac{\Gamma(1-2/a)}{\Gamma^2(1-1/a)}-1}=V_X, \\ b=\dfrac{\Gamma(1-1/a)}{m_X}, \end{cases} \tag{4.74}$$

ahol

$$\Gamma(z)=\int_0^\infty t^{z-1}\cdot \mathrm{e}^{-t}\,\mathrm{d}t \tag{4.75}$$

a gamma-függvény. E függvény értéke csak numerikusan számítható ki, amihez Excelben a statisztikai függvények között megtalálható GAMMA()-t lehet használni. A függvényértékek a kézikönyvekben táblázatos formában is megtalálhatók. Az első egyenletből a értékét a variációs tényező függvényében lehet kiszámítani, vagy pedig azt a 4.20. ábrán látható grafikonból lehet kiolvasni.

E grafikonon a könnyebb leolvasás kedvéért az $1/a$ mennyiség szerepel a vízszintes tengelyen. Láthatjuk, hogy $V_X \to 0$ -ra $a \to \infty$, illetve ha $a \to 2$, akkor $V_X \to \infty$, így a valós esetekben a kettőnél nagyobb szám kell, hogy legyen (egyébként 2-re és annál kisebb pozitív számokra is meg lehet alkotni az eloszlás függvényeit).



4.20. ábra. *Grafikon az „a” paraméter meghatározásához*

Ez az eloszlás csak a $(0,+\infty)$ intervallumon értelmezett, úgyhogy a b paraméter képletének nevezőjében szereplő m_X várható érték nem lehet nulla.

A sűrűségfüggvény módusza

$$m_X=\frac{1}{b}\cdot\left(\frac{a}{a+1}\right)^{1/a}. \tag{4.76}$$

A Fréchet-eloszlás Excelben

Ez esetben a sűrűségfüggvény a növekvő x értékekre hirtelen növekedik, majd a maximuma után eléggé meredeken ereszkedik, azonban a jobb oldalon csak lassan tart a nulla felé. Emiatt a tanulmányozás során az x értékeket érdemesebb a bal oldalon kisebb, a jobb oldalon pedig nagyobb lépéssel felvenni (4.21. ábra).

	A	B	C	D	E	F
1						
2				x	f(x)	F(x)
3				0.05	0.000000	0.000000
4				0.2	0.000000	0.000000
5				0.4	0.003155	0.000051
6				0.5	0.098811	0.003493
7				0.6	0.413989	0.027706
8	a	2.5		0.7	0.759884	0.087227
9	b	1		0.8	0.951577	0.174309
10		módusz		0.874075	0.987432	0.246597
11				1	0.919699	0.367879
12				1.1	0.814419	0.454760
13				1.3	0.593954	0.595134
14				1.5	0.420747	0.695665
15				1.7	0.299304	0.766910
16				2	0.185166	0.837967
17				2.5	0.091454	0.903759
18				3	0.050137	0.937864
19				4	0.018930	0.969233
20				5	0.008786	0.982271

4.21. ábra. *Táblázat a Fréchet-eloszláshoz*

A táblázat adatai $a = 2.5$-re (ami a variációs tényező $V_X \simeq 1.034$ értékének felel meg) és $b = 1$-re ($b \cdot m_X \simeq 1.489$) kerültek kiszámításra. A sűrűségfüggvény értékeit a

$$=\$B\$8*\$B\$9*(POWER((\$B\$9*D3),(-\$B\$8-1)))*\\ EXP(-POWER((\$B\$9*D3),(-\$B\$8))), \qquad (4.77)$$

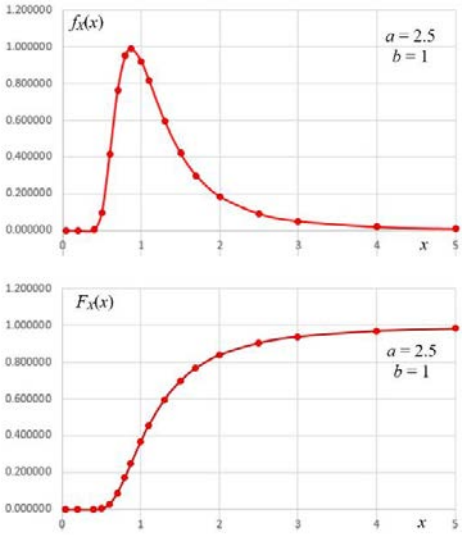
az eloszlásfüggvényét pedig a

$$=EXP(-POWER((\$B\$9*D3),(-\$B\$8))) \qquad (4.78)$$

függvénnyel lehet kiszámítani (ezek a felső sor celláira vonatkoznak), amihez az a paraméter a *B8*, b pedig a *B9* cellából kerül kiolvasásra. Az ábrázoláshoz kiszámoljuk a móduszt is:

$$=(1/\$B\$9)*(\$B\$8/(1+\$B\$8))^(1/\$B\$8) \qquad (4.79)$$

(ez a *D10*-es cellában levő érték), illetve az annak megfelelő függvényértékeket. Az így kapott grafikonok a 4.22. ábrán láthatók.



4.22. ábra. *A Fréchet-eloszlás sűrűségfüggvénye és eloszlásfüggvénye*

A Weibull-eloszlás

A III. típusú Fisher–Tippet-eloszlás a Weibull-féle. Sűrűségfüggvénye

$$f_X(x) = \frac{k}{\lambda} \cdot \left(\frac{x}{\lambda}\right)^{k-1} \cdot \mathrm{e}^{-(x/\lambda)^k}, \qquad (4.80)$$

eloszlásfüggvénye

$$F_X(x) = 1 - \mathrm{e}^{-(x/\lambda)^k}, \qquad (4.81)$$

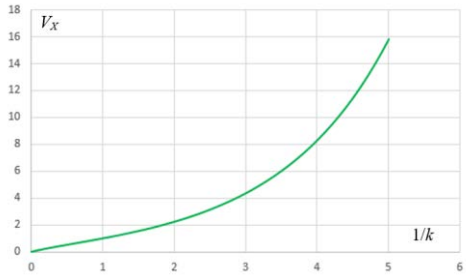
ahol k az alakparaméter (ξ), λ pedig a skálatényező (σ). Az eloszlás értelmezési tartománya a $[0, +\infty)$ intervallum és a minimumok eloszlásának leírásához alkalmazható. Megjegyzendő, hogy $k = 1$-re az exponenciális eloszlás függvényeihez, $k = 2$-re pedig a későbbiekben példázott Rayleigh-eloszlás függvényeihez jutunk.

Ez esetben is be lehet iktatni egy lokalizációs tényezőt a fenti képletekbe „x” helyett „$x - \mu$”-t írván, ami a görbét μ-vel jobbra tolja el.

A két paramétert most is a valószínűségi változó várható értékével és variációs tényezőjével lehet meghatározni:

$$\begin{cases} \sqrt{\dfrac{\Gamma(1+2/k)}{\Gamma^2(1+1/k)}-1}=V_X, \\ \dfrac{1}{\lambda}=\dfrac{\Gamma(1+1/k)}{m_X}, \end{cases} \tag{4.82}$$

ahol $\Gamma(z)$ a már ismert gamma-függvény. Mivel az első összefüggésből k-t analitikusan nem lehet kifejezni, a megoldáshoz valamilyen numerikus módszert kell alkalmaznunk, vagy pedig $V_X(k)$ táblázatos-grafikonos értékei alapján kell interpolálnunk (4.23. ábra). Ha $k \to 0$, akkor $V_X \to \infty$, k nagy értékeire pedig $V_X \to 0$.



4.23. ábra. *Grafikon a „k” paraméter meghatározásához*

Az eloszlás módusza

$$m_X = \lambda \cdot \left(\frac{k-1}{k}\right)^{1/k}. \tag{4.83}$$

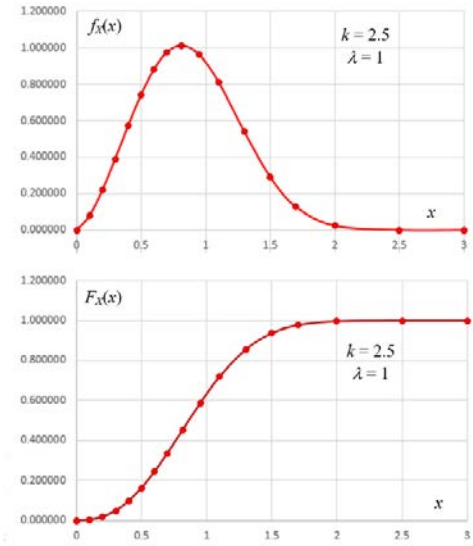
A Weibull-eloszlás Excelben

Ez esetben a WEIBULL.DIST() függvény segítségével könnyebben boldogulhatunk. Paraméterei:

– *X*: az x változó értéke;
– *Alpha*: a k alakparaméter;
– *Beta*: a λ lokalizációs tényező;
– *Cumulative*: ennek *TRUE* (vagy 1-es) értéke az eloszlásfüggvény, *FALSE* (vagy 0) a sűrűségfüggvény értékeinek kiszámítását állítja be.

	A	B	C	D	E	F
1						
2				x	f(x)	F(x)
3				0	0.000000	0.000000
4				0.1	0.078807	0.003157
5				0.2	0.219642	0.017729
6				0.3	0.391033	0.048100
7				0.4	0.571587	0.096241
8	Alpha (k)	2.5		0.5	0.740665	0.162033
9	Beta (lambda)	1		0.6	0.879148	0.243350
10				0.7	0.971722	0.336325
11		módusz		0.815193	1.009843	0.451188
12				0.95	0.960504	0.585071
13				1.1	0.810744	0.718904
14				1.3	0.539529	0.854401
15				1.5	0.291946	0.936434
16				1.7	0.127982	0.976904
17				2	0.024703	0.996507
18				2.5	0.000505	0.999949
19				3	0.000002	1.000000

4.24. ábra. *Táblázat a Weibull-eloszláshoz*



4.25. ábra. *A Weibull-eloszlás sűrűségfüggvénye és eloszlásfüggvénye*

Kipróbálásához hozzuk létre a 4.24. ábrán látható táblázatot, ahol a *B8*-as cella k, a *B9*-es pedig λ értékét tartalmazza. A sűrűségfüggvényhez a *D3*-as cellába írjuk be a

=WEIBULL.DIST(D3,$B$8,$B$9,FALSE), (*4.84*)

az eloszlásfüggvényhez pedig az *E3*-asba a

=WEIBULL.DIST(D3,$B$8,$B$9,TRUE) (*4.85*)

függvényt. Terjesszük ki a képleteket az oszlop többi cellájára. A pontosabb ábrázolás kedvéért számítsuk ki és iktassuk közbe az eloszlás móduszát:

=$B$9*POWER((1-1/$B$8),(1/$B$8)), (*4.86*)

valamint az annak megfelelő függvényértékeket is. Így a 4.25. ábrán levő grafikonokhoz jutunk.

5. TÖBBDIMENZIÓS VALÓSZÍNŰSÉG-ELOSZLÁSOK

Vektorváltozók

Bizonyos esetekben a kísérletileg tanulmányozott jelenség két vagy több valószínűségi változótól is függ, mint

$$Y = Y(X_1, X_2, \dots X_i, \dots X_n)\,. \qquad (5.1)$$

A változók $(X_1, X_2, \dots X_i, \dots X_n)$ halmazát vektorváltozónak is szokták nevezni. Az Y esemény bekövetkezésének valószínűsége, vagyis az, hogy az X_i változók valamilyen x_i értéket szimultán vegyenek fel, szintén valószínűségi változónak tekinthető. Így Y eloszlásfüggvénnyel és sűrűségfüggvénnyel (diszkrét esetben valószínűségi függvénnyel) kell rendelkeznie, és eloszlását ugyanazokkal a jellemzőkkel írhatjuk le, mint az egyszerű valószínűségi változókét. Y-ra azt is mondják, hogy többdimenziós vagy többváltozós valószínűségi változó, az eloszlását pedig az $X_1 \dots X_n$ valószínűségi változók együttes eloszlásaként is említik.

Többváltozós sűrűségfüggvény és eloszlásfüggvény

Ezek szerint, amennyiben a valószínűségi változó folytonos, léteznie kell egy integrálható $f_Y(x_1, x_2, \dots x_n)$ függvénynek, amely sehol sem negatív, és amely egy olyan felületet határoz meg az $n+1$ dimenziós térben, amely alatti térfogat egységnyi:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} f_Y(x_1, x_2, \dots x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \dots \mathrm{d}x_n = 1\,. \qquad (5.2)$$

E feltételek teljesülése esetében $f_Y(x_1, x_2, \dots x_n)$ az Y valószínűségi változó sűrűségfüggvénye lesz.

Akárcsak az egyváltozós esetben, a sűrűségfüggvény integrálja az eloszlásfüggvény:

$$F_Y(x_1, x_2, \dots x_n) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\dots\int_{-\infty}^{x_n} f_Y(x_1, x_2, \dots x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \dots \mathrm{d}x_n\,, \qquad (5.3)$$

amely a következő valószínűséget adja:

$$P(X_1 \le x_1, X_2 \le x_2, \dots X_n \le x_n) = F_Y(x_1, x_2, \dots x_n)\,. \qquad (5.4)$$

Következésképpen az eloszlásfüggvény

$$\frac{\partial^n F_Y(x_1, x_2, \dots x_n)}{\partial x_1 \cdot \partial x_2 \cdot \dots \cdot \partial x_n} = f_Y(x_1, x_2, \dots x_n) \tag{5.5}$$

deriváltja a sűrűségfüggvényt adja vissza.

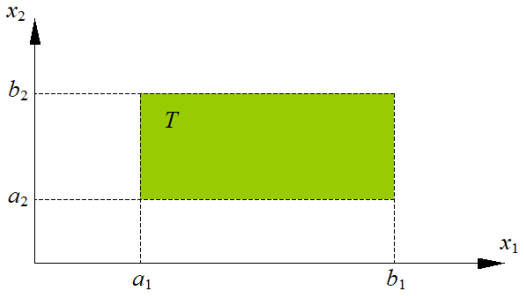
Annak a valószínűsége, hogy a vektorváltozó valamely T tartományon vegyen fel értékeket, a következő kifejezéssel számítható ki:

$$P(a_1 < X_1 \le b_1, a_2 < X_2 \le b_2, \dots a_n < X_n \le b_n) = \\ = \int_{a_1}^{b_1}\int_{a_2}^{b_2}\dots\int_{a_n}^{b_n} f_Y(x_1, x_2, \dots x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \dots \mathrm{d}x_n. \tag{5.6}$$

Megjegyzendő, hogy ez a valószínűség nem egyenlő az $F_Y(b_1, b_2, \dots b_n) - F_Y(a_1, a_2, \dots a_n)$ különbséggel. Példaként vegyük a kétváltozós esetet (5.1. ábra):

$$P(a_1 < X_1 \le b_1, a_2 < X_2 \le b_2) = \int_{a_2}^{b_2}\left(\int_{a_1}^{b_1} f_Y(x_1, x_2)\,\mathrm{d}x_1\right)\mathrm{d}x_2 = \\ = \int_{a_2}^{b_2}\left(\int_{a_1}^{b_1} \frac{\partial^2 F_Y(x_1, x_2)}{\partial x_1 \cdot \partial x_2}\,\mathrm{d}x_1\right)\mathrm{d}x_2 = \int_{a_2}^{b_2}\left(\frac{\partial F_Y(x_1, x_2)}{\partial x_2}\right)\Bigg|_{a_1}^{b_1}\mathrm{d}x_2 = \tag{5.7}$$

$$= \int_{a_2}^{b_2}\left(\frac{\partial F_Y(b_1, x_2)}{\partial x_2} - \frac{\partial F_Y(a_1, x_2)}{\partial x_2}\right)\mathrm{d}x_2 = \left(F_Y(b_1, x_2) - F_Y(a_1, x_2)\right)\Big|_{a_2}^{b_2} = \\ = F_Y(b_1, b_2) - F_Y(a_1, b_2) - F_Y(b_1, a_2) + F_Y(a_1, a_2).$$



5.1. ábra. *Vektorváltozó ($a_1 < X_1 \le b_1$, $a_2 < X_2 \le b_2$) tartománya*

Példánkban a T tartományt egy x_1Ox_2 koordinátasíkban megrajzolt téglalapként ábrázolhatjuk. A kiszámított valószínűség azt jelenti, hogy az X_1 és az X_2 valószínűségi változók szimultán értékeivel felvett pont e téglalap

belsejébe, illetve annak az $X_1 = b_1$, valamint az $X_2 = b_2$ peremére esik (az ábrán a téglalap teteje és annak a jobb oldala).

Az eljárásunkat több dimenzióra is általánosíthatjuk: ha a vektorváltozó n elemű, akkor a T tartomány egy n-dimenziós *hipertéglalap*. Ez egy általánosított geometriai alakzat, amelynek a csúcsokban találkozó oldalai ortogonálisak, az egymással párhuzamos éleik pedig azonos hosszúságúak. Egy hipertéglalapnak 2^n csúcsa (vertexe) van. Négydimenziós hipertéglalapot nem tudunk ábrázolni, viszont a háromdimenziós egy közismert alakzat: ez a téglatest.

Nos, a többváltozós $\int_{a_1}^{b_1}\int_{a_2}^{b_2}\dots\int_{a_n}^{b_n} f_Y(x_1, x_2, \dots x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \dots \mathrm{d}x_n$ integrál kiszámítása ezúttal is változónként történik, tehát pl. először integrálunk x_1 szerint, a b_1 felső és a_1 alsó határértékek behelyettesítésével. Az eredmény két olyan függvény különbsége lesz, amelynek a változói között már nem szerepel az x_1. E két függvény a sűrűségfüggvény $\frac{\partial^{n-1} F_Y(x_2, \dots x_n)}{\partial x_2 \cdot \dots \cdot \partial x_n}$ parciális deriváltja, ahol x_1-et a b_1, illetve az a_1 értékekkel helyettesítettük.

A következő lépésben e két függvény különbségét integráljuk x_2 szerint, az a_2 és a b_2 határok között, aminek az eredménye (mint láthattuk az előbbi példában) négy függvény algebrai összege lesz (két tagja negatív előjelű). Ezek a $\frac{\partial^{n-2} F_Y(x_3, \dots x_n)}{\partial x_3 \cdot \dots \cdot \partial x_n}$ parciális derivált olyan kifejezései, amiket az $x_1 \in \{a_1, b_1\}$, $x_2 \in \{a_2, b_2\}$ halmazok egy-egy elemével képzett értékpár behelyettesítésével kapunk (négy ilyen párunk van).

A harmadik lépésben tehát már négy függvény algebrai összegét integráljuk, ezúttal a $\frac{\partial^{n-3} F_Y(x_4, \dots x_n)}{\partial x_4 \cdot \dots \cdot \partial x_n}$ parciális deriváltba az $x_1 \in \{a_1, b_1\}$, $x_2 \in \{a_2, b_2\}$ és az $x_3 \in \{a_3, b_3\}$ halmazok egy-egy elemével képzett értékhármast kell helyettesítünk (összesen 2^3, tehát nyolc ilyen hármas van). Ha a vektorváltozónk háromelemű, akkor ebben a lépésben a keresett valószínűséget a nyolc behelyettesítési értékkel

$$\begin{aligned} &P(a_1 < X_1 \le b_1, a_2 < X_2 \le b_2, a_3 < X_3 \le b_3) = \\ &= F_Y(b_1, b_2, b_3) - F_Y(a_1, b_2, b_3) - F_Y(b_1, a_2, b_3) + F_Y(a_1, a_2, b_3) - \\ &- F_Y(b_1, b_2, a_3) + F_Y(a_1, b_2, a_3) + F_Y(b_1, a_2, a_3) - F_Y(a_1, a_2, a_3) \end{aligned} \qquad (5.8)$$

gyanánt kapjuk. Az előjelre vonatkozó szabályt is felismerhetjük: az előjelet $(-1)^k$ adja, ahol k a behelyettesített alsó határértékek száma (hány a_i van az argumentumok között).

Ezek után a képletet könnyen általánosíthatjuk:

$$\begin{aligned} P(a_1 < X_1 \le b_1, a_2 < X_2 \le b_2, \dots a_n < X_n \le b_n) = \\ = \int_{a_1}^{b_1}\int_{a_2}^{b_2}\dots\int_{a_n}^{b_n} f_Y(x_1, x_2, \dots x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \dots \mathrm{d}x_n = \\ = \sum_{i=1,2^n} (-1)^{k(i)} \cdot F_Y(c_{i,1}, c_{i,2}, \dots c_{i,n}), \end{aligned} \tag{5.9}$$

ahol az eloszlásfüggvénybe a hipertéglalap i vertexének $(c_{i,1}, c_{i,2}, \dots c_{i,n})$ koordinátáit kell behelyettesíteni (ahol $c_{i,m}$ az a_m és b_m határértékek valamelyike) és $k(i)$ a vertex koordinátái között megjelenő alsó határértékek (a_m tagok) száma.

Amennyiben az integrálási határokat egymáshoz kellőképpen közel vesszük fel,

$$\begin{aligned} P(a_1 - \delta_1/2 < X_1 \le a_1 + \delta_1/2, \dots a_n - \delta_n/2 < X_n \le a_n + \delta_n/2) = \\ = \int_{a_1-\delta_1/2}^{a_1+\delta_1/2} \dots \int_{a_n-\delta_n/2}^{a_n+\delta_n/2} f_Y(x_1, \dots x_n)\,\mathrm{d}x_1 \dots \mathrm{d}x_n \approx f_Y(a_1, a_2, \dots a_n)\cdot\delta_1\cdot\delta_2\cdot\dots\cdot\delta_n, \end{aligned} \tag{5.10}$$

ahol $\delta_1 \cdot \delta_2 \cdot \dots \cdot \delta_n$ az $X_1, X_2, \dots X_n$ változók egyidejűleg felvett $a_1, a_2, \dots a_n$ értékeivel megadott P pont körül felvett hipertéglalap tartománya. A fenti összefüggésben az $f_Y(x_1, x_2, \dots x_n)\cdot\delta_1\cdot\delta_2\cdot\dots\cdot\delta_n$ szorzat tehát annak a valószínűségét közelíti meg, hogy a P pont erre a tartományra essen

E függvények és valószínűségek az X_i változók együttes eloszlását írják le.

Szerkezetméretezési példa

Szemléltetés gyanánt a következő, 5.2. ábrán egy kétváltozós függvény sűrűségfüggvényét mutatjuk be. A megoldandó feladatban az x változó egy szerkezet szilárdságát (R), míg az y a szerkezet terhelését (S) jelenti. Mindkét változó valamilyen eloszlást, például normál eloszlást követő erő.

Mivel e változók egymástól függetlenek, két esemény együttes megjelenésének a valószínűsége az események valószínűségeinek szorzataként írható fel, így például:

$$P(R \le x, S \le y) = P(R \le x)\cdot P(S \le y). \tag{5.11}$$

Ennek alapján a két változó együttes eloszlásfüggvénye a változók eloszlásfüggvényének szorzata:

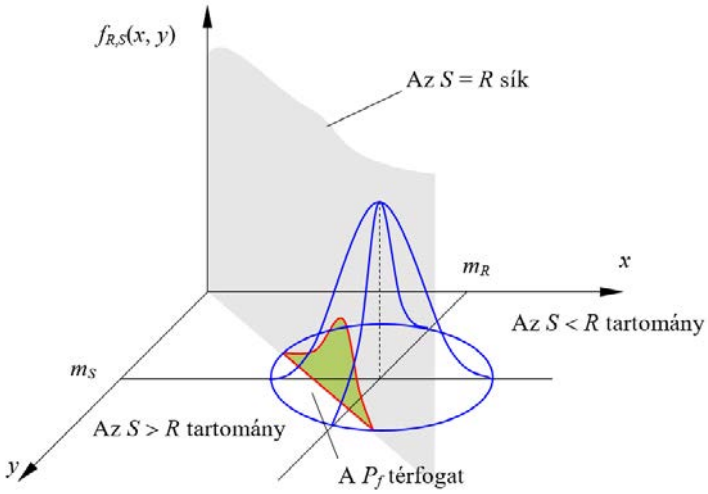
$$F_{R,S}(x,y) = F_R(x) \cdot F_S(y)\,, \tag{5.12}$$

ahonnan következik, hogy

$$\frac{\partial^2 F_{R,S}(x,y)}{\partial x \cdot \partial y} = \frac{\partial^2 \left(F_R(x) \cdot F_S(y)\right)}{\partial x \cdot \partial y} = \frac{\partial F_R(x)}{\partial x} \cdot \frac{\partial F_S(y)}{\partial y}\,, \tag{5.13}$$

tehát az együttes sűrűségfüggvény is az egyszerű sűrűségfüggvények szorzata lesz:

$$f_{R,S}(x,y) = f_R(x) \cdot f_S(y)\,. \tag{5.14}$$



5.2. ábra. *Kétváltozós sűrűségfüggvény*

Az ábrán az $S = R$ sík annak a határesetnek felel meg, amikor a terhelés éppen a szerkezet szilárdságával azonos igénybevételt hoz létre, e síktól balra, ahol a terhelés meghaladja a szerkezet szilárdságát, az minden bizonnyal tönkremegy. E tönkremenetel valószínűségét a harangfelületnek az $S = R$ síkkal lemetszett része alatti térfogat, P_f adja:

$$P_f = \iint\limits_{D_f} f_R(x) \cdot f_S(y)\,\mathrm{d}x\,\mathrm{d}y\,, \tag{5.15}$$

ahol a D_f tartományt az $y = x$ egyenes határolja le:

$$P_f = \int_0^x \int_0^{+\infty} f_R(x) \cdot f_S(y)\, \mathrm{d}x\, \mathrm{d}y \cdot \qquad (5.16)$$

Ha a tönkremenetel valószínűségét Excelben VBA-programozás nélkül akarjuk kiszámítani, akkor a kettős integrál értékét numerikus eljárással kell megközelíteni, egy megfelelően megválasztott Δx és Δy lépéssel kiszámított $f_R(x_i) \cdot f_S(y_j) \cdot \Delta x \cdot \Delta y$ mennyiségek összege gyanánt.

Tegyük fel, hogy mindkét valószínűségi változó normál eloszlású, a terhelés várható értéke $m_S = 3.5\,\mathrm{kN}$, a szilárdságé pedig $m_R = 5.5\,\mathrm{kN}$, a szórások pedig $\sigma_S = 1.25\,\mathrm{kN}$, illetve $\sigma_R = 0.8\,\mathrm{kN}$ (mindezen mennyiségek erőjellegűek). Ekkor a számításokat táblázatos formában az 5.3. ábrán látható módon oldhatjuk meg.

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2					f(x) * f(y)		y	0	0.1	0.2	0.3	0.4
3		S (x)	R (y)				f(y)	2.71782E-11	6.37E-11	1.47E-10	3.34E-10	7.46E-10
4	m	3.5	5.5		x	f(x)						
5	s	1.25	0.8		0	0.006332		1.72102E-13	0	0	0	0
6					0.1	0.007897		2.14624E-13	5.03E-13	0	0	0
7					0.2	0.009785		2.65944E-13	6.23E-13	1.44E-12	0	0
8	Pf	0.094391			0.3	0.012048		3.27434E-13	7.67E-13	1.77E-12	4.02E-12	0
9					0.4	0.014739		4.00569E-13	9.39E-13	2.17E-12	4.92E-12	1.1E-11

5.3. ábra. *A tönkremenetel valószínűségének kiszámítása*

Az *E* oszlopban az x változó (a terhelés), a *2* sorban az y változó (a szilárdság) értékei vannak, $\Delta x = \Delta y = 0.1\,\mathrm{kN}$ lépéssel, $0\,\mathrm{kN}$ és $11\,\mathrm{kN}$ között.

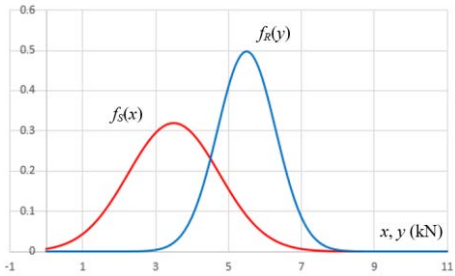
Az *F* oszlop $f_S(x)$, a *3* sor pedig $f_R(y)$ NORM.DIST() függvénnyel kiszámított értékeit tartalmazza. E függvények grafikonját az 5.4. ábrán láthatjuk. A táblázat celláiban e kettő szorzata áll, amennyiben $x \geq y$ (tehát amikor a terhelés eléri vagy meghaladja a szilárdságot). A bal felső (*H5*) cellába beírt függvény

=IF($E5>=H$2,$F5*H$3,0), (5.17)

amelyet át kell másolni a táblázat többi cellájába. A tönkremenetel valószínűségét a cellák tartalmának összegének és a $\Delta x \cdot \Delta y$ területnek a szorzata adja:

=SUM(H5:DN115)*0.1*0.1, (5.18)

ami jelen esetben mintegy 9.4%. Az átlagok és a szórások megváltoztatásával más eseteket is tanulmányozhatunk.



5.4. ábra. *A terhelés (S) és a szilárdság (R) sűrűségfüggvénye*

Az egymástól függő változók esete

Az előbbi kétváltozós példában feltételeztük, hogy a valószínűségi vektorváltozó elemei egymástól függetlenek. Ha ez nem így van, akkor a sűrűségfüggvények és az eloszlásfüggvények közötti összefüggések már nem lesznek éppen ilyen egyszerűek. A szemléltetés kedvéért vegyünk most is egy kétdimenziós példát: legyen X és Y két tetszőleges valószínűségi változó. Emlékezzünk, hogy az $X \le x$ eseménynek az $Y \le y$ eseményre vonatkoztatott feltételes valószínűsége annak az eseménynek a valószínűségét jelenti, hogy $X \le x$ úgy következik be, hogy közben $Y \le y$. Ez a valószínűség

$$P(X \le x \mid Y \le y) = \frac{P(X \le x, Y \le y)}{P(Y \le y)} = \frac{F_{X,Y}(x,y)}{F_Y(y)}, \qquad (5.19)$$

ahol a számlálóban a valószínűségi változók *együttes eloszlásfüggvénye*, a nevezőben pedig Y eloszlásfüggvénye szerepel. Hasonlóképpen írhatjuk fel a

$$P(Y \le y \mid X \le x) = \frac{P(X \le x, Y \le y)}{P(X \le x)} = \frac{F_{X,Y}(x,y)}{F_X(x)} \qquad (5.20)$$

feltételes valószínűséget is, ahol ezúttal a nevezőben az X változó eloszlásfüggvénye fog megjelenni. Ez két feltételes valószínűség legfennebb abban a sajátos esetben egyenlő, amikor $F_X(x) = F_Y(y)$.

Feltételezvén, hogy X és Y a valós számok teljes halmazán értelmezett, az $F_{X,Y}(x,y) = P(X \le x, Y \le y)$ együttes eloszlásfüggvény a következő tulajdonságokkal rendelkezik:

► $$F_X(x) = F_{X,Y}(x,\infty) \text{ és } F_Y(y) = F_{X,Y}(\infty,y), \qquad (5.21)$$

bármely x-re, illetve bármely y-ra. $F_X(x)$, valamint $F_Y(y)$ X és Y együttes eloszlásának a *peremeloszlásai*;

► $$F_{X,Y}(x_0, y_0) = \int_{-\infty}^{y_0} \int_{-\infty}^{x_0} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y\,, \tag{5.22}$$

ahol $f_{X,Y}(x, y)$ az *együttes sűrűségfüggvény*;

► $$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,\mathrm{d}y \text{ és } f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,\mathrm{d}x \tag{5.23}$$

a *peremsűrűségek*;

► $$F_{X,Y}(\infty, \infty) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1\,; \tag{5.24}$$

► $$F_{X,Y}(x, -\infty) = 0 \text{ és } F_{X,Y}(-\infty, y) = 0\,; \tag{5.25}$$

► $$P(x_1 < X \le x_2, y_1 < Y \le y_2) = F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - \\ -F_{X,Y}(x_2, y_1) + F_{X,Y}(x_1, y_1)\,; \tag{5.26}$$

► ha X és Y nem függetlenek, akkor

$$f_{X,Y}(x, y) \neq f_X(x) \cdot f_Y(y) \text{ és } F_{X,Y}(x, y) \neq F_X(x) \cdot F_Y(y)\,. \tag{5.27}$$

A valószínűségeket az $X \le x$ és $Y = y$, illetve az $Y \le y$ és $X = x$ eseményekre is felírhatjuk:

$$P(X \le x \,|\, Y = y) = F_{X,Y}(x \,|\, y)\,, \tag{5.28}$$

$$P(Y \le y \,|\, X = x) = F_{X,Y}(y \,|\, x)\,; \tag{5.29}$$

ezek a megfelelő *feltételes eloszlásfüggvényeket* adják. Ez esetben a számlálóban nem az együttes eloszlásfüggvény szerepel, hanem annak rögzített y-ra, illetve x-re felírt formája, a nevezőben pedig az $f_Y(y)$, illetve az $f_X(x)$ peremeloszlások sűrűségfüggvénye jelenik meg. Bebizonyítható, hogy

$$F_{X,Y}(x \,|\, y) = \frac{\partial F_{X,Y}(x, y)}{\partial y} \cdot \frac{1}{f_Y(y)}\,, \tag{5.30}$$

illetve

$$F_{X,Y}(y \,|\, x) = \frac{\partial F_{X,Y}(x, y)}{\partial x} \cdot \frac{1}{f_X(x)}\,. \tag{5.31}$$

A feltételes eloszlásfüggvények deriváltjai adják a feltételes sűrűségfüggvényeket:

$$f_{X,Y}(x \mid y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{\partial F_{X,Y}(x \mid y)}{\partial x} = \\ = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \cdot \partial y} \cdot \frac{1}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \tag{5.32}$$

illetve

$$f_{X,Y}(y \mid x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{\partial F_{X,Y}(y \mid x)}{\partial y} = \\ = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \cdot \partial y} \cdot \frac{1}{f_X(x)} = \frac{f_{X,Y}(x,y)}{f_X(x)}. \tag{5.33}$$

Alkalmazásként tekintsük az egymástól nem független X és Y valószínűségi változók

$$F_{X,Y}(x,y) = \frac{1}{1 + \mathrm{e}^{-0.5 \cdot x} + \mathrm{e}^{-1.1 \cdot y}} \tag{5.34}$$

együttes sűrűségfüggvényének esetét. E függvény grafikonja az 5.5. ábrán látható. $F_{X,Y}(x,y)$ parciális deriváltjai:

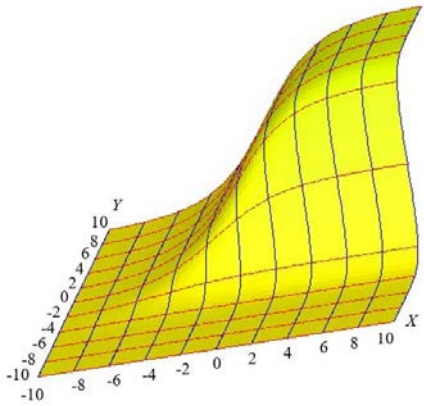
$$\frac{\partial F_{X,Y}(x,y)}{\partial x} = \frac{0.5 \cdot \mathrm{e}^{-0.5 \cdot x}}{\left(1 + \mathrm{e}^{-0.5 \cdot x} + \mathrm{e}^{-1.1 \cdot y}\right)^2}, \tag{5.35}$$

$$\frac{\partial F_{X,Y}(x,y)}{\partial y} = \frac{1.1 \cdot \mathrm{e}^{-1.1 \cdot y}}{\left(1 + \mathrm{e}^{-0.5 \cdot x} + \mathrm{e}^{-1.1 \cdot y}\right)^2}, \tag{5.36}$$

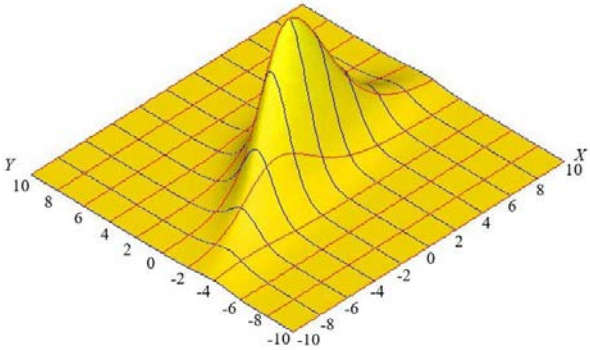
a másodrendű vegyes parciális derivált pedig a két valószínűségi változó együttes sűrűségfüggvényét adja:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \, \partial y} = \frac{1.1 \cdot \mathrm{e}^{-0.5 \cdot x - 1.1 \cdot y}}{\left(1 + \mathrm{e}^{-0.5 \cdot x} + \mathrm{e}^{-1.1 \cdot y}\right)^3} \tag{5.37}$$

(a számlálóban $2 \cdot 0.5 \cdot 1.1 = 1.1$). Ennek a grafikonja az 5.6. ábrán látható.



5.5. ábra. *Az $F_{X,Y}(x,y)$ közös eloszlásfüggvény*



5.6. ábra. *Az $f_{X,Y}(x,y)$ közös sűrűségfüggvény*

A peremeloszlások

$$F_X(x) = F_{X,Y}(x, +\infty) = \frac{1}{1+\mathrm{e}^{-0.5\cdot x}} \tag{5.38}$$

és

$$F_Y(y) = F_{X,Y}(+\infty, y) = \frac{1}{1+\mathrm{e}^{-1.1\cdot y}}, \tag{5.39}$$

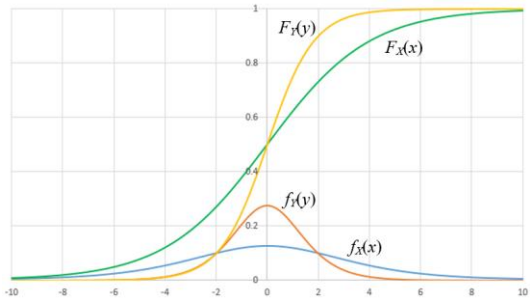
a peremsűrűségek pedig

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \frac{\partial F_{X,Y}(x,y)}{\partial x}\bigg|_{-\infty}^{+\infty} = \frac{0.5\cdot \mathrm{e}^{-0.5\cdot x}}{\left(1+\mathrm{e}^{-0.5\cdot x}\right)^2}, \tag{5.40}$$

illetve

$$f_Y(y)=\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,\mathrm{d}x=\left.\frac{\partial F_{X,Y}(x,y)}{\partial y}\right|_{-\infty}^{+\infty}=\frac{1.1\cdot \mathrm{e}^{-1.1\cdot y}}{\left(1+\mathrm{e}^{-1.1\cdot y}\right)^2} \tag{5.41}$$

(grafikonok az 5.7. ábrán).



5.7. ábra. *A peremsűrűségek és a peremeloszlások*

A fentiek felhasználásával a feltételes eloszlások

$$F_{X,Y}(x\,|\,y)=\frac{\partial F_{X,Y}(x,y)}{\partial y}\cdot\frac{1}{f_Y(y)}=\frac{1+2\cdot \mathrm{e}^{-1.1\cdot y}+\mathrm{e}^{-2.2\cdot y}}{\left(1+\mathrm{e}^{-0.5\cdot x}+\mathrm{e}^{-1.1\cdot y}\right)^2} \tag{5.42}$$

és

$$F_{X,Y}(y\,|\,x)=\frac{\partial F_{X,Y}(x,y)}{\partial x}\cdot\frac{1}{f_X(x)}=\frac{1+2\cdot \mathrm{e}^{-0.5\cdot x}+\mathrm{e}^{-x}}{\left(1+\mathrm{e}^{-0.5\cdot x}+\mathrm{e}^{-1.1\cdot y}\right)^2}, \tag{5.43}$$

a feltételes sűrűségfüggvények pedig

$$f_{X,Y}(x\,|\,y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\frac{\mathrm{e}^{-0.5\cdot x}\cdot\left(1+2\cdot \mathrm{e}^{-1.1\cdot y}+\mathrm{e}^{-2.2\cdot y}\right)}{\left(1+\mathrm{e}^{-0.5\cdot x}+\mathrm{e}^{-1.1\cdot y}\right)^3}, \tag{5.44}$$

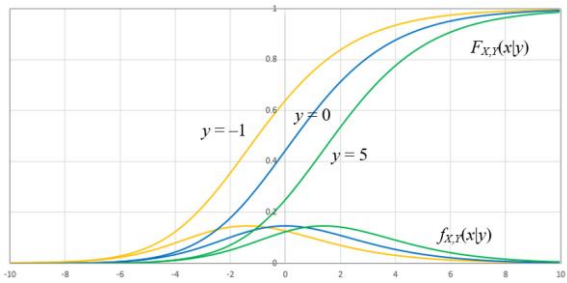
illetve

$$f_{X,Y}(y\,|\,x)=\frac{f_{X,Y}(x,y)}{f_X(x)}=\frac{2.2\cdot \mathrm{e}^{-1.1\cdot y}\cdot\left(1+2\cdot \mathrm{e}^{-0.5\cdot x}+\mathrm{e}^{-x}\right)}{\left(1+\mathrm{e}^{-0.5\cdot x}+\mathrm{e}^{-1.1\cdot y}\right)^3}. \tag{5.45}$$

Hasonlóképpen számíthatjuk ki a fejezet elején említett feltételes valószínűségeket is:

$$P(X\le x\,|\,Y\le y)=\frac{F_{X,Y}(x,y)}{F_Y(y)}=\frac{1+\mathrm{e}^{-1.1\cdot y}}{1+\mathrm{e}^{-0.5\cdot x}+\mathrm{e}^{-1.1\cdot y}}, \tag{5.46}$$

$$P(Y \le y \mid X \le x) = \frac{F_{X,Y}(x,y)}{F_X(x)} = \frac{1+\mathrm{e}^{-0.5\cdot x}}{1+\mathrm{e}^{-0.5\cdot x}+\mathrm{e}^{-1.1\cdot y}}. \qquad (5.47)$$



5.8. ábra. *A feltételes sűrűségfüggvények és eloszlásfüggvények y rögzített –1, 0 és 5 értékére*

A feltételes eloszlás esetében is meghatározhatjuk a leginkább valószínű várható értéket. A feltételes várható értékekre az alábbi kifejezéseket állapíthatjuk meg:

$$E(X \mid Y=y) = \int_{-\infty}^{+\infty} x \cdot f_{X,Y}(x \mid y)\,\mathrm{d}x = \frac{1}{f_Y(y)} \cdot \int_{-\infty}^{+\infty} x \cdot f_{X,Y}(x,y)\,\mathrm{d}x, \qquad (5.48)$$

$$E(Y \mid X=x) = \int_{-\infty}^{+\infty} y \cdot f_{X,Y}(y \mid x)\,\mathrm{d}y = \frac{1}{f_X(x)} \cdot \int_{-\infty}^{+\infty} y \cdot f_{X,Y}(x,y)\,\mathrm{d}y. \qquad (5.49)$$

E mennyiségek állandók, amennyiben a feltételben szereplő valószínűségi változó értékét rögzítjük, egyébként az első y, a második pedig x függvénye lesz, és ekképpen valószínűségi változóknak tekinthetők. Ez esetben bebizonyítható, hogy a feltételes várható értékek várható értéke

$$E\big(E(X \mid Y=y)\big) = m_X = \int_{-\infty}^{+\infty} x \cdot f_X(x)\,\mathrm{d}x, \qquad (5.50)$$

illetve

$$E\big(E(Y \mid X=x)\big) = m_Y = \int_{-\infty}^{+\infty} y \cdot f_Y(y)\,\mathrm{d}y. \qquad (5.51)$$

Példánkat folytatván, a feltételes várható értékeket y, illetve x függvényeként nem tudjuk meghatározni, mivel az integrálokat nem lehet analitikusan kiszámítani, azonban a feltételben szereplő változó rögzített értékeire numerikus integrálással megkapjuk a minket érdeklő értékeket. Így például:

$E(X \mid Y=-1)=-0.774671$, $E(X \mid Y=0)=0.613706$, $E(X \mid Y=1)=1.425329$,
$E(Y \mid X=-1)=0.023566$, $E(Y \mid X=0)=0.278957$, $E(Y \mid X=1)=0.478111$.

Kovariancia és korreláció

Ha egy kísérlet során két mennyiséget az X és az Y valószínűségi változókkal írunk le, akkor feltevődik annak a kérdésnek a megválaszolása, hogy azok függetlenek-e egymástól vagy sem. Másképpen: meg kell vizsgálnunk, hogy azok korrelálatlanok vagy korreláltak-e, s ha igen, akkor milyen mértékben.

A választ a két változó *kovarianciá*jának kiszámításával adhatjuk meg, ami a definíció szerint a két változó átlagos értéktől való eltéréseinek szorzatát veszi alapul, e szorzat várható értékét jelenti:

$$\operatorname{cov}(X,Y)=E\big((X-m_X)\cdot(Y-m_Y)\big), \qquad (5.52)$$

ahol $E(...)$ a várható érték kiszámításának operátora. E definíció alapján $\operatorname{cov}(X,X)=\sigma_X^2$. E mennyiség nagyobb értékei a két változó közötti lineáris függőségre utalnak, azonban nehéz összekötni a kovariancia kiszámolt értékét a lineáris kapcsolat valószínűségével.

Megállapítható, hogy

$$\operatorname{cov}(X,Y)=E(X\cdot Y)-E(X)\cdot E(Y)=m_{X\cdot Y}-m_X\cdot m_Y, \qquad (5.53)$$

ami a kovariancia eltolási tulajdonsága.

A valószínűségi függvény ismeretében diszkrét változók esetében a

$$\operatorname{cov}(X,Y)=\sum_{i=1,n}\sum_{j=1,m} f_{X,Y}(x_i,y_j)\cdot(x_i-m_X)\cdot(y_j-m_Y), \qquad (5.54)$$

folytonos változók esetében pedig a sűrűségfüggvény felhasználásával a

$$\operatorname{cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\cdot(x-m_X)\cdot(y-m_Y)\,\mathrm{d}x\,\mathrm{d}y \qquad (5.55)$$

képlettel adhatjuk azt meg.

Ha a kovariancia értéke nulla (legalábbis nagyon kicsi), akkor a két változó lineárisan független: ez azonban nem jelenti azt, hogy közöttük nem létezhet valamiféle nemlineáris kapcsolat.

A kovariancia pozitív értéke azt mutatja, hogy a két változó együtt változik (pl. ha az egyik növekedik, akkor a másik is követni fogja), a negatív érték pedig fordított irányú változást jelent (az egyik növekedése a másik csökkenését vonja maga után). A negatív érték nem jelent fordított arányosságot; a fordított arányosság egy nemlineáris kapcsolat.

Mivel a kovariancia nem normált mennyiség, a különböző esetekben kiszámított értékek összehasonlítása nem hordoz különösebb információt. Emiatt bevezetésre került egy új fogalom, a *korrelációs együttható*, amely a kovariancia normált értéke. Definíció szerint a normálást a szórások szorzatával való osztással érjük el:

$$\operatorname{corr}(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\cdot\sigma_Y}=\frac{E\left((X-m_X)\cdot(Y-m_Y)\right)}{\sigma_X\cdot\sigma_Y}. \qquad (5.56)$$

E mennyiség értéke a $[-1,+1]$ intervallumra esik ($\operatorname{corr}(X,X)=1$). Nulla, negatív vagy pozitív értéke ugyanolyan jelentéssel bír, mint a kovarianciáé.

Rangkorreláció

Az iménti korrelációs együttható csak a lineáris kapcsolat vizsgálására alkalmas. Ha például a két valószínűségi változó közötti kapcsolat $Y=X^2$, akkor az evidens függőség ellenére a korrelációs együttható ezt a függőséget nem fogja tükrözni. Emiatt a valószínűségi változók nemlineáris kapcsolatának a felderítésére másfajta mennyiség, a *rangkorrelációs együttható* lesz alkalmas. A rangkorreláció azt mutatja, hogy két mennyiség együtt változik-e vagy sem.

Két ilyen fajta korrelációs együttható használata gyakoribb, ezek a Spearman-féle ρ és a Kendall-féle τ. Mindkét esetben a rangkorrelációs együttható értéke a $[-1,+1]$ intervallumon van. Bár a két változó közötti matematikai kapcsolat felderítésére egyik sem alkalmas, legalább egy annyit tudunk, hogy ha az együttható ±1, akkor a két változó rangban teljes mértékben korrelált; a negatív érték azt jelenti, hogy egymással ellentétes irányban változnak. A 0 érték azt mutatja, hogy a két változó rangban nem korrelált, vagy pedig az adatpárok szimmetrikusak valamelyik tengelyre nézve.

A Spearman-féle rangkorreláció

A Spearman-féle rangkorrelációs együtthatót adó összefüggés a korrelációs együtthatóéval analóg, azonban nem a valószínűségi változókra, hanem azok rangjára kiszámolt kovariancia és szórások szerepelnek benne:

$$\rho=\frac{\operatorname{cov}(r_X,r_Y)}{\sigma_{r_X}\cdot\sigma_{r_Y}}. \qquad (5.57)$$

A kiszámításához az X és az Y változók (x_i,y_i) értékpárjait a rangjaikkal, az $(r_{X,i},r_{Y,i})$ párokkal helyettesítjük, majd ezeknek számítjuk ki a korrelációs

együtthatóját. Valamely x_i érték rangját a csökkenő sorrendbe állított értékek sorában elfoglalt helye jelenti, a legkisebb rangja (ami 1) a legnagyobb értéknek van, a legnagyobb rangja pedig a legkisebb értéknek van (ami egyenlő n-nel, a párok számával).

A Spearman-féle rangkorreláció Excelben

Egy táblázatban (5.9. ábra) rendezzük az összehasonlítandó valószínűségi változók mintavételezésével nyert x és y értékeket két egymás melletti oszlopba. Példaként vehetjük az

$$\begin{aligned} x_i &= -0.5 + rnd_1, \\ y_i &= k \cdot x_i + c \cdot (-0.5 + rnd_2) \end{aligned} \qquad (5.58)$$

mesterségesen előállított véletlen számokat, amelyek között a k együttható teremt kapcsolatot. Észrevehetjük, hogy ha $k = 0$, akkor elméletileg a két véletlen szám egymástól független kell legyen. Ha pedig $c = 0$ (és k nem), akkor a kettő lineáris kapcsolatban áll egymással.

Az r_X és az r_Y rangokat az Excel RANK.AVG() függvényével számíthatjuk ki. Ennek három paramétere van:

– *number*: az x vagy y változó, aminek a rangját keressük;

– *ref*: az intervallum (az oszlopban), amelyben a sorba állítandó elemek vannak;

– *order*: az opcionálisan megadott FALSE vagy 0 értéke csökkenő, ellenben növekvő sorrendben keres.

Az ábrán levő táblázat r_X oszlopában a

=RANK.AVG(B3,$B$3:$B$12,0), (5.59)

r_Y oszlopában a

=RANK.AVG(C3,$C$3:$C$12,0) (5.60)

függvényeket kell használni.

A korrelációs együttható kiszámítására szintén létezik beépített Excel-függvény, éspedig a CORREL(). Ennek két paramétere van, *Array1* és *Array2*, amelyek a két valószínűségi változó értékeinek intervallumát jelentik.

Esetünkben ezt felhasználhatjuk az X és Y valószínűségi változók $\mathrm{corr}(X, Y)$ korrelációs együtthatójának kiszámítására (*G9*-es cella):

=CORREL(B3:B12,C3:C12), (*5.61*)

valamint az r_X és az r_Y rangok ρ korrelációjának kiszámítására is (*G12*-es cella):

=CORREL(D3:D12,E3:E12). (*5.62*)

	A	B	C	D	E	F	G
1							
2	i	x	y	rx	ry		k
3	1	0.177202	-0.20796	5	6		0.5
4	2	-0.11664	-0.45743	10	9		
5	3	0.393498	-0.22073	2	7		c
6	4	0.006083	-0.06768	7	5		1
7	5	-0.03963	-0.23262	8	8		
8	6	-0.10464	-0.46089	9	10		corr(X,Y)
9	7	0.374391	0.453792	3	3		0.582411
10	8	0.433358	0.706877	1	1		
11	9	0.326997	-0.0043	4	4		ρ
12	10	0.011196	0.470854	6	2		0.709091

5.9. ábra. *Spearman-féle rangkorreláció Excelben*

Különböző eseteket tanulmányozva azt tapasztaljuk, hogy mindkét együttható az adatok bizonyos fokú kapcsolatára utal (tehát hogy azok nem teljesen függetlenek). Ha a minták n számát megnöveljük (további x_i és y_i értékeket tartalmazó sorokat adunk a táblázathoz), akkor $k=0$-ra azt tapasztalhatjuk, hogy n növekedésével az X és Y valószínűségi változók közötti korreláltság egyre inkább kisebbnek mutatkozik.

A Kendall-féle rangkorreláció

A Kendall-féle együttható kiszámításához az (x_i, y_i) értékpárokat x növekvő sorrendjébe rendezzük, azaz $x_1 \le x_2 \le x_3 \ldots \le x_n$. A párokat összehasonlítjuk: az elsőt az összes többivel, a másodikat a harmadiktól kezdve egész n-ig, tehát az i-edik párt minden $j > i$, $j \le n$-re. Az utolsó párt már nem kell összehasonlítani semmivel sem. Ha összesen n párunk van, akkor az összehasonlítások száma $(n-1)+(n-2)+\ldots+1$, ami egy számtani haladvány összege. Másképpen: ez a szám n elem másodosztályú kombinációinak a száma, azaz C_n^2.

A párokról a következőket állapítjuk meg:

– ha $x_i < x_j$ és $y_i < y_j$, akkor ez a két pár konkordáns; az ilyen esetek száma n_c,

– ha $x_i < x_j$ és $y_i > y_j$, akkor ez a két pár diszkordáns; az ilyen esetek száma n_d,

– ha $x_i = x_j$ és $y_i \neq y_j$, akkor e két pár x-ben kötött; az ilyen esetek száma t_x,

– ha $x_i \neq x_j$ és $y_i = y_j$, akkor e két pár y-ban kötött; az ilyen esetek száma t_y,

– ha $x_i = x_j$ és $y_i = y_j$, akkor e két pár x-ben is és y-ban is kötött; az ilyen eseteket nem vesszük figyelembe. Ezek száma t_{xy}.

Ha a tanulmányozott adatsorban nincsenek kötött párok, akkor a Kendall-féle rangkorrelációs együttható

$$\tau = \frac{n_c - n_d}{n \cdot (n-1)/2} \qquad (5.63)$$

(a nevezőben n a párok száma, míg maga a nevező az említett számtani haladvány összegeként az összehasonlítások számát adja).

Az előbbi esettel ellentétben, amikor x-ben vagy y-ban kötött párok is vannak, akkor a szakirodalomban többféle képletet is találunk. Az egyik lehetőség a következő:

$$\tau = \frac{n_c - n_d}{\sqrt{\frac{n' \cdot (n'-1)}{2} - \frac{t_x \cdot (t_x - 1)}{2}} \cdot \sqrt{\frac{n' \cdot (n'-1)}{2} - \frac{t_y \cdot (t_y - 1)}{2}}}, \qquad (5.64)$$

ahol az $n' = n - t_{xy}$-be nem számoljuk bele a kétszeresen kötött párok összehasonlítását, a kivont mennyiségek pedig az x-ben, illetve az y-ban kötött párokkal végzett összehasonlítások számát jelentik. A nevező maga a gyök alatti mennyiségek mértani közepe arányosa.

Egy egyszerűbb verzióban

$$\tau = \frac{n_c - n_d}{\sqrt{n_c + n_d + t_x} \cdot \sqrt{n_c + n_d + t_y}}. \qquad (5.65)$$

A Kendall-rangkorreláció kiszámítására Excelben nincs beépített függvény. A minták nagyobb n számára a számításokat VBA-programozás vagy külső könyvtárak nélkül nehéz lenne elvégezni, a gondot a párok összehasonlítása okozza.

Valószínűségi változók függvénye

Az eddigiekben a valószínűségi változók együttes eloszlását tanulmányoztuk, azaz a kísérlet eredményét több, esetleg egymástól nem független változóval írtuk

le, majd azok megjelenő értékeiből vonunk le következtetéseket (például azt, hogy a szerkezet állja-e vagy sem a terhelést).

Előfordulhat egy olyan eset is, amikor a kísérlet eredményét jelölő Y mennyiség két vagy több egyszerű változó függvényeként írható le: ekkor nem az együttes eloszlást tanulmányozzuk, hanem az

$$Y = Y(X_1, X_2, \dots X_i, \dots X_n) \tag{5.66}$$

függvény értékeinek eloszlását. Ilyen egyszerűbb függvények például:

$$Y = X_1 + X_2, \quad Y = X_1 \cdot X_2, \quad Y = X_1 / X_2. \tag{5.67}$$

Y értékei nyilván egy valószínűségi változó értékeiként mutatkoznak. Felvetődik tehát az a kérdés, hogy milyen kapcsolatot állapíthatunk meg az Y függvénynek és a változóinak az eloszlása között. A továbbiakban tegyük fel, hogy az X_i változók független változók.

Ha az átlag

$$m_Y = \int_{-\infty}^{+\infty} y \cdot f_Y(y)\,\mathrm{d}x \tag{5.68}$$

és a szórás

$$\sigma_Y = \sqrt{\int_{-\infty}^{+\infty} f_Y(y) \cdot (y - m_Y)^2\,\mathrm{d}y} \tag{5.69}$$

képletébe behelyettesítjük az Y változót függvényként megadó képletet és Y sűrűségfüggvényét, akkor e mennyiségeket analitikusan is kiszámíthatjuk. Néhány egyszerűbb esetben e számítások könnyebben elvégezhetők, és a következő táblázatban feltüntetett eredményekhez jutunk.

X	m_X	σ_X
c	c	0
$a \cdot X + c$	$a \cdot m_X + c$	$a \cdot \sigma_X$
$a \cdot X \pm b \cdot Y$	$a \cdot m_X \pm b \cdot m_Y$	$\sqrt{a^2 \cdot \sigma_X^2 + b^2 \cdot \sigma_Y^2}$
$a \cdot X \cdot Y$	$a \cdot m_X \cdot m_Y$	$a \cdot \sqrt{m_X^2 \cdot \sigma_X^2 + m_Y^2 \cdot \sigma_Y^2 + \sigma_X^2 \cdot \sigma_Y^2}$
$a \cdot X / Y$	$a \cdot m_X \cdot m_{1/Y}$	$a \cdot \sqrt{m_X^2 \cdot \sigma_X^2 + m_{1/Y}^2 \cdot \sigma_{1/Y}^2 + \sigma_X^2 \cdot \sigma_{1/Y}^2} = m_{X^2} \cdot m_{1/Y^2} - m_X^2 \cdot m_{1/Y}^2$

Az X/Y hányadosra vonatkozóan megjegyzendő, hogy azt X és $1/Y$ szorzataként értelmezzük, valamint azt is, hogy $m_{1/Y} \neq 1/m_Y$ és $\sigma_{1/Y} \neq 1/\sigma_Y$.

Más esetben a számítások már nem ennyire egyszerűek, ilyenkor a megoldást az Y függvény Taylor-sorba fejtése jelentheti, amelyet az

$$m = (m_{X_1}, m_{X_2}, \ldots, m_{X_i}, \ldots, m_{X_n}) \tag{5.70}$$

középpontban végzünk el:

$$Y \simeq Y(m_{X_1}, m_{X_2}, \ldots, m_{X_i}, \ldots, m_{X_n}) + \sum_{i=1}^{n} \left(\frac{\partial Y}{\partial X_i} \right)_m \cdot (X_i - m_{X_i}) + \ldots \cdot \tag{5.71}$$

E kifejtés első tagja az Y függvénynek az m pontban számított értékét jelenti, a következő tagokban pedig együtthatóként a változók szerinti parciális deriváltak értéke szerepel, szintén az m pontban.

Ha a Taylor-sorba fejtésnél csak a felírt tagokat vesszük tekintetbe, Y átlagos értéke a következőképpen közelíthető meg:

$$m_Y \simeq Y(m_{X_1}, m_{X_2}, \ldots, m_{X_i}, \ldots, m_{X_n}), \tag{5.72}$$

ugyanis a parciális derivált utáni zárójel átlaga zéró (X_i átlagából, ami m_{X_i}, kivonjuk m_{X_i}-t). Y szórásának megközelítése a táblázatban feltüntetett összegképlet alapján:

$$\sigma_Y \simeq \sqrt{\left(\frac{\partial Y}{\partial X_1} \right)_m^2 \cdot \sigma_{X_1}^2 + \left(\frac{\partial Y}{\partial X_2} \right)_m^2 \cdot \sigma_{X_2}^2 + \ldots + \left(\frac{\partial Y}{\partial X_n} \right)_m^2 \cdot \sigma_{X_n}^2}, \tag{5.73}$$

mivel a Taylor-sorba fejtés első tagja konstans, és annak szórása zéró.

Például ha

$$Y = \frac{a \cdot X_1^{\,2} + b \cdot X_1 + c}{X_2 \cdot \cos X_3}, \tag{5.74}$$

akkor az Y valószínűségi függvény várható értékét az

$$m_Y = \frac{a \cdot m_{X_1}^2 + b \cdot m_{X_1} + c}{m_{X_2} \cdot \cos m_{X_3}}, \tag{5.75}$$

a szórásnégyzetét pedig a

$$\sigma_Y^2 = \left(\frac{2 \cdot a \cdot m_{X_1} + b}{m_{X_2} \cdot \cos m_{X_3}} \right)^2 \cdot \sigma_{X_1}^2 + \left(-\frac{a \cdot m_{X_1}^2 + b \cdot m_{X_1} + c}{m_{X_2}^2 \cdot \cos m_{X_3}} \right)^2 \cdot \sigma_{X_2}^2 + \left(\frac{(a \cdot m_{X_1}^2 + b \cdot m_{X_1} + c) \cdot \sin m_{X_3}}{m_{X_2} \cdot \cos^2 m_{X_3}} \right)^2 \cdot \sigma_{X_3}^2 \tag{5.76}$$

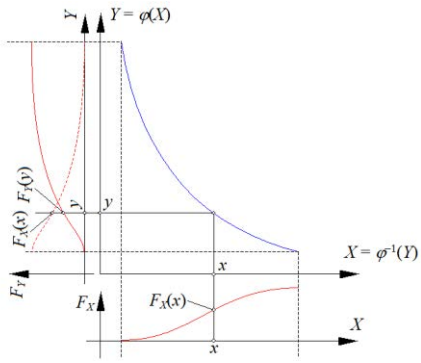
kifejezésekkel közelíthetjük.

Az átlag és a szórás meghatározása mellett felmerülhet a sűrűség, illetve az eloszlásfüggvény meghatározásának igénye is.

A legegyszerűbb esetben Y csak egyetlen X változó függvénye, $Y=\varphi(X)$. Ekkor, feltételezvén, hogy $\varphi(X)$ invertálható, Y eloszlásfüggvénye

$$F_Y(y)=P(Y\le y)=P(\varphi(X)\le y)= \\ =\begin{cases} P(X\le\varphi^{-1}(y))=F_X(\varphi^{-1}(y)) & \text{ha } \varphi^{-1}(Y) \text{ növekvő,} \\ P(X\ge\varphi^{-1}(y))=1-F_X(\varphi^{-1}(y)) & \text{ha } \varphi^{-1}(Y) \text{ csökkenő.} \end{cases} \tag{5.77}$$

Az invertálhatóság feltétele φ bijektivitása, ugyanakkor ez a függvény folytonos – következésképpen φ is, és az inverze is szigorúan monoton, innen következik a fenti képlet kifejtése. E tulajdonságok geometriai értelmezését, az eloszlásfüggvények kapcsolatát az 5.10. ábrán láthatjuk, a csökkenő inverz függvény esetét példázva.



5.10. ábra. *Az Y = φ(X) transzformáció eloszlásfüggvényeinek kapcsolata*

Mivel a sűrűségfüggvény az eloszlásfüggvény deriváltja, amennyiben a $\varphi^{-1}(Y)$ inverz függvény deriválható, és az növekvő, az összetett függvény deriválási szabályát alkalmazva

$$f_Y(y)=\frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}=\frac{\mathrm{d}F_X\left(\varphi^{-1}(y)\right)}{\mathrm{d}y}=\frac{\mathrm{d}F_X\left(\varphi^{-1}(y)\right)}{\mathrm{d}\left(\varphi^{-1}(y)\right)}\cdot\frac{\mathrm{d}\left(\varphi^{-1}(y)\right)}{\mathrm{d}y}= \\ =\frac{\mathrm{d}F_X(x)}{\mathrm{d}x}\cdot\frac{\mathrm{d}x}{\mathrm{d}y}=f_X(x)\cdot\frac{1}{\mathrm{d}y/\mathrm{d}x}=f_X(x)\cdot\frac{1}{\varphi'(x)}. \tag{5.78}$$

Amennyiben az $\varphi^{-1}(Y)$ inverz függvény csökkenő, akkor a levezetés megismétlésével kapott képlet csak egy „–” előjelben fog különbözni.

Ha $\varphi^{-1}(Y)$ növekvő, akkor a szigorúan monoton tulajdonsága miatt $\varphi(X)$ is növekvő, és ekkor $\varphi'(x) > 0$, így az előbbi relációval kapott $f_Y(y)$ sűrűségfüggvény sehol sem lehet negatív.

Ha $\varphi^{-1}(Y)$ csökkenő, akkor $\varphi(X)$ is csökkenő, és $\varphi'(x) < 0$, ami a „–" előjellel szintén egy sehol sem negatív $f_Y(y)$ sűrűségfüggvényhez vezet.

Ha a derivált abszolút értékét vesszük, a két esetet egyetlen összefüggésben foglalhatjuk össze:

$$f_Y(y) = f_X(x) \cdot \frac{1}{\left|\frac{\mathrm{d}y(x)}{\mathrm{d}x}\right|} . \qquad (5.79)$$

Ezeknek az összefüggéseknek a legfontosabb alkalmazásai a következők:

– amikor az x értékek egyszerű eltolásáról van szó:

$$Y = X + c \quad \rightarrow \quad \begin{cases} y = \varphi(x) = x + c, \quad x = \varphi^{-1}(y) = y - c, \\ \dfrac{\mathrm{d}y}{\mathrm{d}x} = 1, \\ F_Y(y) = F_X(y - c), \\ f_Y(y) = f_X(y - c); \end{cases} \qquad (5.80)$$

– amikor x léptékének megváltoztatásáról van szó:

$$Y = c \cdot X \quad \rightarrow \quad \begin{cases} y = \varphi(x) = c \cdot x, \quad x = \varphi^{-1}(y) = y / c, \\ \dfrac{\mathrm{d}y}{\mathrm{d}x} = c, \\ F_Y(y) = F_X(y / c), \\ f_Y(y) = \dfrac{1}{c} \cdot f_X(y / c). \end{cases} \qquad (5.81)$$

E két egyszerű műveletet használtuk a fontosabb eloszlások tárgyalásakor, például az $Y = (X - m_X) / \sigma_X$ normalizált változó előállításakor.

Számításainkat a többváltozós esetre is kiterjeszthetnénk, azonban a számításokat nehéz lenne elvégezni. Ezek miatt a mérnöki gyakorlatban gyakran egyszerűsítő feltételezésekbe bocsátkozunk a többváltozós eloszlást leíró függvényeket illetően. Például, ha $X_1, X_2, \ldots X_n$ normál eloszlást követő független valószínűségi változók, akkor azok bármely $Y = a_1 \cdot X_1 + a_2 \cdot X_2 + \ldots + a_n \cdot X_n$ lineáris kombinációja szintén normál eloszlásúnak tekinthető, az előbbi táblázat alapján kiszámítható átlaggal és szórásnégyzettel.

Hasonlóképpen feltételezhetjük, hogy ha $X_1, X_2, \ldots X_n$ lognormál eloszlású független változók, akkor a $Y = a_1 \cdot X_1^{\alpha_1} \cdot a_2 \cdot X_2^{\alpha_2} \cdot \ldots \cdot a_n \cdot X_n^{\alpha_n}$ szintén jó közelítéssel lognormál eloszlású lesz.

Ha meghatározzuk (például az említett közelítő képletekkel) Y átlagát és szórását, akkor fel tudjuk írni a feltételezett, közelítő sűrűség- és eloszlásfüggvényt is.

Példa: a Rayleigh-eloszlás

A Rayleigh-eloszlás egy folytonos

$$R = \sqrt{X^2 + Y^2} \tag{5.82}$$

valószínűségi változó eloszlása, ahol X és Y normál eloszlású, nulla átlagú, azonos szórású és egymástól független változók.

Ezt az eloszlást véletlenszerűen változó vektoriális mennyiségek nagyságának leírására használjuk, amikor e mennyiség összetevői (a tengelyekre eső vetületei) normál eloszlást követnek. Tipikus alkalmazása a szél sebességének vizsgálata, amikor e sebességet két egymásra merőleges irány szerint mérjük.

Az említett feltételek mellett a két összetevő sűrűségfüggvénye

$$f_X(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \mathrm{e}^{x^2/(2 \cdot \sigma^2)}, \quad f_Y(y) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \mathrm{e}^{y^2/(2 \cdot \sigma^2)}, \tag{5.83}$$

együttes eloszlásuk pedig

$$F_{X,Y}(x, y) = \int_D f_X(x) \cdot f_Y(y)\, \mathrm{d}x\, \mathrm{d}y = \frac{1}{2 \cdot \pi \cdot \sigma^2} \cdot \int_D \mathrm{e}^{(x^2+y^2)/(2 \cdot \sigma^2)}\, \mathrm{d}x\, \mathrm{d}y. \tag{5.84}$$

A kitűzött cél a vektor hosszának (az R valószínűségi változónak) a tanulmányozása, így a D tartományt a

$$\rho = \sqrt{x^2 + y^2} \le r \tag{5.85}$$

feltétellel írhatjuk le. Ekképpen a számításokat egyszerűbb polárkoordinátákban elvégezni:

$$F_R(r) = \frac{1}{2 \cdot \pi \cdot \sigma^2} \cdot \int_0^{2 \cdot \pi} \int_0^r \rho \cdot \mathrm{e}^{\rho^2/(2 \cdot \sigma^2)}\, \mathrm{d}\rho\, \mathrm{d}\varphi = \frac{1}{\sigma^2} \cdot \int_0^r \rho \cdot \mathrm{e}^{\rho^2/(2 \cdot \sigma^2)}\, \mathrm{d}\rho, \tag{5.86}$$

ahol $\mathrm{d}A = (\rho \cdot \mathrm{d}\varphi) \cdot \mathrm{d}\rho$, az eredmény pedig a Rayleigh-eloszlás eloszlásfüggvénye. A megfelelő sűrűségfüggvény pedig e kifejezés deriválásával

$$f_R(r) = \frac{r}{\sigma^2} \cdot \mathrm{e}^{r^2/(2 \cdot \sigma^2)}. \tag{5.87}$$

A központi határeloszlás tétele

Van a fentieknél általánosabb érvényű és az eloszlás függvényeinek említett mérnöki megközelítését alátámasztó tétel is. Ez a *központi határeloszlás tétele*, mely szerint ha $X_1, X_2, \dots X_n$ egymástól független, tetszőleges, de azonos eloszlást követő valószínűségi változók, amelyek közül egyiknek sincs meghatározó hatása, akkor az $Y = X_1 + X_2 + \dots + X_n$ összeg aszimptotikusan normál eloszlású lesz.

Az aszimptotikus azt jelenti, hogy n-nek eléggé nagynak kell lennie. Azt, hogy egyik változónak sincs meghatározó hatása, úgy kell értelmezni, hogy átlagaik közel azonos nagyságúak. Ez utóbbi feltétel ritkábban teljesül, ilyenkor a központi határeloszlás tételét az $Y = \sum` X_i$ változó standardizált alakjára mondják ki, és eszerint a standardizált változók összegének eloszlása lesz aszimptotikusan normál eloszlású. Az összeg tagjainak standardizált alakját már ismerjük:

$$X_i^s = \frac{X_i - m_{X_i}}{\sigma_{X_i}}, \qquad (5.88)$$

ezek átlaga bármely i-re nulla, szórásuk pedig egységnyi, s ekképpen egyiknek sem lehet meghatározó hatása az összegük eloszlására nézve. Az összeg átlaga az átlagok $\sum_{i=1,n}` m_{X_i}$ összege (ld. a fenti táblázatot), emiatt a standardizált változók $Y^s = X_1^s + X_2^s + \dots + X_n^s$ összegének átlaga is nulla lesz ($m_{Y^s} = 0$). Az összeg szórása a táblázat alapján $\sqrt{\sum_{i=1,n} \left(\sigma_{X_i}^s\right)^2}$, és mivel a szórások mind egységnyiek, $\sigma_{Y^s} = \sqrt{n}$. Ezek szerint a normalizált változók összegének az eloszlása egy nulla átlagú, de nem egységnyi szórású normál eloszlás felé közelít.

Van a központi határeloszlásnak egy még általánosabb formája is (Ljapunov), amely szerint a tetszőleges (tehát változónként nem feltétlenül azonos típusú) eloszlást követő valószínűségi változóknak akkor normál eloszlású az összege, ha minden X_i változó átlagos értéke, szórása, abszolút eltérése és $\varepsilon_{X_i} = | x_i - m_{X_i} |^3$ harmadrendű momentuma véges nagyságú, és ha

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \varepsilon_{X_i}^3}{\sqrt{\sum_{i=1}^{n} \sigma_{X_i}^2}} = 0 \cdot \qquad (5.89)$$

Hasonló összefüggéseket állapíthatunk meg a valószínűségi változók szorzatára is: amennyiben azok lognormális eloszlást követnek, a felsorolt feltételek mellett a szorzat is (standard) lognormális eloszlást fog mutatni.

E tételek gyakorlati alkalmazását megnehezítik bizonyos dolgok: egyfelől csak összegre, illetve szorzatra alkalmazhatjuk, másrészt feltevődik a kérdés, hogy vajon mekkora kell, hogy legyen n, teljesülnek-e a tétel alkalmazásának a feltételei, stb.

6. A KÍSÉRLETI EREDMÉNYEK STATISZTIKAI FELDOLGOZÁSA

Az adatokat rendszerint *vizsgálat*tal, azaz megszemléléssel vagy méréssel nyerjük.

A *megszemlélés* a helyzet felméréséből, az egyedek megszámlálásából áll.

A tényleges *mérés* során a megmérendő mennyiséget összehasonlítjuk a *mértékegység*gel, a kettő hányadosa a *mérőszám,* avagy a *mérték*. Az összehasonlításhoz valamilyen *mérőeszköz*t használunk. A mérés eredményét, annak pontosságát több tényező is befolyásolja, így a mérőeszköz vagy az összehasonlításnál használt etalon pontossága, a leolvasási hibák, valamint a mérés folyamata során fellépő hibák. Ezek miatt a meghatározott mérték még az ismételten megmért mennyiségek esetében is változó mennyiségként fog jelentkezni: azt mondjuk, hogy a megmért mennyiséget *mérési hibák* terhelik.

A mérés során fellépő hibákat három kategóriába szokás felosztani:

–*a*). *szisztematikus hibák* (vagy rendszeres hibák): a mérőeszközhöz vagy a mérési eljáráshoz köthető determinisztikus jellegű eltérések, amelyek nagysága állandó és meghatározható. Ezeknek a hibáknak a forrása lehet például a mérőeszköz elállítódása(a nullpont eltolódása egy tolómérce esetében) vagy pedig a hőmérséklet-változás okozta térfogatváltozás (amikor egy piknométerrel a sűrűséget nem azon a hőmérsékleten mérjük, amelyre az eszközt kalibrálták);

– *b*). *durva hibák*: ezek forrása például a tévedés (pl. téves leolvasás), valamilyen erős környezeti hatás, a mérés folyamata során bekövetkező baleset (pl. rossz érintkezés elektromos mennyiségek mérésénél), az adatátviteli vagy rögzítési hiba és más ehhez hasonló körülmények lehetnek;

– *c*). *véletlen hibák*: eredetük a mérés körülményeit befolyásoló jelenségek komplexitása miatt ismeretlen, valószínűségi változóként foghatók fel.

A szisztematikus hibákat az előidéző okok ismeretében korrigálhatjuk, ilyen módon a hatásuk kioltható.

A durva hibák egy részét akár szemrevételezéssel is felfedezhetjük, ilyen esetben az átlagtól való jelentős eltérést tekintjük durva hibának. Itt persze egy jó kérdés az, hogy mit tekintünk jelentős eltérésnek, vagyis hogy hol a határ a durva hibák és a véletlen hibák között. Erre a választ a véletlen hibák statisztikai eloszlásáról felállított hipotézis alapján adhatunk: ha a méréssel szerzett adatok halmaza nem illeszkedik a hipotetikus eloszláshoz, akkor vagy a durva hibák által befolyásolt adatok miatt történik mindez (és akkor meg kell keresni, hogy melyek

azok: nem feltétlenül a megmért legnagyobb vagy legkisebb értékekről van szó), vagy pedig az eloszlásról felállított hipotézisünk nem állja meg a helyét.

Ha csak kevés megmért adat áll a rendelkezésre, akkor a hiba eloszlásáról nem lehet fogalmat alkotni, így a durva hibák elkülönítése bizonytalanságokba ütközik. Ilyenkor megoldást jelenthet a túl nagynak és a túl kicsinek tűnő értékek törlése, vagy pedig a megmért legnagyobb és legkisebb értékek figyelmen kívül hagyása. Ha például egy tengely átmérőjét ötször mérjük meg, akkor a legkisebb és a legnagyobb értékek eltávolítása után a megmaradt három adatból számíthatunk egy átlagot, amelyet a tengely átmérőjének tekintünk.

Statisztikai sokaság

Egy kísérlet során valamilyen jelenséget vizsgálunk, rögzített körülmények között. Mind a körülményeket, mind a kísérleti eredményeket mennyiségileg valamilyen mérhető paraméterrel írjuk le. Ha a kísérlet valamilyen véletlen folyamatot tanulmányoz, akkor még szigorúan beállított és ellenőrzött körülmények között is a mért eredmények bizonyos változatosságot, véletlen értékeket mutatnak. A kísérleti eredmények ilyen halmazát *statisztikai sokaság*nak vagy *populáció*nak nevezzük, amely a tanulmányozott mennyiséget leíró valószínűségi változó különböző értékeit tartalmazza. A tanulmányozott jelenség leírásához e valószínűségi változó pontos meghatározására (tehát eloszlásának meghatározására) van szükségünk.

A nagy számok törvényének alapján e pontos meghatározáshoz nagyszámú kísérlet elvégzésére lenne szükség, azonban a gyakorlatban erre legtöbbször nincs lehetőség, csak korlátozott számú egyedből álló statisztikai sokaság tanulmányozására szorítkozhatunk. Ezen sokaság (mért eredmények) alapján rajzoljuk meg a hisztogramokkal közelített empirikus eloszlásfüggvényt és sűrűségfüggvényt.

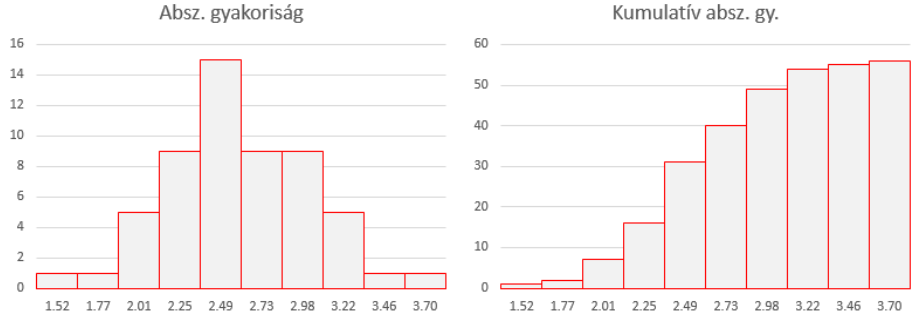
Minta, visszatevéses és visszatevés nélküli mintavétel

Sok esetben a statisztikai sokaság elemeinek oly nagy, hogy minden egyes elemének tanulmányozására nincsen lehetőség. Ezt a mondatot könnyebben értelmezhetjük, ha például valamilyen automata gép által gyártott alkatrész méreteinek ellenőrzésére gondolunk: a gép igen sok alkatrészt gyárthat rövid idő alatt, azonban nekünk nincs lehetőségünk minden egyes alkatrész megvizsgálására. Ilyenkor a statisztikai sokaságot jelentő halmazból egy kényelmesebben tanulmányozható részhalmazt, úgynevezett *mintá*t különítünk

el, és vizsgálatainkban e minta elemeire szorítkozunk. A *mintavétel* véletlenszerűen történik, azaz a statisztikai sokaság bármely eleme azonos valószínűséggel kerülhet a kiválasztott minta halmazainak elemei közé. A mintavétel lehet *visszatevéses*, amikor az adott egyed megvizsgálása után visszakerül abba a sokaságba, ahonnan a következő mintát vesszük (tehát megvan annak a valószínűsége, hogy az másodszorra is a kezünkbe kerül); és lehet *visszatevés nélküli*, amikor egy-egy egyed csak egyszer kerül a megvizsgált mintába. Amennyiben a statisztikai sokaság egyedeinek száma nagy, akkor a visszatevés nélküli és a visszatevéses mintavétel közel azonos eredményekhez vezet (mivel valamely egyed kiválasztásának valószínűsége igen kicsi). Ellenben a két eset tárgyalását szét kell választanunk, hiszen valamely érték megjelenésének empirikus valószínűsége valamely minta többszöri kiválasztásának megnövekedett valószínűsége miatt nagyobb lesz a tényleges valószínűségnél.

Adatok feldolgozása

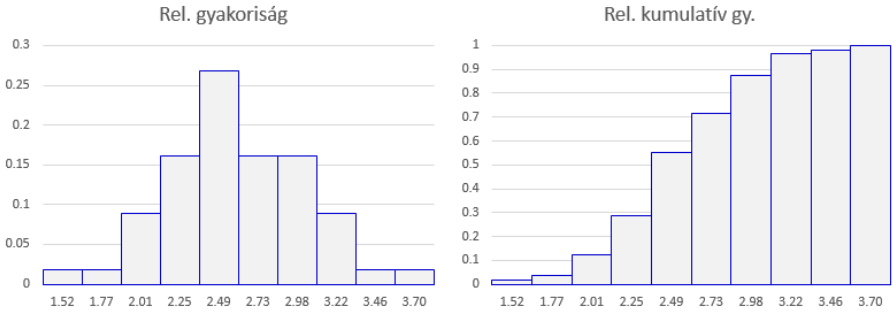
Mintánkat a könnyebb feldolgozás céljából rendezzük, vagyis az észlelt értékeket növekvő sorrendbe állítjuk. Az empirikus sűrűségfüggvény és eloszlásfüggvény megrajzolásához rendezett mintánkat kétféleképpen is feldolgozhatjuk. Így a már ismertetett eljárás szerint a legkisebb és legnagyobb észlelt érték által lehatárolt intervallumot részintervallumokra (osztályokra) osztjuk és megszámoljuk, hogy egy-egy osztályba hány elem esik. A részintervallumok hosszát általában egyenlőnek vesszük, mindenik részintervallumot a középértékével jellemezhetünk.



6.1. ábra. *Az abszolút gyakoriságok hisztogramjai*

A 6.1. ábrán egy $M = 56$ adatból álló minta hisztogramjait látjuk: a kísérleti értékek a legkisebb és legnagyobb megmért adatok által lehatárolt $[1.40, 3.82]$

intervallumon helyezkednek el, amelyet $Z=10$ egyenlő, $\Delta x=0.242$ hosszúságú osztályra bontottunk. A 6.2. ábra a normalizált hisztogramokat mutatja – ezek szemlátomást a normál eloszlás sűrűség- és eloszlásfüggvényeire hasonlítanak.



6.2. ábra. *A relatív gyakoriságok hisztogramjai*

Az 56 adatból álló rendezett minta a következő:

i	1	2	3	4	5	6	7	8	9	10
x_i	1.403	1.679	1.896	2.066	2.075	2.102	2.122	2.162	2.173	2.217
i	11	12	13	14	15	16	17	18	19	20
x_i	2.230	2.251	2.304	2.315	2.323	2.354	2.380	2.382	2.423	2.432
i	21	22	23	24	25	26	27	28	29	30
x_i	2.448	2.477	2.487	2.498	2.508	2.533	2.533	2.534	2.560	2.585
i	31	32	33	34	35	36	37	38	39	40
x_i	2.612	2.622	2.638	2.718	2.718	2.742	2.749	2.752	2.780	2.843
i	41	42	43	44	45	46	47	48	49	50
x_i	2.888	2.915	2.927	2.947	2.974	2.990	3.022	3.050	3.060	3.106
i	51	52	53	54	55	56				
x_i	3.115	3.160	3.187	3.271	3.576	3.821				

Ezeket a „kísérleti” adatokat az Excel véletlenszám-generátorával előállított adatok alapján a Box–Müller-eljárással állapítottuk meg; a valószínűségi változó elméleti átlaga 2.5, a szórása pedig 0.7.

A második módszer alkalmazása során a megjelent értékeket egyenként tekintjük, tehát egy osztályba csak azonos értékek kerülhetnek. Így a részintervallumok hossza változó lesz.

Az empirikus átlag és az empirikus szórás

A mintánk alapján az ismertetett Σ-t tartalmazó képletekkel (tehát nem az integrálokat tartalmazó képletekkel) megállapíthatjuk a tanulmányozott mennyiség empirikus átlagos értékét (várható értékét), az empirikus szórását és egyéb empirikus jellemzőit. Amennyiben az egyedek (mért értékek) M száma

nagy, joggal elvárhatjuk, hogy ezen empirikus jellemzők a tanulmányozott valószínűségi változó jellemzőinek jó közelítését (becslését) adják.

Példánkban a mintát alkotó egyedek száma $M = 56$, azok összege $\sum_{k=1}^{M} x_k = 145.64$, így az empirikus átlag

$$m_X = \bar{x} = \frac{\sum_{k=1}^{M} x_k}{M} = \frac{145.64}{56} = 2.601. \qquad (6.1)$$

Ezt az átlagot az Excel AVERAGE() függvényével is kiszámíthatjuk, amelynek a paramétere a mintát tartalmazó tömb.

Amennyiben az osztályokkal dolgozunk, a $Z = 10$ osztályon számított $\sum_{i=1}^{Z} n_i \cdot x_i = 145.57$ összeg valamivel kisebb az előbbinél, így az empirikus átlagra is valamivel kisebb értéket kapunk:

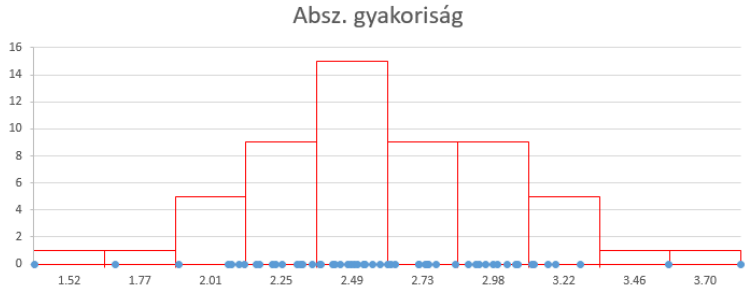
$$m_X = \frac{\sum_{i=1}^{Z} n_i \cdot x_i}{\sum_{i=1}^{Z} n_i} = \frac{145.57}{56} = 2.599. \qquad (6.2)$$

A különbségek onnan származnak, hogy a második képlet esetében az $n_i \cdot x_i$ szorzatok függetlenek attól, hogy az i-edik osztály Δx hosszúságú intervallumán az adatok tulajdonképpen hol is helyezkednek el. Ha egy osztályban az adatok egyenletesen oszlanak el, akkor annak az x_i közepe jobban megközelíti az illető osztályban szereplő adatok átlagát. Ha viszont például egy osztályban az adatok inkább a felső határérték körül tömörülnek, akkor az $n_i \cdot x_i$ szorzat kisebb lesz az elemek összegénél, és így a kiszámított empirikus átlag is kisebb lesz az első képlettel meghatározotthoz viszonyítva. A 6.3. ábrán észrevehetjük, hogy például a harmadik, 2.01 középértékű osztályban az elemek túlnyomó része inkább a felső határérték közelében helyezkedik el. Emiatt pontosabb eredményhez vezet az első képlet használata, ami egyébként nem igényli a hisztogramok megrajzolását sem.

Az empirikus szórásnégyzet meghatározásához először kiszámítjuk az egyedek átlagtól való eltérésének négyzetét, majd azok összegzésével és átlagolásával a

$$\sigma_X^2 = \frac{\sum_{k=1}^{M} (x_k - m_X)^2}{M} = \frac{10.69}{56} = 0.191 \qquad (6.3)$$

értékhez jutunk. Az empirikus szórást (a fenti mennyiség négyzetgyökét) a STDEV.P() függvénnyel is kiszámíthatjuk, amelynek a paramétere a mintát tartalmazó tömb.



6.3. ábra. *Az egyedek elhelyezkedése a hisztogram osztályain belül*

Amennyiben ugyanannak a valószínűségi változó eloszlásának valamely a paraméterét (átlagos értékét, szórásnégyzetét stb.) más és más minták alapján számítjuk ki, azt fogjuk tapasztalni, hogy e paraméter értéke mintánként más és más lehet. Ezek szerint az a paraméter is egy változó mennyiség lesz, melynek értéke véletlenszerűen fog változni a mintavételezés többszöri megismétlése során. Az ilyen paramétereket *statisztikai függvény*eknek nevezik.

Ezek szerint a statisztikai függvénynek is van valószínűség-eloszlása (sűrűségfüggvénye és eloszlásfüggvénye), átlagos értéke, szórása és így tovább. Az a paraméter becslését *torzítatlan*nak mondják, ha átlagos értéke megegyezik az a paraméter elméleti értékével; ellenben a becslés *torzított*. Tehát ha a az X változó átlagos értékének a különböző mintákból kiszámított értékét jeleni, akkor a abban az esetben lesz az m_X várható érték torzítatlan becslése, ha

$$m_a = m_X \,. \tag{6.4}$$

Könnyen belátható (a Σ-ás képletek felhasználásával), hogy az empirikus átlag torzítatlan becslése a tényleges átlagnak, ugyanis

$$m_{m_x} = \frac{\overbrace{\sum_{i=1}^{n}\frac{x_i}{n}+\sum_{i=1}^{n}\frac{x_i}{n}+\ldots+\sum_{i=1}^{n}\frac{x_i}{n}}^{k\text{-szor}}}{k} = \sum_{i=1}^{M}\frac{x_i}{M}, \quad \text{ahol } n=\frac{M}{k} \tag{6.5}$$

(az empirikus átlagokat azonos számú egyedet tartalmazó mintából számítjuk, és akkor az átlagok átlaga egyenlő az M egyed átlagával). Azonban a szórásnégyzet becslésénél más a helyzet, ugyanis a mintákon számított empirikus

szórásnégyzetek átlaga nem egyezik meg az M egyed figyelembevételével számított szórásnégyzettel, ha a számításokat elvégezzük, akkor az

$$m_{\sigma_X^2} = \frac{n-1}{n} \cdot \sigma_X^2 \qquad (6.6)$$

összefüggéshez jutunk. E szerint a populáció szórásnégyzete nagyobb a minta szórásnégyzeténél. Ha n (a minta elemeinek száma) eléggé nagy, akkor az $(n-1)/n$ hányados közel áll az n-hez, és akkor az empirikus szórásnégyzet az igazi torzítatlan becslésének tekinthető. Ellenben, ha a minta elemeinek száma kicsi, akkor empirikus szórásnégyzetünket korrigálnunk kell. Ha a szórásnégyzet becslésében a megengedhető hibát 1%-nak vesszük, akkor n legalább 100 kell, hogy legyen ahhoz, hogy ne kelljen korrigálnunk az empirikus szórásnégyzetet, ellenben a fenti formulát kell alkalmaznunk: a korrigált szórásnégyzetet adó képlet

$$\sigma_X^2 * = \frac{n}{n-1} \cdot \sigma_X^2 \qquad (6.7)$$

lesz.

Excelben a korrigált szórást (a fenti mennyiség négyzetgyökét) a STDEV.S() függvénnyel számíthatjuk ki. A függvény nevében az „S” a mintára („sample”) utal. A STDEV.P() függvénnyel a korrigálatlan szórást határozhatjuk meg; a nevében a „P” a populációra utal (az argumentumként beadott számsort populációnak tekinti).

Példánkban a minta elemeinek száma 56 volt, így a kiszámított empirikus szórásnégyzet korrigálásra szorul:

$$\sigma_X^2 * = \frac{n}{n-1} \cdot \sigma_X^2 = \frac{56}{55} \cdot 0.191 = 0.194 . \qquad (6.8)$$

Ha több (N) méréssorozatot végzünk, vagy több mintát veszünk, akkor az eljárás tehát a következő lehet: minden méréssorozat vagy minta alapján kiszámítunk egy-egy átlagos $m_{X,i}$ értéket és empirikus $\sigma_{X,i}^2$ szórásnégyzetet, majd a mért X mennyiség várható (átlagos) értékét az

$$m_X = \frac{1}{N} \cdot \sum_{i=1}^{N} m_{X,i} \qquad (6.9)$$

átlag, szórásnégyzetét pedig az

$$\sigma_X^2 = \frac{N}{N-1} \cdot \left(\frac{1}{N} \cdot \sum_{i=1}^{N} \sigma_{X,i}^2 \right) = \frac{1}{N-1} \cdot \sum_{i=1}^{N} \sigma_{X,i}^2 \qquad (6.10)$$

korrigált átlag formájában határozzuk meg.

Konfidencia-intervallumok

Mivel a minta alapján kiszámított a empirikus mennyiségek (a kiszámított m_X átlag és a σ_X^2 szórásnégyzet) szintén valószínűségi változók, felvetődik az a kérdés, hogy mennyire pontosan határoztuk meg az értéküket. E változók bizonyos valószínűségi eloszlással, tehát sűrűség- és eloszlásfüggvénnyel (f_a, illetve F_a) rendelkeznek. E függvények ismeretében az a mennyiség lehetséges értékeinek intervallumán elhatárolhatunk egy olyan tartományt, amelyen kívül a csak valamely p valószínűséggel lesz megtalálható (tehát $1-p$ valószínűséggel van azon belül). Ezt a tartományt a átlagos értéke körül szimmetrikusan veszik fel:

$$P(m_a - \varepsilon \le a \le m_a + \varepsilon) = 1 - p\,, \qquad (6.11)$$

az $[m_a - \varepsilon, m_a + \varepsilon]$ intervallumot pedig az $1-p$ *megbízhatósági szint*nek megfelelő *konfidencia-intervallum*nak nevezzük. Az a mennyiség, amelynek konfidencia-intervallumát meg szoktuk határozni, rendszerint a tanulmányozott valószínűségi változó átlaga vagy várható értéke szokott lenni. p valószínűség értéke a gyakorlatban 0.001, 0.01, 0.02, 0.05 vagy 0.10 szokott lenni, a megbízhatósági szintet pedig százalékban szokták megadni: az a felsorolt öt valószínűségnek megfelelően rendre 99.9, 99, 98, 95, illetve 90%.

Az intervallum határait megadó ε mennyiség kiszámítása X (a tanulmányozott mennyiség, statisztikai sokaság) és a (a statisztikai jellemző) eloszlásának ismeretében történhet.

Konfidencia-intervallum normál eloszlású sokaság esetén, amikor ismert a szórás

Tegyük fel, hogy a tanulmányozott X valószínűségi változó normál eloszlást követ. Átlagát (várható értékét) egy n elemű mintából kiszámított empirikus

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad (6.12)$$

átlagos értékkel, szórását pedig az előbbiekben elmondottak alapján megállapított

$$\sigma^* = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2} \qquad (6.13)$$

korrigált empirikus szórással közelíthetjük meg.

A mintát alkotó $x_1, x_2, \dots x_n$ mennyiségek mind ugyanazt az m_X átlagú és σ_X szórású normál eloszlást követik, így az $y = x_1 + x_2 + \dots + x_n$ kompozíciójuk (algebrai összegük) átlaga $n \cdot m_X$, szórása pedig $\sqrt{n} \cdot \sigma_X$ lesz. E kompozíció normál eloszlásúnak tekinthető, a sűrűségfüggvénye:

$$f(y) = \frac{1}{\sqrt{n} \cdot \sigma_X \cdot \sqrt{2 \cdot \pi}} \cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{y - n \cdot m_X}{\sqrt{n} \cdot \sigma_X}\right)^2}. \qquad (6.14)$$

Ha az említett kompozíciót elosztjuk a mintát alkotó elemek n számával, akkor az X valószínűségi változó empirikus átlagát kapjuk. Az $\bar{x} = (x_1 + x_2 + \dots + x_n)/n$ valószínűségi változó átlaga és szórása $x_1 + x_2 + \dots + x_n$ átlagának és szórásának n-ed része (m_X, illetve $\sigma_X / \sqrt{n}$), tehát sűrűségfüggvénye

$$g(\bar{x}) = \frac{\sqrt{n}}{\sigma_X \cdot \sqrt{2 \cdot \pi}} \cdot \mathrm{e}^{-\frac{n}{2}\left(\frac{\bar{x} - m_X}{\sigma_X}\right)^2}. \qquad (6.15)$$

Ezek szerint az X valószínűségi változó várható értéke is normál eloszlású lesz, az előbbi $g(x)$ sűrűségfüggvénnyel és annak integrálásával (de a normál eloszlás ismert eloszlásfüggvényéből is) előállítható $G(x)$ eloszlásfüggvénnyel. Ezt az eloszlást standard formára hozhatjuk, a

$$z = \frac{\bar{x} - m_X}{\sigma_X / \sqrt{n}} \qquad (6.16)$$

változócsere bevezetésével, mely standard formának a továbbiakban Φ-vel jelölt eloszlásfüggvényét az előbbi fejezetekből már ismerjük. E függvény tulajdonságait felhasználván, annak valószínűségét, hogy a z mennyiség egy bizonyos, a zérus átlagnál kisebb $-\lambda$ határértéknél kisebb legyen,

$$P(z \le -\lambda) = \Phi(-\lambda) \qquad (6.17)$$

formában írjuk fel. Annak valószínűsége, hogy z a zérus átlaghoz viszonyítva szimmetrikusan felvett λ határértéknél nagyobb legyen,

$$P(z \ge \lambda) = 1 - \Phi(\lambda). \qquad (6.18)$$

A standard normál eloszlás sűrűségfüggvénye szimmetrikus, tehát az előbbi két valószínűség egyenlő egymással. Ennek következtében annak valószínűsége, hogy z valamely $[-\lambda, +\lambda]$ intervallumon belül legyen:

$$P(-\lambda \le z \le \lambda) = 1-[P(z \le -\lambda) + P(z \ge \lambda)] = \\ = 1-2\cdot\big(1-\Phi(\lambda)\big) = 2\cdot\Phi(\lambda)-1. \qquad (6.19)$$

Ezek szerint

$$P\left(|\,\overline{x} - m_X| \le \lambda \cdot \frac{\sigma_X}{\sqrt{n}}\right) = 2\cdot\Phi(\lambda)-1, \qquad (6.20)$$

avagy

$$P\left(\overline{x} - \lambda\cdot\frac{\sigma_X}{\sqrt{n}} \le m_X \le \overline{x} + \lambda\cdot\frac{\sigma_X}{\sqrt{n}}\right) = 2\cdot\Phi(\lambda)-1\cdot \qquad (6.21)$$

Ez utóbbi kifejezésben az egyenlőtlenség két oldalán a konfidencia-intervallum határait ismerjük fel, a λ paramétert pedig úgy kell meghatároznunk, hogy e valószínűség a megbízhatósági szinttel legyen egyenlő. Ehhez a standard normáleloszlás Φ eloszlásfüggvényéből kiindulva meghatározzuk a $\Phi(\lambda) = 1 - p/2$ egyenlőségnek eleget tevő λ-t. Ezt hagyományosan a kézikönyvekben, segédletekben, de interneten is közzétett függvényértékek táblázatából keresték ki, és az adatokat rendszerint interpolálni kellett, mivel a függvényértékek csak bizonyos lépéssel szerepeltek. A keresgélés és interpolálás helyett használhatjuk az interneten fellelhető számos „kalkulátor" valamelyikét, amelyek rendszerint tetszetős ábrákon szemléltetik az eloszlásokkal kapcsolatos fogalmakat (akár magyar nyelven is). A meghatározott λ értékkel $\varepsilon = \lambda\cdot\frac{\sigma_X}{\sqrt{n}}$, a konfidencia-intervallum határai pedig $\overline{x} \pm \varepsilon$.

Példánkat folytatván tegyük fel, hogy a meghatározott korrigált $\sigma_X^2{}^* = 0.194$ empirikus szórásnégyzet elfogadható a tényleges szórásnégyzet kellő pontosságú közelítéseként, a konfidencia-intervallumot pedig 95%-os megbízhatósággal szeretnénk meghatározni ($p = 0.05$). Ehhez a megbízhatósági szinthez azt a λ értéket kell megkeressük, amelyre $\Phi(\lambda) = 0.975$. Ha ehhez Excelt szeretnénk használni, akkor a standard normál eloszlás sűrűségfüggvényének argumentumát a megadott valószínűségnek megfelelően visszaadó NORM.S.INV() függvényt használhatnánk. Ennek csak egy paramétere van, a *Probability*nek nevezett valószínűség. A megadott megbízhatósági szintnek megfelelő valószínűséggel e függvény a $\lambda = 1.960$ értéket téríti vissza, amellyel

$$\varepsilon = \lambda\cdot\frac{\sigma_X}{\sqrt{n}} = 1.960\cdot\frac{\sqrt{0.194}}{\sqrt{56}} = 0.115\cdot \qquad (6.22)$$

Ezzel az átlag 95%-os megbízhatósággal a $[2.601-0.115, 2.601+0.115]$, vagyis a $[2.486, 2.716]$ intervallumon keresendő.

Az Excelben azonban van egy CONFIDENCE.NORM() függvény is, amellyel egy pár műveletet megspórolhatunk. Ez a függvény közvetlenül ε értékét adja vissza, a három paraméternek megfelelően:

– *Alpha*: a p valószínűség (ami tehát 1 – a megbízhatósági szint);
– *Standard_dev*: a minta σ szórása;
– *Size*: a minta elemeinek n száma.

Segítségével 99%-os megbízhatósággal (*Alpha* 0.01) $\varepsilon = 0.152$, 90%-os megbízhatósággal (*Alpha* 0.1) $\varepsilon = 0.097$ értékéhez jutunk. *Standard_dev* megadott értéke $\sqrt{0.194}$, *Size* pedig 56 kell legyen.

Az átlag konfidencia-intervalluma tehát 99%-os megbízhatósággal $[2.449, 2.753]$, 90%-os megbízhatósággal pedig $[2.504, 2.698]$. A megbízhatóság növekedésével az intervallum hossza egyre nagyobb lesz:

– 99%-os megbízhatósággal (*Alpha* 0.01) $\varepsilon = 0.152$, $m_X \in [2.449, 2.753]$,
– 95%-os megbízhatósággal (*Alpha* 0.05) $\varepsilon = 0.115$, $m_X \in [2.486, 2.716]$,
– 90%-os megbízhatósággal (*Alpha* 0.1) $\varepsilon = 0.097$, $m_X \in [2.504, 2.698]$.

Konfidencia-intervallumok normál eloszlású sokaság esetén, amikor nem ismert a szórás

Az előbbi részben elvégzett levezetés feltételezi σ_X ismeretét, azonban az rendszerint nem ismert. Helyette használhatjuk a σ^* korrigált empirikus szórást, de ekkor a

$$z = \frac{\bar{x} - m_X}{\sigma^* / \sqrt{n}} \qquad (6.23)$$

mennyiség már nem standard normál eloszlású, hanem az egy úgynevezett $n-1$ szabadságfokú Student-eloszlást követő valószínűségi változó lesz, mivel σ^* nem egy szám, hanem egy feltételezetten normál eloszlású valószínűségi változó értéke.

A Student- (vagy „*t*”) eloszlás a valószínűségszámítás fontos eloszlása, amelyet a

$$t = \frac{\sqrt{n}\cdot\eta}{\sqrt{\xi_1+\xi_2+\ldots+\xi_n}} = \frac{\sqrt{n}\cdot\eta}{\chi} \qquad (6.24)$$

valószínűségi változó követ (ez n szabadságfokú eloszlás lesz), ahol ξ_i és η standard normál eloszlású valószínűségi változók. Az ordinátatengelyre szimmetrikus sűrűségfüggvénye

$$h_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\cdot\pi}\cdot\Gamma\left(\frac{n}{2}\right)\cdot\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}, \qquad (6.25)$$

ahol

$$\Gamma(z) = \int_0^\infty t^{z-1}\cdot \mathrm{e}^{-t}\,\mathrm{d}t \qquad (6.26)$$

a már ismert gamma-függvény.

Nos, bevezetvén a

$$t = \frac{\sqrt{n}\cdot(\overline{x}-m_X)}{\sigma^*} \qquad (6.27)$$

jelölést, mely mennyiség bizonyíthatóan $n-1$ szabadságfokú Student-eloszlású valószínűségi változó, annak valószínűsége hogy e mennyiség a $[-\lambda,+\lambda]$ az intervallumon belül tartózkodjon,

$$P\left(-\lambda \le \sqrt{n}\cdot\frac{\overline{x}-m_X}{\sigma^*} < \lambda\right) = 1-p = P(|t|\le\lambda) \qquad (6.28)$$

kell, hogy legyen. E valószínűségnek megfelelő λ meghatározása úgy történik, hogy a Student-eloszlás $h_{n-1}(t)$ sűrűségfüggvényének numerikus integrálásával kiszámítjuk a $H_{n-1}(t)$ eloszlásfüggvény értékeit, majd ezen értékeket pontosan úgy használjuk a számításainkban, mint az előbbi eset standard normál eloszlásának Φ eloszlásfüggvényét. A kézzel történő számítások egyszerűsítése végett a táblázatokban általában nem H értékeit adják meg, hanem olyanokat, amelyek az adott n és p értékeknek megfelelő, $P(|t|\ge t_p) = p = S_n(t_p)$ egyenlőséget kielégítő t_p értéket adják vissza. A táblázatokban nem szereplő mennyiségeket lineárisan interpolálhatjuk. A kiolvasott vagy interpolált $S_{n-1}(\lambda)$ értékkel

$$P(|t| \le \lambda) = 1 - S_{n-1}(\lambda) \qquad (6.29)$$

A táblázatból kiolvasott vagy numerikusan meghatározott λ-val a konfidencia-intervallumot az

$$\overline{x} - \lambda \cdot \frac{\sigma_X}{\sqrt{n}} \le m_X \le \overline{x} + \lambda \cdot \frac{\sigma_X}{\sqrt{n}} \qquad (6.30)$$

egyenlőtlenség írja le. E képletben a korrigálatlan empirikus szórás szerepel.

Excelben a Student-eloszlást a T.DIST() és T.INV() beépített függvényekkel, azok különböző verzióival tanulmányozhatjuk, azonban a konfidencia-intervallumok kiszámításához ebben az esetben is a rendelkezésünkre áll egy speciális függvény: ez a CONFIDENCE.T() függvény, amelynek szintén három argumentuma van:

– *Alpha*: a p valószínűség (1 – a megbízhatósági szint);
– *Standard_dev*: a minta korrigálatlan σ szórása;
– *Size*: a minta elemeinek n száma.

Alkalmazásával, az előbbi példát folytatván, a következő eredményekhez jutunk (*Standard_dev* megadott értéke $\sqrt{0.191}$, *Size* pedig 56):

– 99%-os megbízhatósággal (*Alpha* 0.01) $\varepsilon = 0.156$, $m_X \in [2.445, 2.757]$,
– 95%-os megbízhatósággal (*Alpha* 0.05) $\varepsilon = 0.117$, $m_X \in [2.484, 2.718]$,
– 90%-os megbízhatósággal (*Alpha* 0.1) $\varepsilon = 0.098$, $m_X \in [2.503, 2.699]$.

A konfidencia-intervallumot nemcsak az átlag, hanem a szórás várható értékére is megállapíthatjuk; ekkor a

$$\lambda_1 < (n-1) \cdot \frac{\sigma^{*2}}{\sigma_X^2} < \lambda_2 \qquad (6.31)$$

egyenlőtlenségből indulunk ki. Az egyenlőtlenségben szereplő mennyiség $n-1$ szabadságfokú Pearson-féle χ^2 eloszlást követ. Ez az eloszlás, a Student-t eloszláshoz hasonlóan, fontos eszköze a valószínűségszámításnak: amennyiben ξ_i standard normál eloszlású változók ($i = 1, n$), akkor a $\chi^2 = \sum_{i=1,n} \xi_i^2$ valószínűségi változó n szabadságfokú χ^2 eloszlást követ, melynek sűrűségfüggvénye

$$f_n(x) = \frac{x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)}, \quad \text{ha } x > 0, \tag{6.32}$$

eloszlásfüggvénye pedig ennek az integrálja. A χ^2 változó sűrűségfüggvénye nem szimmetrikus és negatív argumentumokra nincs értelmezve, ezért a konfidencia-intervallumot valamilyen pozitív λ_1 és λ_2 értékeknek megfelelően kell keresgélnünk.

A kézikönyvekben ez esetben olyan táblázatokat kapunk, amelyekből a különböző n és p értékeknek megfelelő, a $P(\chi^2 > x_p) = 1 - F_n(x_p) = p$ egyenlőséget kielégítő x_p-t kapjuk meg. Ekkor λ_1-et és λ_2-t úgy határozzuk meg, hogy teljesüljenek a

$$P(\chi^2 > \lambda_1) = 1 - \frac{p}{2} \text{ és } P(\chi^2 > \lambda_2) = \frac{p}{2} \tag{6.33}$$

egyenlőtlenségek, a σ_X szórás konfidencia-intervalluma pedig

$$\left[\sqrt{n} \cdot \frac{\sigma}{\sqrt{\lambda_2}}, \sqrt{n} \cdot \frac{\sigma}{\sqrt{\lambda_1}} \right] \tag{6.34}$$

lesz. A képletben σ az empirikus szórás korrigálatlan értéke.

Excelben sajnos nincs olyan függvény, amivel a konfidencia-intervallum széleit meghatározó paramétereket közvetlenül kiszámíthatnánk, ezért a χ^2 eloszlásfüggvényhez kell folyamodnunk. A sűrűségfüggvény és az eloszlásfüggvény értékeit a CHISQ.DIST() függvénnyel számíthatjuk ki, amelynek a paraméterei:

– *X*: a valószínűségi változó értéke;
– *Deg_freedom*: a szabadságfokok száma (1 vagy annál nagyobb egész);
– *Cumulative*: *FALSE* értéke a sűrűségfüggvény, *TRUE* értéke pedig az eloszlásfüggvény értékének kiszámítását állítja be.

Az eloszlásfüggvény inverze a CHISQ.INV() függvény, amelynek a paraméterei:

– *Probability*: az a valószínűség, amelynek megfelelően keressük az X változó értékét;
– *Deg_freedom*: a szabadságfokok száma.

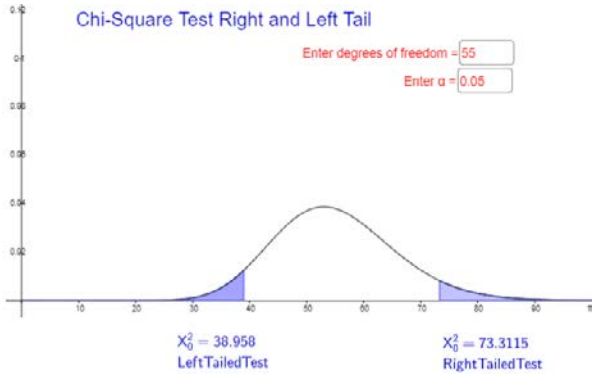
Létezik egy CHISQ.DIST.RT() és egy CHISQ.INV.RT() függvény is, amelyek az eloszlásfüggvény „jobb oldali" („right-tailed") valószínűségével dolgoznak (a p a jobb oldali valószínűségnek a $q = 1 - p$ bal oldali, „left-tailed" valószínűség felel meg).

Példánkban a szabadságfokok száma (*Deg_freedom*) $n - 1 = 55$, a megbízhatósági szinteknek megfelelő *Probability* értékeire a jobb oldali CHISQ.INV.RT() függvénnyel kiszámolt adatok pedig a következők ($\sigma^2 = 0.191$):

megbízhatóság	p	$1-p/2$	λ_1	$p/2$	λ_2	σ_x konfidencia-intervalluma
99%	0.01	0.995	31.735	0.005	85.749	[0.353, 0.581]
95%	0.05	0.975	36.398	0.025	77.380	[0.372, 0.542]
90%	0.10	0.950	38.958	0.050	73.311	[0.382, 0.524]

(megjegyzendő, hogy λ_1 értékét a bal oldali CHISQ.INV () függvénnyel is kiszámolhattuk volna a $p/2$ valószínűségnek megfelelően).

A jobb és bal oldali valószínűségek értelmezéséhez tekintsük a GeoGebrában végrehajtott χ^2-próba ablakát (6.4. ábra). A próba elvégzése során a valószínűségi változónak azt a χ_0^2 értékét keressük, amelyre az a paraméterrel megadott p valószínűséggel a bal oldalon $P(\chi^2 \le \chi_0^2) = p$, a jobb oldalon pedig $1 - P(\chi^2 \ge \chi_0^2) = p$. Ez utóbbi természetesen egyenértékű a $P(\chi^2 \le \chi_0^2) = 1 - p$ feltétellel.



6.4. ábra. *A χ^2-próba GeoGebrában*

A p kétoldali valószínűség (amelyet például a *t*-próbánál alkalmazunk) azt jelenti, hogy a valószínűségi változó $p/2$ valószínűséggel esik a bal oldali

x_b határérték alá, illetve $p/2$ valószínűséggel esik a jobb oldali x_j határérték fölé: $P(X \le x_b) + P(X \ge x_j) = p$, ugyanakkor $P(X \le x_b) = P(X \ge x_j) = p/2$. Nulla átlagú szimmetrikus eloszlások esetén $x_b = -x_j$, és a képleteket némiképp egyszerűsíthetjük.

Konfidencia-intervallumok, amikor a sokaság eloszlása ismeretlen

Ez esetben a konfidencia-intervallumok meghatározása egy jóval tüskésebb probléma. Bebizonyítható, hogy minél nagyobb a minta elemeinek n száma, a

$$t = \frac{\sqrt{n} \cdot (\bar{x} - m_X)}{\sigma^*} \qquad (6.35)$$

valószínűségi változó (amiről az előbbiekben megállapítottuk, hogy Student-eloszlású) eloszlása egyre közelebb áll a normálishoz, függetlenül attól, hogy X milyen eloszlást követ. A normál eloszlás ugyanis a Student-eloszlás határesete, amikor a szabadságfokok száma végtelenül nagy.

Hasonló megjegyzést tehetünk a Pearson-féle χ^2 eloszlásról, amelynek a szabadságfokok végtelenjére számított határesete szintén a normál eloszlás.

Ezzel az észrevétellel, amennyiben a minta száma kellőképpen nagy, az m_X átlag és a σ_X szórás konfidencia-intervallumát az előbbi alfejezetben bemutatott ismeretlen szórású normál eloszlás esetével azonos módon közelíthetjük meg. A „kellőképpen nagy” szám általában legalább 30, de ha az eloszlás nagyon ferde vagy torz, akkor még nagyobb mintát kell számításba venni.

Statisztikai próbák

A konfidencia-intervallumokat egy adott minta alapján kiszámított empirikus jellemzőkkel állapítottuk meg. Felvetődik az a kérdés, hogy az így megállapított konfidencia-intervallumok mennyire elfogadhatóak, azaz mennyire fedik a tényleges átlagos értéket és szórásnégyzetet. Ahhoz, hogy erről meggyőződhessünk, rendszerint egy második minta feldolgozására is szükség van. A feltételezések ilyen ellenőrzését statisztikai próbának nevezzük.

A következőkben lássunk néhány ilyen statisztikai próbát.

Az u-próba (vagy z-próba)

Tegyük fel, hogy az X mennyiség normál eloszlást követ, és annak ismerjük a σ_X szórását és empirikus $\bar{x}$ átlagát. Tegyük fel azt is, hogy bizonyos

megfontolások alapján az X valószínűségi változó átlagos értékére vonatkozóan azt a feltételezést tesszük, hogy $m_X = m_0$. Felvetődik az a kérdés, hogy ez a hipotézisünk helyes-e vagy sem. E feltételezés helyességét az u próbával lehet elvégezni: bebizonyítható, hogy amennyiben a feltételezésünk helyes, akkor az

$$u = \sqrt{n} \cdot \frac{\bar{x} - m_0}{\sigma_X} \qquad (6.36)$$

mennyiség (amit még z-vel is jelölnek) standard normál eloszlású lesz, ahol n a minta elemeinek száma, amelyből $\bar{x}$-et kiszámítottuk. Természetesen a különböző minták alapján kiszámított empirikus $\bar{x}$ átlagok egymástól különbözőek lehetnek, de azok valahol m_X közelében kell legyenek. Ha ezt az m_X értéket megközelítő, feltételezett m_0 átlagot helyesen határoztuk meg, akkor u standard normál eloszlású lesz, ellenben u eloszlása attól különbözni fog. A próba alkalmazása a következőképpen történik: bizonyos $1-p$ szignifikanciaszintnek megfelelően a standard normál eloszlás eloszlásfüggvényéből meghatározzuk a

$$P(|u| < u_p) = 2 \cdot \Phi(u_p) - 1 = 1 - p \qquad (6.37)$$

egyenletet kielégítő u_p értéket (ez a valószínűség azt jelenti, hogy u csak p valószínűséggel esik kívül a $\pm u_p$ értékekkel lehatárolt intervallumon). Amennyiben az így meghatározott u_p mennyiséggel és az előbbi formulával meghatározott u-val felírt

$$-u_p \le u \le u_p \qquad (6.38)$$

egyenlőtlenség igaz, úgy az átlagra tett m_0 hipotézisünk p szignifikancia mellett elfogadható, ellenben a feltételezésünket el kell vetni. Az m_0-ra vonatkozó feltételezés ebben az egymintás próbában nyilván valamilyen gyakorlati megfigyelésre kell, hogy alapozzon, valamilyen okok miatt feltételezzük, hogy X átlagos értéke kb. ennyi kell, hogy legyen. A kísérletezésekben gyakran nincs semmi kapaszkodónk ezen érték megválasztásában, ilyenkor további méréssorozatokat is el kell végeznünk.

Excelben u_p kiszámításához ez esetben is használhatnánk a NORM.S.INV() függvényt, azonban a próba elvégzéséhez létezik egy beépített Z.TEST() függvény is. Ennek három paramétere van, és azt a p valószínűséget adja vissza, amelyre még teljesül a hipotézis:

– *Array*: a mintát tartalmazó tömb X változó értékét;
– *x*: az átlagra tett m_0 hipotézis;
– *Sigma*: a minta szórása. Ez az adat opcionális, és ha azt nem adjuk meg, akkor a mintából kiszámított szórást veszi figyelembe.

Az eddig használt kísérleti adatainkkal ez a függvény a $p = 0.044$ értéket téríti vissza. A visszatérített érték féloldali, így azt kettővel kell megszoroznunk, vagyis az $m_0 = 2.5$ érték 91.2%-os megbízhatósággal fogadható el az X változó átlagos értékeként.

A t-próba

Az *u* próbánál feltételeztük, hogy a valószínűségi változó szórása ismert. Ez a gyakorlatban azonban nem mindig van így, legfennebb a hipotézist alkalmazhatjuk, hogy eléggé nagy n esetében a korrigált szórás jó közelítését adja a tényleges szórásnak.

Ellenben, ha a szórás tényleges értéke ismeretlen, akkor egy másfajta statisztikai próbát, a *t* próbát kell alkalmaznunk. Eszerint, ha a valószínűségi változó átlagára vonatkozó hipotézisünk helyes, akkor a

$$t = \sqrt{n} \cdot \frac{\bar{x} - m_0}{\sigma^*} \qquad (6.39)$$

valószínűségi változó $n-1$ szabadságfokú Student *t*-eloszlást követ, ahol σ^* a korrigált empirikus szórás.

A próba alkalmazása az előbbiekben ismertetett módon történik: bizonyos $1-p$ szignifikanciaszintnek megfelelően a

$$P(|t| < t_p) = 1 - S_{n-1}(t_p) = 1 - p \quad \Rightarrow \quad S_{n-1}(t_p) = p \qquad (6.40)$$

egyenlőségből meghatározzuk t_p értékét és azt összehasonlítjuk az előbbi formulával kiszámított t értékkel. Ha

$$-t_p \le t \le t_p, \qquad (6.41)$$

akkor az átlagra vonatkozó feltételezésünk az adott szignifikanciaszintnek megfelelően helyes, ellenben azt el kell vetnünk.

Itt meg kell jegyeznünk azt, hogy e képletben az $S_n(t)$ függvény nem azonos az eloszlásfüggvénnyel: a gyakorlati szempontokat szem előtt tartva, a *t*-próba táblázatát eleve a „kétoldali" valószínűségeknek megfelelően alkották meg. Emiatt

az *u*-próba 6.37. és a *t*-próba 6.40. összefüggései között különbség van, bár a kettő csak a próbastatisztika (u, illetve t) képletében különbözik egymástól.

Az Excelben van egy T.TEST() függvény, azonban ez a kétmintás *t* próbához való. Emiatt a Student-eloszlás T.INV.2T() függvényét kell használnunk, amely az eloszlásfüggvény egy megadott p kétoldali valószínűségnek megfelelő argumentumát téríti vissza, vagyis azt a határértéket, amelyre $P(X \le -t) = P(X \ge t) = p/2$:

– *Probability*: a valószínűség, amelyre az X változó értékét keressük;
– *Deg_freedom*: a szabadságfokok száma.

95%-os szignifikanciaszint mellett ez a valószínűség $p = 0.05$, példánknak megfelelően a szabadságfokok száma $n-1 = 55$. Ekkor a függvény által visszatérített érték $t_p = 2.004$. Az átlag értékére felállított $m_0 = 2.5$ hipotézisünkre a példa adataival

$$t = \sqrt{n} \cdot \frac{\bar{x} - m_0}{\sigma^*} = \sqrt{56} \cdot \frac{2.601 - 2.5}{\sqrt{0.194}} = 1.716\,, \qquad (6.42)$$

tehát teljesül a $-t_p \le t \le t_p$ feltétel, vagyis e szignifikanciaszint mellett az átlagra vonatkozó hipotézisünk elfogadható.

A kétmintás u-próba

Van amikor azt kell eldöntenünk, hogy két különböző kísérletben tanulmányozott X és Y valószínűségi változók várható értékei tekinthetők-e egymással egyenlőnek. Ekkor, ha a két változó normál eloszlású, és mindkettőnek ismert a szórása, a kétmintás *u*-próbát kell alkalmaznunk, ahol az

$$u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \qquad (6.43)$$

mennyiség eloszlását vizsgáljuk: amennyiben a két változó átlaga (m_X és m_Y) közel azonos, akkor az u mennyiség standard normál eloszlású lesz. A képletben n_X és n_Y a két minta elemeinek száma.

Sokszor X és Y ugyanannak a fizikai mennyiségnek véletlen értékeit takarják, amelyeket két különböző kísérlet során rögzítünk. Ekkor elfogadható az a hipotézis, hogy $\sigma_X = \sigma_Y$, és akkor az előbbi formula némileg leegyszerűsödik.

Az *u*-próba alkalmazása a szokott módon történik, csakhogy u értékét az előbbi képlettel számítjuk ki. Amennyiben a

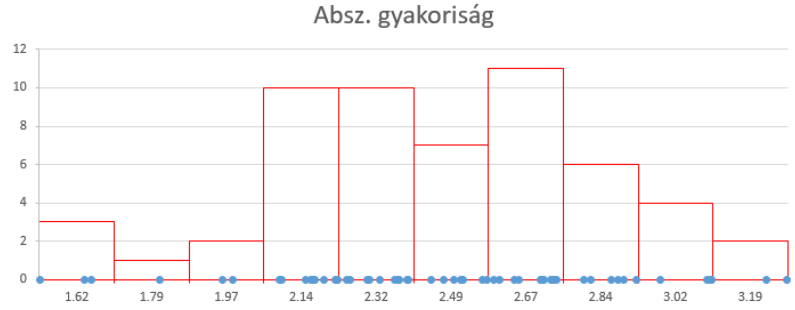
$$-u_p \le u \le u_p \qquad (6.44)$$

feltétel nem teljesül, az azt jelenti hogy X és Y várható értékei között szignifikáns eltérés van: X és Y nem ugyanannak a mennyiségnek a változását jelenti, vagy pedig a kísérleti körülmények megváltozása miatt valamilyen figyelmen kívül hagyott paraméter hatását tapasztalhatjuk.

A példánk folytatásához a már alkalmazott eljárással, az elsővel azonos elméleti átlaggal és szórással rendelkező új adatsort hoztunk létre:

i	1	2	3	4	5	6	7	8	9	10
y_i	1.531	1.636	1.651	1.813	1.958	1.982	2.093	2.097	2.152	2.166
i	11	12	13	14	15	16	17	18	19	20
y_i	2.167	2.173	2.175	2.196	2.222	2.228	2.250	2.258	2.298	2.302
i	21	22	23	24	25	26	27	28	29	30
y_i	2.327	2.361	2.369	2.372	2.392	2.392	2.446	2.477	2.500	2.515
i	31	32	33	34	35	36	37	38	39	40
y_i	2.521	2.566	2.577	2.593	2.606	2.640	2.653	2.704	2.706	2.711
i	41	42	43	44	45	46	47	48	49	50
y_i	2.726	2.729	2.736	2.741	2.805	2.819	2.868	2.884	2.897	2.926
i	51	52	53	54	55	56				
y_i	2.983	3.093	3.100	3.103	3.230	3.280				

Ennek empirikus átlaga $\bar{y} = 2.477$, empirikus szórása $\sigma_Y^2 = 0.152$, a korrigált empirikus szórása pedig $\sigma_Y^2{}^* = 0.155$. Ezt az adatsort is $n_Y = 56$ egyed alkotja, abszolút gyakoriságainak hisztogramja és az egyedek elhelyezkedése a 6.5. ábrán látható. Észrevehetjük, hogy bár ugyanazzal az eljárással hoztuk létre, az eloszlás már nem igazodik olyan szépen a normál eloszlás haranggörbéjéhez.



6.5. ábra. *A második adatsor eloszlása*

Kétmintás *u*-próbához közvetlenül alkalmazható függvény az Excelben nincs, ezért ki kell számítanunk az u mennyiség értékét. Adatainkkal

$$u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} = \frac{2.601 - 2.477}{\sqrt{\frac{0.194}{56} + \frac{0.155}{56}}} = 1.571 \cdot \qquad (6.45)$$

95%-os szignifikanciaszint mellett a standard normál eloszlás eloszlásfüggvényének az $1 - p/2 = 0.975$ valószínűségnek megfelelő u_p argumentumát keressük a NORM.S.INV() függvény alkalmazásával, amelynek az egyetlen paramétere az említett valószínűség. Ez az érték $u_p = 1.960$. Mivel az $-u_p \le u \le u_p$ teljesül, az adott szignifikanciaszint mellett állíthatjuk, hogy a két minta (X és Y) eloszlása azonos lehet (ami pl. az abszolút gyakoriságok hisztogramjainak egyszerű összehasonlításával már nem annyira evidens).

A kétmintás t-próba

A kétmintás *t*-próbát a kétmintás *u*-próbához hasonlóan alkalmazzuk, amikor a két valószínűségi változó szórásai (σ_X és σ_Y) ismeretlenek. Ekkor, amennyiben $m_X = m_Y$ feltételezésünk helyes, a

$$t = \sqrt{\frac{n_X \cdot n_Y \cdot (n_X + n_Y - 2)}{n_X + n_Y}} \cdot \frac{\bar{x} - \bar{y}}{\sqrt{(n_X - 1) \cdot {\sigma_X}^{*2} + (n_Y - 1) \cdot {\sigma_Y}^{*2}}} \qquad (6.46)$$

valószínűségi változó követ $(n_X + n_Y - 2)$ szabadságfokú Student-eloszlást.

Excelben a kétmintás *t*-próba elvégzésére létezik egy T.TEST() függvény, amelynek a használata bonyolultabb az eddigiekénél. E függvény argumentumai a következők:

– *Array1* és *Array2*: a mintákat tartalmazó tömbök;

– *Tails*: ezzel állítjuk be a visszatérítendő valószínűség értelmét, esetünkben ez 2 kell, hogy legyen (a kétoldali valószínűségnek megfelelően; a másik lehetséges beállítás a 1-es érték, amikor a visszatérített érték az iménti fele lenne);

– *Type*: a teszt típusa. A lehetséges értékek: 1 – amikor a két minta egyedszáma azonos, 2 – a két minta szórása azonos, 3 – a két minta szórása eltérő. Ezt az utóbbi típust fogjuk használni.

Adatainkkal a függvény által visszatérített érték $p = 0.120$. Érdekességként, ha az 1-es típust használjuk, akkor a visszatérített érték 0.140, a 2-es típusra pedig

ugyancsak a 0.120 értéket kapjuk. Megjegyzendő, hogy ezek kerekített eredmények, a 2-es és a 3-as típusok közötti különbség az ötödik tizedesnél jelentkezik. A kapott eredmények szerint 86-88%-os szignifikancia mellett lehet elfogadni azt, hogy a két mintát azonos eloszlású valószínűségi változók eredményezték.

A függvény nincs kellőképpen dokumentálva, így eléggé homályos, hogy a beállítások és a visszatérített eredmények mit takarnak, így a biztonság kedvéért a fenti megadott formulával kiszámoljuk t értékét:

$$t = \sqrt{\frac{56 \cdot 56 \cdot (56+56-2)}{56+56}} \cdot \frac{2.601 - 2.477}{\sqrt{(56-1)\cdot 0.194 + (56-1)\cdot 0.155}} = 1.571. \qquad (6.47)$$

95%-os szignifikanciaszint mellett a valószínűség $p = 0.05$, példánknak megfelelően a szabadságfokok száma $n_X + n_Y - 2 = 56 + 56 - 2 = 110$. Ekkor a T.INV.2T() függvény által visszatérített érték $t_p = 1.982$, miszerint a $-t_p \le t \le t_p$ feltétel teljesül, és ekképpen a két mintának ugyanabból az eloszlásból való származásának hipotézise elfogadhatónak bizonyul.

Az F-próba

Az *F*-próbát annak a hipotézisnek a vizsgálatához használjuk, hogy két normál eloszlású alapsokaság szórásai tekinthetők-e egyenlőknek vagy sem (az előbbi kétmintás *u*- és *t*-próbák a két alapsokaság átlagára vonatkoztak).

Bebizonyítható, hogy az

$$F = \frac{{\sigma_X}^{*2}}{{\sigma_Y}^{*2}} \qquad (6.48)$$

valószínűségi változó (Fisher-kritérium), amennyiben X és Y szórása közel azonos, $(n_X - 1)$, $(n_Y - 1)$ szabadságfokú Snedecor-féle F eloszlást követ. A fenti hányados felírásánál úgy kell eljárnunk, hogy az egységnél nagyobb legyen (vagyis ${\sigma_X}^{*2} \ge {\sigma_Y}^{*2}$).

A próbát ugyanúgy alkalmazzuk, mint az előbbieket, az *F*-eloszlás táblázataiból kiolvasott értéket hasonlítjuk össze az előbbi formulával kiszámítottal.

Excelben az *F* eloszlás függvényei megtalálhatók (F.DIST(), F.INV() jobb és bal oldali verzióban), azonban a próba elvégzéséhez a rendelkezésünkre áll a F.TEST() függvény is. Ennek argumentumai *Array1* és *Array2*: a mintákat tartalmazó tömbök.

Példánk mintáit alkalmazván, a függvény által visszatérített valószínűség $p = 0.405$, vagyis a szórások egyenlőségét kb. 41%-os szignifikanciaszint mellett fogadhatjuk el.

Hasonló eredményre jutunk az

$$F = \frac{\sigma_X^{*2}}{\sigma_Y^{*2}} = \frac{0.194}{0.155} = 1.252 \qquad (6.49)$$

érték kiszámításával. A szabadságfokok száma 55, 55, a használt függvények pedig az F.INV() bal és az F.INV.RT() jobb oldali inverz-eloszlásfüggvények, amelyek paraméterei:

– *Probability*: a valószínűség, amelyre keressük az eloszlásfüggvény argumentumát (esetünkben ez a p valószínűség fele);

– *Deg_freedom1* és *Deg_freedom2*: a két szabadságfok.

A visszatérített F_1 bal oldali érték a sűrűségfüggvénynek a $p/2$ valószínűségnek megfelelő argumentuma. A jobb oldali, F_2 érték a $1-p/2$ valószínűségeknek megfelelő argumentumot jelenti. Ha a szignifikanciaszint 80%, akkor, amire $F_1 = 0.935$ és $F_2 = 1.070$. Mivel a kiszámított F értékünk az $[F_1, F_2]$ intervallumon kívül esik, ezen a szinten a szórások egyenlőségének hipotézise nem fogadható el. 40%-os szignifikancia esetén ez az intervallum $[0.800, 1.251]$-re bővül, tehát ez esetben a hipotézisünk már elfogadható.

Mint láthatjuk, a szórások egyenlőségére kapott megbízhatósági szint az átlagokra megállapítottak alatt marad. Ennek az oka az egyedek számának relatív kis száma a minta szórásához viszonyítva.

Illeszkedésvizsgálat

A kísérleti eredmények feldolgozása során a tanulmányozott valószínűségi változó eloszlásával kapcsolatban bizonyos feltételezésekbe bocsátkozhatunk: például az empirikus sűrűségfüggvény és eloszlásfüggvény görbéjének (jobban mondva az azokat megközelítő hisztogramok) alapján vagy valamilyen egyéb megfontolásra vagy analógiára támaszkodva feltételezzük, hogy az valamilyen elméleti (ismert formulával leírt) eloszlást követ. A méréseredmények alapján meghatározott empirikus eloszlás sohasem esik egybe a pontos, elméleti eloszlással, éppen ezért felvetődik az a kérdés, hogy milyen kockázattal fogadható

el az általunk felállított hipotézis. Ennek eldöntésére a χ^2-próbát használjuk, a következőkben leírt módon.

A valószínűségi változó legkisebb és legnagyobb értéke által lehatárolt tartományt a hisztogramok megrajzolásakor n darab egyenlő hosszúságú részintervallumokra, osztályokra osztottuk. Minden osztályba egy bizonyos N_i számú megfigyelt egyed (méréseredmény) esett, mely számot abszolút gyakoriságnak neveztünk. E gyakoriságok összege az elvégzett mérések, megfigyelések számával egyenlő, $\sum N_i = N$. Minden osztályt az x_i középértékével jellemezhetünk, az elméleti sűrűségfüggvényből pedig kiszámítjuk az illető osztályba eső egyedek megjelenésének valószínűségét, ami $p_i \approx f_X(x_i) \cdot \Delta x_i$. Ezzel a valószínűséggel az adott Δx_i szélességű i osztályba eső egyedek pszeudo-elméleti száma

$$N_i^* = N \cdot f_X(x_i) \cdot \Delta x_i \qquad (6.50)$$

kellene legyen, ahol $\sum N_i^* = N$. Az N_i^* mennyiség az elméleti hisztogram egy téglalapjának területét jelenti.

Pontosabb eredményt kapunk az elméleti $F_X(x)$ eloszlásfüggvény használatával: ekkor az i osztályba eső egyedek pszeudo-elméleti számát az

$$N_i^* = N \cdot \left(F(x_{i,f}) - F(x_{i,a})\right) \qquad (6.51)$$

összefüggés adja, ahol az $x_{i,a}$ és az $x_{i,f}$ mennyiségek az osztály intervallumának alsó és felső határát jelentik.

Bebizonyítható, hogy a

$$\chi^2 = \sum_{i=1}^{n} \frac{(N_i - N_i^*)^2}{N_i^*} \qquad (6.52)$$

valószínűségi változó $n-1-r$ szabadságfokú χ^2 eloszlást követ, ahol r a feltételezett eloszlás megbecsült paramétereinek száma (normál eloszlás esetén két megbecsült paraméterünk van, az átlag és a szórás).

Nos, a kiszámított χ^2 mennyiségünkre elvégezzük a χ^2-próbát: valamely megválasztott p szignifikanciaszintnek megfelelően a táblázatából kiolvassuk a $P(\chi^2 > \lambda) = p$ egyenlőséget kielégítő λ értéket. Ha a kiszámított χ^2 érték kisebb, mint a meghatározott λ értéke, akkor az adott szignifikanciaszint mellett a feltételezett elméleti eloszlás alkalmazhatónak bizonyul.

Az ismertetett illeszkedésvizsgálatot bármilyen feltételezett elméleti eloszlás tesztelésére használhatjuk.

Tételezzük fel, hogy a példánkban használt X minta normál eloszlású. Az $m_X = 0.601$, $\sigma_X^2 * = 0.194$ paraméterű normál eloszlás sűrűségfüggvényének felhasználásával a $\chi^2 = 3.029$, az eloszlásfüggvény alapján pedig a pontosabbnak tekinthető $\chi^2 = 3.019$ értékekhez jutunk. A szabadságfokok száma $n-1-r=7$, mivel tíz osztályunk és két becsült paraméterünk van. 95%-os megbízhatósági szintnek megfelelően ($p = 0.95$) a már ismert CHISQ.INV.RT() függvény által visszatérített érték 2.167, ami nagyobb az általunk kiszámítottaknál, így a minta ezen a szinten nem tekinthető normál eloszlásúnak. A feltétel csak 88%-os szinten teljesül.

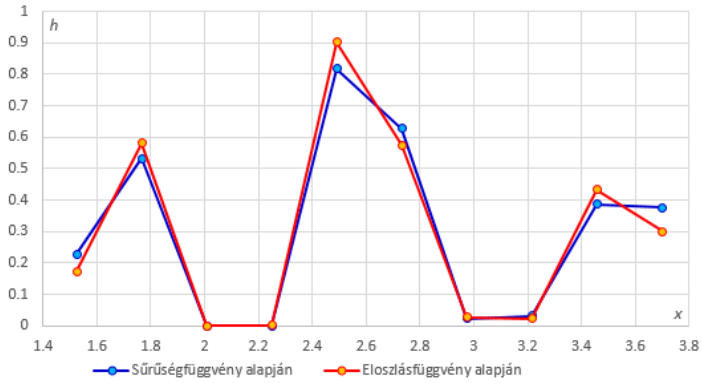
Az Excelben egyébként létezik egy CHISQ.TEST() függvény is, amelynek a két argumentuma *Actual_range* és *Expected_range*, a mintát és az elméleti értékeket tartalmazó tömbök. Ha ezek a tömbök az egyedek valódi, illetve a pszeudo-elméleti számát tartalmazzák, akkor a teszt kb. 96%-os megbízhatósági szintet eredményez.

A durva hibák szűrése

Van ennek az illeszkedésvizsgálatnak egy másik lehetséges hasznosítása is: a durva hibák kiszűrése. Ezek a hibák sokszor igen eltérő értékükkel tűnnek ki (éppen ezért a kismintás méréseknél a legkisebb és a legnagyobb eredményt gyakran figyelmen kívül hagyják), azonban egy pontosabb méréssorozat esetében elég nehezen különíthetők el a véletlen hibáktól.

A véletlen méréshibákról azonban tudjuk (vagy legalábbis meggyőződéssel feltételezzük), hogy valamilyen, pl. normál eloszlást követnek. Ha a rendszeres hibák korrigálása és a szemlátomást durva hibák által érintett elemek kiiktatása után a méréshibákra elvégezzük a χ^2-próba műveleteit, akkor azt tapasztalhatjuk, hogy az szignifikáns eltérést mutat a tényleges normál eloszlástól. Ilyenkor azt feltételezhetjük, hogy a minta elemei között vannak olyanok is, amelyeket nem véletlen, hanem durva hiba befolyásolt, és azokat el kell távolítanunk az adott mintából. Gyakorlatilag ez úgy történhet, hogy a legnagyobb $h_i = (N_i - N_i^*)^2 / N_i^*$ hányadosnak megfelelő osztályokból annyi elemet távolítunk el, hogy a χ^2-próba az adott szignifikancia mellett teljesüljön. Az eltávolított elemeket durva hibák által befolyásolt méréseredményeknek tekintjük. Példánkban e hányadosok

grafikonja a 6.6. ábrán látható: e grafikon alapján a durva hibák keresését az ötödik osztályban kell kezdenünk.



6.6. ábra. *A h_i hányadosok, osztályonként*

7. SZTOCHASZTIKUS FOLYAMATOK

A kutatások során előfordul, hogy a tanulmányozott jelenség egy időben lezajló véletlen folyamat: az ilyen folyamatokat sztochasztikusnak nevezik.

Markov-láncok

A diszkrét sztochasztikus folyamatok során a tanulmányozott mennyiséget leíró változó lépésekben változik. Ezek közül Markov-tulajdonságúnak nevezzük azokat, amelyek esetében a jelenséget leíró változó jövőbeni értéke, az adott jelenlegi állapot mellett, nem függ a múltbeliektől (legalábbis nem közvetlenül). A Galton-deszkán legördülő golyók története is egy ilyen eseményláncolatot példáz: egy adott szeget az oda érkező golyó jobbról vagy balról kerülhet meg, ami befolyásolja annak a további útját, viszont a jövőbeni sorsa nem függ attól, hogy miként jutott el az adott szegig. Az ilyen tulajdonsággal rendelkező *Markov-lánc* a sztochasztikus folyamatok klasszikus alapmodelljeként tekinthető.

A legegyszerűbb verzióban az $A_1, A_2, \ldots A_i, \ldots A_n$ események egy rendszer lehetséges diszkrét állapotai. A rendszer állapotát a szintén diszkrét $t = 0,1,\ldots$ pillanatokban tanulmányozzuk. Bármely pillanatban a rendszer az A_i állapotok valamelyikében található. A rendszer állapotát a k pillanatban (lépésben) tehát egy olyan diszkrét X_k valószínűségi változóval írhatjuk le, amelynek az értéke $X_k = i, i \in [1, n]$, amennyiben ebben a pillanatban a rendszer az i állapotban van (i helyett használhatunk az állapotokat egyértelműen azonosító és azokat mennyiségileg vagy minőségileg leíró, nem feltétlenül számszerű adatokat is).

Annak a valószínűsége, hogy a rendszer a következő, $k+1$-edik pillanatban valamely x állapotban legyen, definíció szerint a következő feltételes valószínűségként írható fel:

$$P(X_{k+1} = x \mid X_k = x_k, X_{k-1} = x_{k-1}, \ldots X_0 = x_0) = P(X_{k+1} = x \mid X_k = x_k). \quad (7.1)$$

Ez az egyenlőség a Markov-tulajdonság képletbe öntött megfogalmazása, miszerint ez a feltételes valószínűség nem függ a múltbeli, k előtti állapotok egyikétől sem.

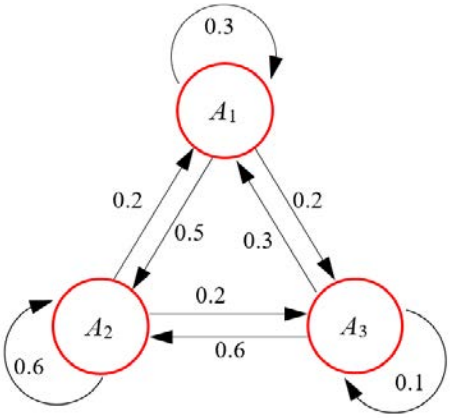
Amennyiben a

$$P(X_{k+1} = x_j \mid X_k = x_i) = p_{i,j} \quad (7.2)$$

átmenet-valószínűség nem függ az időtől, akkor az illető Markov-láncot *stacionárius*nak (vagy homogénnek) nevezik. A $p_{i,j}$ átmenet-valószínűség annak az esélye, hogy a rendszer az i állapotból a következő lépésben a j állapotba kerüljön. Ezeket a valószínűségeket egy mátrixba szokás rendezni.

Nyilvánvaló, hogy az átmenet-valószínűségeket bármely i-re és minden j-re összegezve, az eredmény $\sum_j p_{i,j} = 1$.

Példaként tekintsük azt az esetet, amikor egy termék iránti kereslet heti ingadozását vizsgáljuk. A kereslet lehet alacsony (A_1), átlagos (A_2) és magas (A_3), ami három lehetséges állapotot jelent. A 7.1. ábra az egymást követő heteken tapasztalható átmenetek valószínűségét sematizálja. A nyilakra az egy lépés adott kezdeti és végső állapota közötti átmenet valószínűsége van ráírva. Tegyük fel, hogy a folyamat stacionárius, tehát az átmeneti valószínűségek sohasem változnak meg.



7.1. ábra. *Egy termék iránti kereslet heti alakulását bemutató gráf*

Az átmeneti valószínűségeket egy mátrixba tömöríthetjük, amelynek egy $p_{i,j}$ eleme az i állapotból a j állapotba való átmenet-valószínűsége (i a sor, j az oszlop indexe):

$$[P] = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.6 & 0.1 \end{bmatrix}. \qquad (7.3)$$

Az egylépéses átmenet-valószínűség mintájára meg lehet alkotni a többlépéses átmenet-valószínűség fogalmát is:

$$P(X_k = x_j \mid X_0 = x_i) = p_{i,j}^{(k)}. \qquad (7.4)$$

Ez annak a valószínűsége, hogy az i állapotban levő rendszer k lépés után kerül a j állapotba.

Jelölje $[p^{(k)}]$ azoknak a valószínűségeknek sormátrixát, amelyek azt adják vissza, hogy a k-adik lépésben a rendszer milyen valószínűséggel van valamelyik lehetséges állapotában:

$$[p^{(k)}] = \left[p(X_k = 1) \quad p(X_k = 2) \quad \ldots \quad p(X_k = n) \right]. \qquad (7.5)$$

Ez a folyamatot leíró Markov-lánc eloszlása a k-adik pillanatban. $[p^{(0)}]$ a rendszer kezdeti állapotát, a Markov-lánc kezdeti eloszlását jelenti.

Példánkban a rendszer legyen a kezdetben egy jól meghatározott állapotban, aminek a valószínűsége tehát egységnyi:

$$[p^{(0)}] = \left[1.0 \quad 0.0 \quad 0.0 \right], \qquad (7.6)$$

vagyis az első héten a termék iránti kereslet alacsony. A második héten a kereslet alakulását az egylépéses átmeneti valószínűségekkel a

$$[p^{(1)}] = [p^{(0)}] \cdot [P] = \\ = \left[1.0 \quad 0.0 \quad 0.0 \right] \cdot \begin{vmatrix} 0.3 & 0.5 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.6 & 0.1 \end{vmatrix} = \left[0.3 \quad 0.5 \quad 0.2 \right] \qquad (7.7)$$

szorzat formájában lehet kiszámítani, tehát az 30%-os valószínűséggel marad alacsony, 50%-os valószínűséggel lesz átlagos, és csak 20%-os valószínűséggel lesz magas. Mivel kezdetben a rendszer egy bizonyos, jól meghatározott állapotban volt (az elsőben), az első lépésben a rendszer lehetséges állapotainak valószínűsége a megfelelő átmeneti valószínűségekkel egyenlő (tehát $p_{1,j}$-vel, ahol $j \in [1,3]$).

A harmadik héten, a második lépés végén várható állapotot hasonlóképpen számítjuk ki, a második héten érvényes, megelőző állapotból kiindulva. Az első lépés végén, tehát a második héten, a rendszer a lehetséges állapotainak bármelyikében kerülhetett (a kereslet alakulása bármilyen lehetett). Ezekben az állapotokban egy bizonyos valószínűséggel volt megtalálható (a kereslet bizonyos valószínűséggel volt alacsony, átlagos vagy magas). Amennyiben valamelyik átmeneti valószínűség nem nulla (a példánkban egyik sem az), akkor ebben a

lépésben az bármelyik lehetséges állapotból kiindulva bármelyik lehetséges állapotába eljuthat. A második lépésben annak a valószínűségét, hogy a rendszer valamelyik i állapotból kiindulva a j-be jusson, a $p_i^{(1)}$ valószínűség (amellyel a lépés kezdetén az i állapotban volt) és a $p_{i,j}$ átmeneti valószínűség szorzata adja. Mivel e lépésben a rendszer bármely i állapotból eljuthat a j-be, a $p_i^{(1)} \cdot p_{i,j}$ szorzatokat összegeznünk kell az összes i-re, a lehetőségek teljességének leírásához ezeket pedig ki kell számítanunk minden j-re. Tömören, mátrixos formában ez ekképpen néz ki:

$$[p^{(2)}] = [p^{(1)}] \cdot [P] = \\ = \begin{bmatrix} 0.3 & 0.5 & 0.2 \end{bmatrix} \cdot \begin{vmatrix} 0.3 & 0.5 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.6 & 0.1 \end{vmatrix} = \begin{bmatrix} 0.25 & 0.57 & 0.18 \end{bmatrix}, \qquad (7.8)$$

ami még a következő formában is felírható:

$$[p^{(2)}] = [p^{(0)}] \cdot [P]^2 . \qquad (7.9)$$

A gondolatsor hasonló módon folytatható. Az ötödik héten például

$$[p^{(4)}] = [p^{(0)}] \cdot [P]^4 = \begin{bmatrix} 0.2425 & 0.5757 & 0.1818 \end{bmatrix}, \qquad (7.10)$$

és így tovább. Észrevehetjük, hogy példánkban az idő múlásával egyre valószínűbbé válik az, hogy a kereslet átlagosan alakul, mivel a $[P]$ mátrixban a középső oszlopban vannak a nagyobb valószínűségek.

A példánkban a több- (k-) lépéses átmenet-valószínűségeket a $[P]^k$ mátrix elemei jelentik.

Stacionárius esetben tehát a kezdeti eloszlással és az átmeneti valószínűségek mátrixával a k-adik lépésben a Markov-lánc eloszlása

$$[p^{(k)}] = [p^{(0)}] \cdot [P]^k , \qquad (7.11)$$

mivel $[P]$ időtől független állandókat tartalmaz. Ha a lánc nem stacionárius, akkor az átmeneti valószínűségek minden lépés során megváltozhatnak, és akkor az eloszlást a

$$[p^{(k)}] = [p^{(0)}] \cdot [P(t=k-1)] \cdot [P(t=k-2)] \cdot \ldots \cdot [P(t=0)] = \\ = [p^{(0)}] \cdot \prod_{i=k-1,0} [P(t=i)] \qquad (7.12)$$

szorzattal lehet kiszámítani.

Belátható, hogy ha m egy köztes lépés ($0<m<k$), akkor a többlépéses átmenet-valószínűségek között felírható a

$$p_{i,j}^{(k)} = \sum_{r=1,n} p_{i,r}^{(m)} \cdot p_{r,j}^{(k-m)} \qquad (7.13)$$

összefüggés, ahol n a lehetséges állapotok száma. Ez az összefüggés a Chapman–Kolmogorov-egyenlőség.

Folytonos idejű Markov-láncok

A Markov-lánc időparamétere az eddigiekben diszkrét volt. Ez azonban lehet folytonos is: ekkor a rendszer állapota nem lépésekben, hanem folyamatosan változik. Tekintsük azt az esetet, amikor a lehetséges állapotok továbbra is diszkrétek. A rendszert leíró mennyiség az $X(t)$ valószínűségi függvény, amelynek értékét egy adott t pillanatban diszkrét valószínűségi változóként foghatjuk fel. Az ilyen láncot néha *Markov-folyamat*nak nevezik.

A folyamat átmenet-valószínűségét

$$p_{i,j}(t,t+\Delta t) = P(X(t+\Delta t)=j \mid X(t)=i) \qquad (7.14)$$

formában adhatjuk meg, ahol az állapotokat a t és a $t+\Delta t$ pillanatokban tekintjük. Homogén folyamat esetében $p_{i,j}$ független az időtől, azonban függ a két pillanat között eltelt Δt időtartam hosszától.

Amennyiben $\Delta t \to 0$, az átmenet-valószínűség határértékét az *átmenet intenzitás*ának nevezzük:

$$q_{i,j} = \lim_{\Delta t \to 0} p_{i,j}(t,t+\Delta t) = \lim_{\Delta t \to 0} P(X(t+\Delta t)=j \mid X(t)=i)/\Delta t\,, \qquad (7.15)$$

az intenzitások pedig nem negatív számok, amelyek az i állapotból a j állapotba jutás sebességét adják. Értelmezése szerint $i=j$-re az intenzitás nulla lenne, azonban egyezményesen a matematikai szempontból előnyösebb

$$q_{i,i} = -\sum_{\substack{j=1,n \\ j\neq i}} q_{i,j} \qquad (7.16)$$

értékkel szokták meghatározni, ugyanis ekkor bármely rögzített i-re az intenzitások összege mindig nulla ($q_{i,i}$ az összes többi tag megfordított előjelű összege). Homogén folyamatok esetében az intenzitások állandók.

Az átmenet-valószínűségek mátrixával analóg módon a $q_{i,j}$ intenzitásokat is egy $[Q]$ mátrixban tárolhatjuk.

Jelölje $[p(t)]$ azoknak a valószínűségeknek sormátrixát, amelyekkel a rendszer a t pillanatban valamelyik diszkrét állapotában van:

$$[p(t)] = \left[p(X_k = 1) \quad p(X_k = 2) \quad \ldots \quad p(X_k = n) \right]. \tag{7.17}$$

Bebizonyítható, hogy

$$\frac{\mathrm{d}}{\mathrm{d}t}[p(t)] = [p(t)] \cdot [Q], \tag{7.18}$$

vagyis

$$\frac{\mathrm{d}}{\mathrm{d}t} p_j(t) = q_{j,j} \cdot p_j(t) + \sum_{\substack{i=1,n \\ i \neq j}} q_{i,j} \cdot p_i(t) \cdot \tag{7.19}$$

Annak a valószínűségét, hogy a rendszer a t pillanatban valamely j állapotban legyen, ennek a lineáris differenciál-egyenletrendszernek a megoldása adja. A megoldáshoz szükséges kezdeti feltétel a rendszer kezdeti állapotát megadó $[p(0)]$ kezdeti eloszlás. Elméletileg a megoldást a

$$[p(t)] = \mathrm{e}^{t\cdot[Q]} \tag{7.20}$$

mátrixexponens adja, azonban a tényleges megoldáshoz inkább numerikus eljárásokat használhatunk.

Folytonos állapotterű Markov-láncok

A teljesség kedvéért említhető az az eset, amikor a rendszer állapotát egy folytonos valószínűségi változó írja le, így valamely lépésben vagy pillanatban a rendszer lehetséges állapotainak száma végtelen nagy. Emiatt az átmenet-valószínűséget nem lehet mátrixos formában megadni. A Chapman–Kolmogorov-egyenletet összegzés helyett integrálással írhatjuk fel:

$$p(x_3, t_3 \mid x_1, t_1) = \int_{-\infty}^{+\infty} p(x_3, t_3 \mid x_2, t_2) \cdot p(x_2, t_2 \mid x_1, t_1)\, \mathrm{d}x_2 \cdot \tag{7.21}$$

Ezt a kifejezést a következőképpen kell értelmezni: a bal oldal annak a valószínűségét jelenti, hogy a kezdetben $(t = t_1)$ a valószínűségi változó $X = x_1$ értékével megadott kezdeti állapotából a későbbi $t_3 > t_1$ pillanatban az $X = x_3$ végső állapotba jusson. A jobb oldalon a szorzat két tagja hasonló valószínűségeket ír le, ahol $t_1 < t_2 < t_3$. Maga a szorzat azt a valószínűséget adja, hogy a rendszer az $X = x_2$ értékkel leírt köztes állapoton keresztül jusson el a

kezdeti állapotból a végsőbe, az integrálás pedig összegzi ezeket a valószínűségeket a lehetséges x_2 értékekre.

A Chapman–Kolmogorov-egyenlet tulajdonképpen egy feltételes sűrűségfüggvény definíciója; megállapítható a megfelelő eloszlásfüggvény is.

A gyakorlatban az ilyen folyamatokat nem könnyű kezelni. Megjegyezhető, hogy a számításokat a folytonos időkoordináta helyett inkább lépésekben lehet elvégezni, így tulajdonképpen visszalépünk a diszkrét idejű láncok esetéhez.

Egy sztochasztikus folyamat statisztikai jellemzői

Ez esetben tehát nem egy valószínűségi változó különböző véletlen értékeivel van dolgunk, hanem egy véletlenszerűen változó $X(t)$ függvény különböző időbeni lefutását követjük. Egy i lefutás alkalmával az $X(t)$ függvényt mintavételezzük (technikailag ez bizonyos időlépéssel megmért adatok rögzítését jelenti), a megmért adatok pedig az idő függvényeként, az $X(t)$ véletlen folyamat egy-egy $x_i(t)$ megvalósításából, annak mintavételezéséből származnak.

Az $x_i(t)$ függvények valamely adott $t=\tau$ pillanatban rögzített $x_i(\tau)$ értékei valószínűségi változóknak tekinthetők, így kiszámíthatjuk azok átlagát és szórását, meghatározhatjuk azok sűrűség és eloszlásfüggvényét. Általában (de nem minden esetben) azt tapasztaljuk, hogy ezek a mennyiségek és függvények a megválasztott τ pillanattól függenek, tehát azok is az idő függvényei. Így definiálhatjuk a sztochasztikus folyamat $m_X(t)$ átlagát és $\sigma_X(t)$ szórását, valamint

$$F_X(x,t)=P(x(t)\le x) \tag{7.22}$$

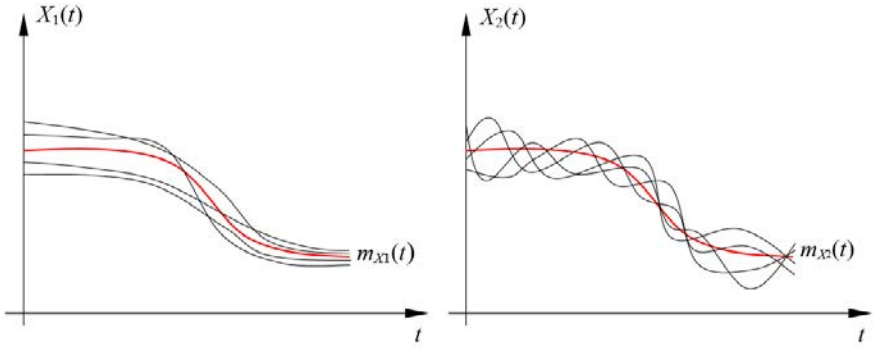
eloszlás- és

$$f_X(x,t)=\frac{\partial F_X(x,t)}{\partial x} \tag{7.23}$$

sűrűségfüggvényét.

Azt gondolhatnánk, hogy ha egy valószínűségi változó leírásakor ezek a mennyiségek teljesen leírják a tanulmányozott mennyiséget, akkor az időbeni változásukat adó függvények is teljesen le kell, hogy írják a megfigyelt sztochasztikus folyamatot. Azonban ez nincs így, és ha megszemléljük a 7.2. ábrát, akkor rájövünk, hogy miért: az ábrán két olyan folyamatot ábrázoltunk, amelyek átlaga, szórása és eloszlása azonos, azonban az időbeni lefutásuk más-más természetű jelenséget takar. A bal oldali ábra görbéi lassan változnak és bár eltérnek az átlagot jelentő görbétől, egy-egy időbeni lefutás az átlagot jelentő

görbét többé-kevésbé jól követő vonalat eredményez. A jobb oldali ábra ezzel szemben egy időben gyorsan változó folyamatot mutat, amelynél egy-egy időbeni lefutás alatt az átlag körüli erős ingadozást tapasztalhatjuk.



7.2. ábra. *Azonos átlagú sztochasztikus folyamatok*

Felvetődik tehát annak az igénye, hogy a valószínűségi változóknál bevezetett mennyiségek mellett további jellemzőket is definiáljunk, amelyek a sztochasztikus folyamat belső szerkezetét jellemezzék.

Autokovariancia és autokorreláció

Az első ilyen lehetőség a kovariancia fogalmának adaptálása lenne.

A kovarianciát és a korrelációt két, egymástól különböző valószínűségi változó kapcsolatának vizsgálatára vezették be. E mennyiségeket a sztochasztikus folyamat két egymástól különböző, t_1 és t_2 pillanatokban rögzített értékeire is megállapíthatjuk, s ezeket *autokovarianciá*nak és *autokorreláció*nak nevezzük:

$$K_X(t_1,t_2)=\frac{1}{n}\cdot\sum_{i=1}^{n}[x_i(t_1)-m_X(t_1)]\cdot[y_i(t_2)-m_Y(t_2)], \qquad (7.24)$$

$$R_X(t_1,t_2)=\frac{K_X(t_1,t_2)}{\sigma_X(t_1)\cdot\sigma_X(t_2)}. \qquad (7.25)$$

Amennyiben a sztochasztikus folyamat autokorrelációja távolabbi t_1 és t_2 pillanatpárosra alacsony, zérushoz közeli érték, akkor az illető folyamat valamilyen időben véletlenszerűen erősen ingadozó jelenséget takar, míg magasabb autokorreláció-értékek kevésbé ingadozó, simább lefutást jelentenek.

Stacionárius és ergodikus folyamatok

Vannak olyan sztochasztikus folyamatok, amelyek átlaga, szórása, sűrűség- és eloszlásfüggvénye időben állandó (azonban maga a folyamat továbbra is véletlenül változik az átlag körül). Ezeket a folyamatokat stacionáriusoknak nevezzük, autokovarianciájuk és autokorrelációjuk csak a t_2 és t_1 pillanatok közötti $\tau = t_2 - t_1$ különbségtől függ.

Amennyiben a stacionárius folyamat különböző pillanatokban rögzített $x(t_i)$ értékekből meghatározott átlag, szórás, eloszlás és autokorreláció azonos a folyamat átlagával, szórásával, eloszlásával és autokorrelációjával, akkor az illető stacionárius folyamatot *ergodikus*nak nevezzük.

Az ergodikus folyamat egyetlen rögzített görbéjéből tehát minden minket érdeklő mennyiséget ki tudunk számítani.

Spektrális sűrűség

Egy időben lejátszódó folyamat belső szerkezetének leírására gyakran a Fourier-transzformációt alkalmazzák. A Fourier-tétel alapján bármely periodikus függvény tetszőleges pontossággal megközelíthető harmonikus (például csak koszinusz) függvények összegeként. Amennyiben ismerjük e harmonikus függvények amplitúdóját és frekvenciáját, következtetéseket vonhatunk le a megközelített függvényre vonatkozólag. Ezt a fajta megközelítést a szerkezetek számításában használják előszeretettel, például a rezonancia fellépésének lehetőségét ellenőrző eljárásokban (ilyenkor a szerkezet sajátfrekvenciájával megegyező harmonikus összetevők amplitúdóját vizsgáljuk). A tételben szereplő Fourier-sor végtelen, és az csak periodikus függvények megközelítésére használható.

A sztochasztikus folyamatok esetében is használhatjuk a Fourier-transzformáción alapuló harmonikus analízist. A gyakorlatban az időbeni folyamat rendszerint nem periodikus, és a számításokban nincs lehetőség végtelen sok harmonikus tagot figyelembe venni. Ilyenkor, a mérnöki számításokban a véges hosszúságú minta hosszát a megközelítendő függvény egyetlen, egész periódusának tekintjük (mintha a rögzített görbe folyamatosan ismétlődne), és csupán csak az első néhány, alacsonyabb frekvenciájú komponenst vesszük figyelembe (ez a közelítés azért fogadható el, mert egyrészt a magasabb sajátfrekvencián történő rezgések a tapasztalat szerint erősen csillapítottak, másrészt az első néhány tag már kielégítő pontossággal megközelíti

a valódi függvényt). Az amplitúdók ábrázolásával egy diagramot kapunk, amely a folyamat különböző frekvenciájú komponenseinek relatív nagyságát ábrázolja. Ez az *amplitúdóspektrum*. Mivel $X(t)$ lefutásairól bizonyos időlépéssel elvégzett mintavételezéssel nyert, tehát diszkrét adatsorok állnak a rendelkezésre, az amplitúdóspektrumot a Fourier-transzformációt annak diszkrét változatában, annak is a „gyors" verziójában (FFT, Fast Fourier Transform) lehet végrehajtani.

A mérnöki gyakorlatban, a sztochasztikus folyamatok leírásában a *spektrális teljesítménysűrűség* használata az elterjedtebb. Maga a fogalom az elektromosságtan területéről, általánosítással származik, ahol eredetileg a jel tényleges teljesítményének frekvencia-összetevőkre való leosztása volt a cél. Átvitt értelemben a „teljesítményen" azt kell érteni, hogy a vizsgált $X(t)$ mennyiség négyzetes átlagának a spektrumára vonatkozik, a „sűrűségen" pedig azt, hogy a spektrum az egységnyi frekvenciájú sávszélességre van normalizálva. A teljesítmény általánosított fogalma tehát nincs konkrét fizikai jelentéssel felruházva. A spektrális teljesítménysűrűség mértékegysége a vizsgált jel mértékegységének a négyzete elosztva a frekvencia mértékegységével. Ha pl. $X(t)$ az egy dinamikai szabadságfokkal rendelkező rendszer gyorsulása, akkor e függvény spektrális teljesítménysűrűségét $(\mathrm{m/s^2})^2/\mathrm{Hz}$ -ben mérjük.

Ebben a meglátásban a vizsgált mennyiség teljesítménye:

$$P = \frac{1}{T} \cdot \int_0^T |X(t)|^2 \, \mathrm{d}t \,, \tag{7.26}$$

ahol T a rögzített adatok teljes időtartama. Ha ezt a képletet általánosítani szeretnénk, akkor az integrálási határokat $\pm\infty$-nek kell vennünk, és ekképpen a teljesítményt a

$$P = \lim_{T\to\infty} \frac{1}{T} \cdot \int_{-\infty}^{+\infty} |X(t)|^2 \, \mathrm{d}t \,, \tag{7.27}$$

határérték formájában definiáljuk.

A Fourier-transzformációval kapcsolatban létezik egy Parseval-féle egyenlőség, miszerint ha $\hat{X}(f)$ az $X(t)$ függvény transzformáltja, akkor

$$\int_{-\infty}^{+\infty} |X(t)|^2 \, \mathrm{d}t = \int_{-\infty}^{+\infty} |\hat{X}(f)|^2 \, \mathrm{d}f \tag{7.28}$$

(esetünkben a fizikai jelentése az, hogy a teljesítmény nagysága független a szemléletmódtól).

E két utóbbi összefüggés alapján a teljesítményt a következőképpen is kifejezhetjük:

$$P = \lim_{T\to\infty} \frac{1}{T} \cdot \int_{-\infty}^{+\infty} |\hat{X}(f)|^2 \, \mathrm{d}f \,. \tag{7.29}$$

Definíció szerint a spektrális teljesítménysűrűség e mennyiség integrandusza:

$$S_X(f) = \lim_{T\to\infty} \frac{1}{T} \cdot |\hat{X}(f)|^2 \,. \tag{7.30}$$

Stacionárius folyamatok esetében a Wiener–Khinchin (magyar helyesírással, fonetikusan „Hincsin”)-tétel értelmében a dolgok leegyszerűsödnek, ugyanis a spektrális sűrűséget az autokorrelációs függvényből is előállíthatjuk, a következő transzformációval:

$$S_X(f) = \int_{-\infty}^{+\infty} R_X(\tau) \cdot \mathrm{e}^{-i\cdot(2\cdot\pi\cdot f)\cdot\tau} \, \mathrm{d}\tau \,, \tag{7.31}$$

$$R_X(\tau) = \int_{-\infty}^{+\infty} S_X(f) \cdot \mathrm{e}^{i\cdot(2\cdot\pi\cdot f)\cdot\tau} \, \mathrm{d}f \,, \tag{7.32}$$

ahol a hatványkitevőben levő zárójel az $\omega = 2\cdot\pi\cdot f$ körfrekvencia. A második összefüggés a Wiener–Khinchin-tételben szereplő inverz transzformáció, amit most csak a teljesség kedvéért írtunk ide.

E tételből egyébként az következik, hogy a nulla frekvenciának megfelelő spektrális teljesítménysűrűség az autokorrelációs függvény grafikonja alatti területtel arányos:

$$S_X(0) = \int_{-\infty}^{+\infty} R_X(\tau) \, \mathrm{d}\tau \,. \tag{7.33}$$

A fordított irányú transzformációból viszont az következik, hogy a spektrális teljesítménysűrűség grafikonja alatti terület az autokorrelációs függvény origóban számított értékével azonos:

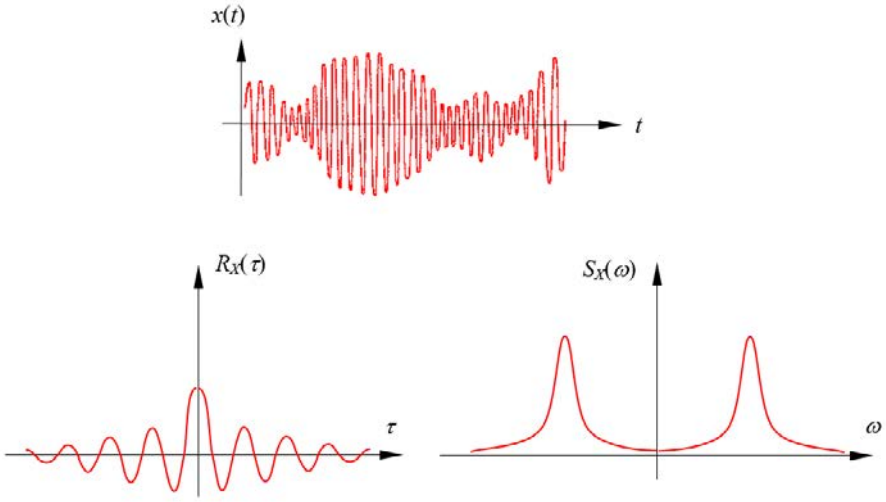
$$\int_{-\infty}^{+\infty} S_X(f) \, \mathrm{d}f = R_X(0) \,. \tag{7.34}$$

$S_X(f)$ értéke az f frekvenciájú komponens $\sigma_X^2(f)$ szórásnégyzetét adja.

A spektrális teljesítménysűrűség-függvény páros, tehát $S_X(-f) = S_X(+f)$, a grafikonja pedig az ordinátára nézve szimmetrikus, és emiatt gyakran csak a pozitív felét rajzolják meg. E függvény pozitív, tehát $S_X(f) \geq 0$, a frekvencia bármely értékére.

A Wiener–Khinchin-transzformációt numerikus integrálással lehet kiszámolni. A végtelenbe nyúló integrálási határokat és a frekvencia tartományát le kell szűkítenünk, azonban a klasszikus numerikus integrálás így is igen sok időt vesz igénybe, úgyhogy ez esetben is a gyors Fourier-transzformáció alkalmazása jelenti a megoldást.

A 7.3. ábrán egy időben lejátszódó folyamatot, annak autokorrelációs függvényét (amit a fenti definíció szerint egy rögzített t_1 és növekvő t_2 értékből álló párokra számítottunk ki), valamint a Wiener–Khinchin-tétel direkt transzformációjával meghatározott spektrális sűrűségét láthatjuk. Észrevehető, hogy a rögzített folyamat tulajdonképpen egy amplitúdómodulált jel, ahol a harmonikus hordozóra egy véletlenül változó jel tevődik rá. Ezért az autokorreláció görbéje meglehetősen elnyúlt, a spektrális sűrűség görbéjének pedig a vivőjel körfrekvenciájának megfelelő pontban határozottan kiemelkedő csúcsa van.



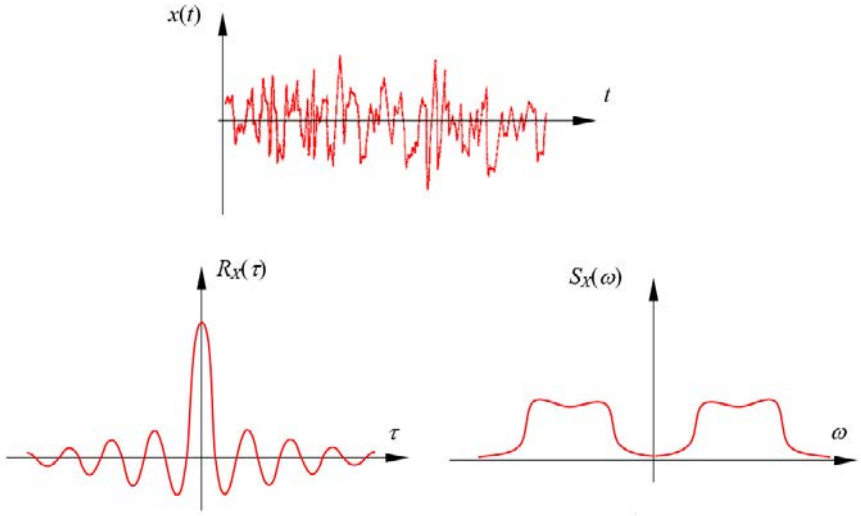
7.3. ábra. *Amplitúdómodulált jel x(t) időbeni lefutása,*
RX(τ) autokorrelációs függvénye és SX(ω) spektrális teljesítménysűrűsége

Más esetben a sztochasztikus folyamat nem rendelkezik kiemelkedő fontosságú (predomináns) harmonikus összetevőkkel, mint a 7.4. ábrán látható „fehér zaj”. A fehér zaj esetében az összetevők ideálisan egyenletesen oszlanak el a teljes frekvenciatartományon. Ennek mintájára a „rózsaszín zaj” esetében az intenzitás a frekvenciával arányosan csökken, „a vörös zaj” pedig a rózsaszínnek

az a változata, amikor a spektrum csak az alacsonyabb frekvenciatartományra korlátozódik. Hasonlóképpen definiálhatunk egyéb „színű” zajokat is.

A fehér zaj esetében az autokorrelációs függvény τ zérus értéke környékén rendelkezik kiemelkedő értékkel, attól távolodva erőteljesen csökken. A spektrális sűrűség görbéje kiemelkedő része viszont egy egész intervallumot felölel, és az ott lapos, nincs kiemelkedő csúcsa.

Amennyiben a fehér zaj harmonikus összetevőinek frekvenciatartománya igen széles (végtelen), akkor az autokorrelációs függvény a $\tau = 0$ pontban egy igen magas (végtelen), tűszerűen kiemelkedő ugrást mutat, a spektrális sűrűség görbéje pedig az abszcisszához közelít (azzal egybeesik).



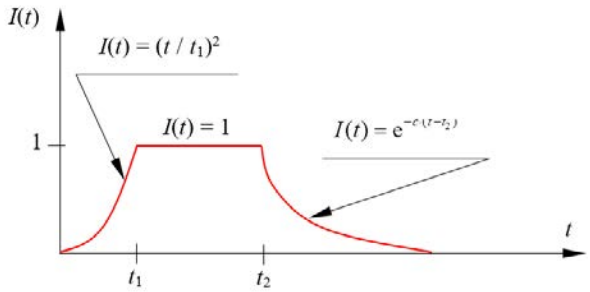
7.4. ábra. *Fehér zaj időbeni lefutása, autokorrelációs függvénye és spektrális sűrűsége*

Normál eloszlású sztochasztikus folyamat szimulálása

A 4.3. fejezet végén megadott átlagú és szórású normál eloszlású pszeudovéletlen számok számítógéppel történő előállításának egy módszerét ismertettük. Ezeket az x_i számokat tekinthetjük egy stacionárius sztochasztikus folyamat valamely Δt lépéssel történő mintavételezéséből származó adatoknak. Ez a folyamat elméletileg az adott átlaggal és szórással fog rendelkezni.

Az így keletkezett adatsort például szintetikus gyorsulásgörbe gyanánt használhatjuk a szerkezetek számításában a földrengések szimulációjához. Ezeknek a görbéknek az átlaga nulla kell, hogy legyen, hiszen az egyensúlyi állapot

körüli lengésekről van szó. A természetben a földrengések gyorsulásgörbéjének van egy nem-stacionárius jellege: annak van egy aránylag gyorsan növekedő amplitúdójú (szórású) kezdeti szakasza, egy rövidebb vagy hosszabb stacionárius szakasza, majd egy aszimptotikusan lecsengő befejező része. Ennek figyelembevétele pl. a 7.5. ábrán látható intenzitásgörbével oldható meg.



7.5. ábra. *Intenzitásgörbe, szintetikus szeizmogram előállításához*

Ez a módszer nagyon egyszerű, de nem biztos, hogy célravezető, ugyanis a szintetikus gyorsulásgörbe egy bizonyos spektrális teljesítménysűrűséggel kellene, hogy rendelkezzen, azonban az ismertetett módszerrel előállított adatsor spektrális teljesítménysűrűsége inkább a fehér zajéhoz fog hasonlítani. A földrengések elméleti tanulmányozása, a természetes folyamatok fizikai hátterének megismerése alapján felmerült a gyorsulásgörbe szűrésének ötlete, amire a legegyszerűbben egy másodrendű lineáris szűrőt lehet alkalmazni:

$$\ddot{x} + 2 \cdot \zeta \cdot \omega_0 \cdot \dot{x} + \omega_0^2 \cdot x = -w(t) \,. \qquad (7.35)$$

Az egyenlet az egy dinamikai szabadságfokkal rendelkező, viszkózus csillapítással rendelkező lineáris rendszer kényszerlengéseinek mozgásegyenlete, amikor azt támaszgerjesztés éri. A képletben ζ (a csillapítási hányados) és ω_0 (a csillapítatlan sajátkörfrekvencia) a szűrő paraméterei, $w(t)$ pedig pl. a Box–Müller-eljárással előállított értékek sorozata által adott, Δt időlépéssel mintavételezett gerjesztés (a bázis gyorsulása).

Az egyenlet numerikus megoldása a szintén Δt időlépéssel mintavételezett $\ddot{x}(t)$ szintetikus gyorsulásgörbe lesz. A megoldás lépésenkénti integrálást jelent. Egy gyakran használt eljárásban (közvetlen implicit integrálás) feltételezzük, hogy egy lépés alatt a gyorsulás lineárisan változik, azaz $\dddot{x} = k$ állandó. A szűrő egyenletében szereplő mennyiségeket integrálással kapott rekurzív összefüggésekkel fejezzük ki, a kinematikából jól ismert összefüggések alapján.

Megállapítjuk a gyorsulás, a sebesség és az elmozdulás kifejezését a lépés elején érvényes értékekre támaszkodva:

$$a = \int k \,\mathrm{d}t = a_0 + k \cdot t \Rightarrow \ddot{x}_{i+1} = \ddot{x}_i + k \cdot \Delta t \rightarrow k = \frac{\ddot{x}_{i+1} - \ddot{x}_i}{\Delta t}, \tag{7.36}$$

$$v = \int a \,\mathrm{d}t = v_0 + a_0 \cdot t + k \cdot \frac{t^2}{2} \Rightarrow \dot{x}_{i+1} = \dot{x}_i + \frac{\Delta t}{2} \cdot (\ddot{x}_i + \ddot{x}_{i+1}), \tag{7.37}$$

$$\begin{gathered} d = \int v \,\mathrm{d}t = d_0 + v_0 \cdot t + a_0 \cdot \frac{t^2}{2} + k \cdot \frac{t^3}{6} \Rightarrow \\ \Rightarrow x_{i+1} = x_i + \Delta t \cdot \dot{x}_i + \frac{\Delta t^2}{3} \cdot \ddot{x}_i + \frac{\Delta t^2}{6} \cdot \ddot{x}_{i+1}. \end{gathered} \tag{7.38}$$

Ha ezeket behelyettesítjük a szűrő differenciálegyenletébe, akkor egy olyan algebrai egyenlethez jutunk, amelyben – ha ismerjük a lépés elején érvényes kinematikai mennyiségeket – az egyedüli ismeretlen a lépés végén érvényes $\ddot{x}_{i+1}$ gyorsulás lesz:

$$\begin{gathered} \left(1 + \zeta \cdot \omega_0 \cdot \Delta t + \omega_0^2 \cdot \frac{\Delta t^2}{6}\right) \cdot \ddot{x}_{i+1} = \\ = w_{i+1} - \zeta \cdot \omega_0 \cdot \left(2 \cdot \dot{x}_i + \Delta t \cdot \ddot{x}_i\right) - \omega_0^2 \cdot \left(x_i + \Delta t \cdot \dot{x}_i + \frac{\Delta t^2}{3} \cdot \ddot{x}_i\right). \end{gathered} \tag{7.39}$$

Ez egy rekurzív egyenlet, amit $\ddot{x}_{i+1}$-ben oldunk meg. A jobb oldalon ismert mennyiségek állnak: w_{i+1} a Δt lépéssel mintavételezett gerjesztés (a szűrő bemeneti függvényének) aktuális értéke, az x_i, $\dot{x}_i$ és $\ddot{x}_i$ értékeket pedig az előző lépésben számítottuk ki. Az első lépésben ez utóbbiakat kezdeti feltételként kell megadnunk, esetünkben a kezdeti nyugalmi állapotból kimozduló rendszernek megfelelő $x_0 = 0$, $\dot{x}_0 = 0$, $\ddot{x}_0 = w_0$ formában.

Miután megkaptuk a gyorsulás új $\ddot{x}_{i+1}$ értékét, a megállapított rekurzív képletekkel kiszámoljuk az $\dot{x}_{i+1}$ sebességet és az x_{i+1} elmozdulást is.

Az ekképpen nyert gyorsulások adatsorának pl. a fenti intenzitásgörbével adhatunk nem-stacionárius jelleget. A stacionárius szakasz spektrális teljesítménysűrűsége

$$S(f) = S_0 \cdot \frac{1}{\omega_0^2} \cdot \frac{1}{\left[1 - \left(\frac{\omega}{\omega_0}\right)^2\right]^2 + 4 \cdot \zeta^2 \cdot \left(\frac{\omega}{\omega_0}\right)^2} \tag{7.40}$$

lesz (ahol $\omega = 2 \cdot \pi \cdot f$), ami igen közel áll a földrengések megfigyeléséből származó Kanai–Tajimi spektrális sűrűséghez. E képletben S_0 / ω_0^2 az $f = 0$ gerjesztő frekvenciának megfelelő érték, a képlet pedig egy dinamikai szabadságfokkal rendelkező, harmonikusan gerjesztett lineáris rendszer rezonanciafüggvényével analóg.

A kipróbáláshoz Excelben hozzuk létre a 7.6. ábrán látható táblázatot.

	A	B	C	D	E	F	G	H	I	J
1	**N**	**n**	**Fehér szám**	**Fehér szám**	**t (s)**	**w (m/s²)**	**a (m/s²)**	**v (m/s)**	**d (m)**	
2	1024	1	0.472684968	0.926868349	0.000	0.178261039	0.178261	0	0	
3		2	0.731763355	0.020730981	0.010	0.023351895	0.020068	0.000992	6.28E-06	
4	**Átlag (m/s²)**	3	0.168248595	0.931460404	0.020	0.019538778	0.014284	0.001163	1.71E-05	
5	0	4	0.224836463	0.76792656	0.030	-0.172734265	-0.17735	0.000348	2.63E-05	
6		5	0.632811073	0.789878107	0.040	-0.309726303	-0.30744	-0.00208	1.87E-05	
7	**Szórás (m/s²)**	6	0.419936131	0.466197722	0.050	0.066788787	0.076136	-0.00323	-1.1E-05	
8	0.7	7	0.162962697	0.40247201	0.060	-0.169195683	-0.15425	-0.00362	-4.3E-05	
9		8	0.268786006	0.706181951	0.070	-0.598942278	-0.56801	-0.00723	-9.4E-05	
10	**Δt (s)**	9	0.425088476	0.130305073	0.080	0.163438261	0.211287	-0.00902	-0.00018	
11	0.01	10	0.295789118	0.778075954	0.090	-0.916856498	-0.84737	-0.0122	-0.00028	
12		11	0.862532237	0.39257111	0.100	-0.438198595	-0.33213	-0.0181	-0.00044	
13	**ζ**	12	0.095295012	0.372215138	0.110	0.453808354	0.58312	-0.01684	-0.00062	
14	0.1	13	0.909787118	0.543035766	0.120	0.165837003	0.305208	-0.0124	-0.00076	
15		14	0.411589857	0.802735273	0.130	-0.482043797	-0.32526	-0.0125	-0.00088	
16	**ω0 (rad/s)**	15	0.370339493	0.672759797	0.140	-0.271252866	-0.08972	-0.01457	-0.00102	
17	12	16	0.139711695	0.364810938	0.150	0.308413768	0.505171	-0.0125	-0.00116	
18		17	0.806343188	0.360407031	0.160	0.165285456	0.366277	-0.00814	-0.00126	

7.6. ábra. *Szintetikus sztochasztikus gyorsulásgörbe létrehozása*

E táblázatban a *C* oszlopban a RAND() függvénnyel 0 és 1 közötti egyenletes eloszlású véletlen számokat hozunk létre, amelyeket átmásolunk a *D* oszlopba, hogy az Excel ne számolja minden frissítés során újra a teljes táblázatot.

Az *E* oszlopban 0-val kezdődően, a megadott Δt lépéssel, létrehozzuk a t_i időkoordinátát (t_i az i-edik a minta vételének a pillanata). A táblázatban 1-gyel kezdődő, a *B* oszlopban levő n sorszámmal $i = n - 1$, $i \in [0, N-1]$.

A *D* oszlopban levő véletlen számokra támaszkodva az *E* oszlopban a már ismert Box–Müller-eljárással normál eloszlású számokat hozunk létre, a megadott átlaggal és szórással. Esetünkben, a feladat természetéből fakadóan az átlagnak most nullának kell lennie. Ezek a normál eloszlású véletlen számok a mintavételezett bemeneti jel w_i értékeit, a gerjesztést jelentik.

A *G*, *H* és *I* oszlopokban a gyorsulás, a sebesség és az elmozdulás *i* pillanatokban kiszámított értékei fognak állni. A legfelső értékek a kezdeti feltételeket jelentik: $a_0 = \ddot{x}_0 = w_0$, $v_0 = \dot{x}_0 = 0$, $d_0 = x_0 = 0$.

A *G* oszlopban, a második értéktől kezdve alkalmazzuk az $\ddot{x}_{i+1}$ gyorsulásra levezetett 7.36. képletet, amely Excel-függvény formájában, a szükséges paraméterek 7.6. ábrán látható lokalizációjával a következőképpen néz ki:

```
=(F3-$A$14*$A$17*(2*H2+$A$11*G2)-
$A$17*$A$17*(I2+$A$11*H2+$A$11*$A$11*G2/3))/
(1+$A$14*$A$17*$A$11+$A$17*$A$17*$A$11*$A$11/6)        (7.41)
```

(némiképp elegánsabb táblázathoz és rövidebb számítási időhöz jutunk, ha a konstans szorzókat külön cellákban számoljuk ki, és a képletekben azokra támaszkodunk).

A *H* oszlopban, szintén a második értéktől kezdve, a sebesség kiszámítását a következő függvénnyel oldjuk meg, a 7.37. képlet alapján:

```
=H2+$A$11*(G2+G3)/2                                     (7.42)
```

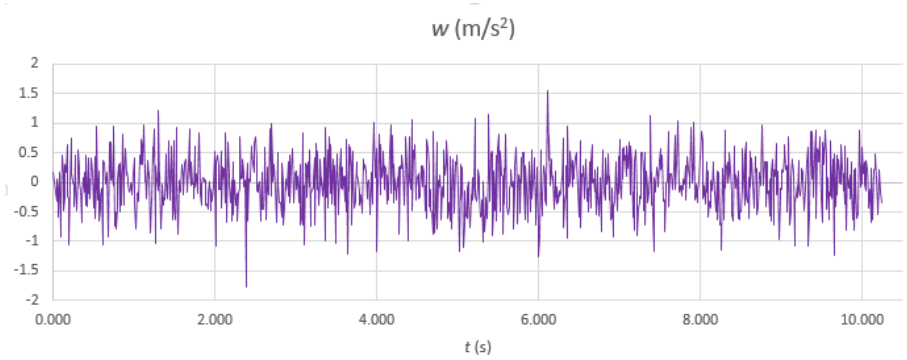
(itt már nincs olyan sok szorzó, de a $\Delta t / 2$ mennyiséget is ki lehetne számítani egy külön cellában). Végül az *I* oszlopban az elmozdulásra az

```
=I2+$A$11*H2+$A$11*$A$11*G2/3+$A$11*$A$11*G3/6          (7.43)
```

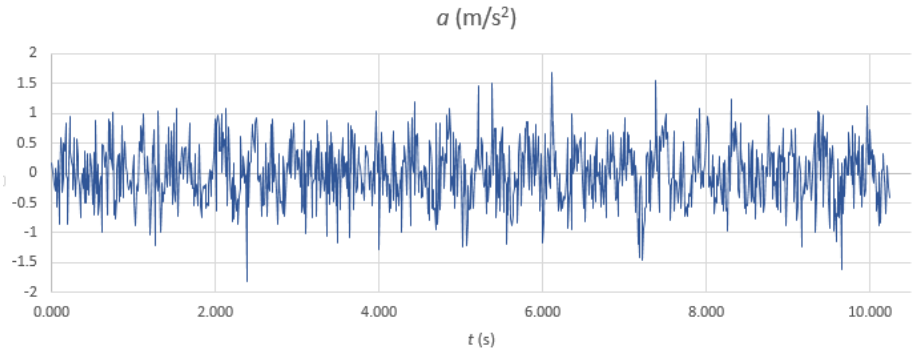
függvényt alkalmazzuk (7.38. összefüggés), amelyet szintén csinosítani lehetne a már elmondott módon.

A számítások eredményét a következő ábrákon láthatjuk: a 7.7. ábrán a szűrő bemeneteként meghatározott gyorsulásgörbét, a 7.8. ábrán a kimeneti, szűrt gyorsulásgörbét, a 7.9. ábrán a sebesség, a 7.10. ábrán pedig az elmozdulás időbeni lefolyását. Mindezeket a diszkrét pillanatokban kiszámolt értékek egyenes vonalakkal való összekötésével kapjuk, úgyhogy ezek tulajdonképpen a tényleges függvények bizonyos pontosságú közelítései.

A bemeneti és a kimeneti gyorsulásgörbék közötti különbséget ránézéssel elég nehéz megállapítani, meggyőzőbb lenne a spektrális teljesítménysűrűségek vizsgálata.

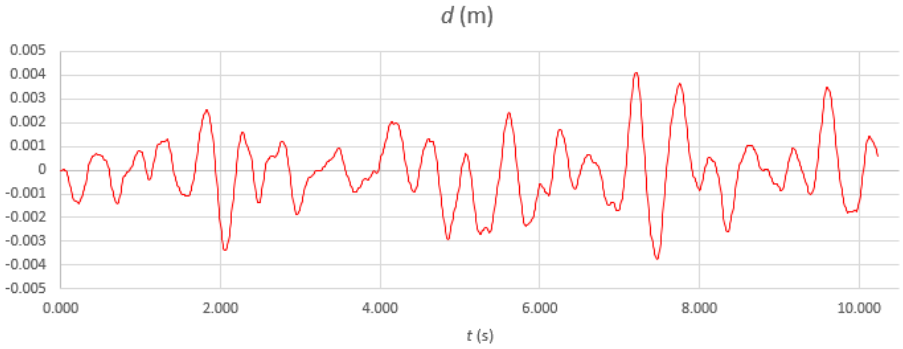


7.7. ábra. *A bemeneti gyorsulásgörbe*



7.8. ábra. *A szűrt, kimeneti gyorsulásgörbe*

7.9. ábra. *A sebességgörbe*



7.10. ábra. *Az elmozdulásgörbe*

A spektrális teljesítménysűrűséget a diszkrét Fourier-transzformációra (FFT) alapozva lehet kiszámítani. A gyors transzformációt a $c(t)$ függvény az állandó Δt lépéssel mintavételezett c_i függvényértékeinken lehet elvégezni. Ezeknek a pontoknak a száma N, amely az algoritmus hátterében álló elvek miatt 2-nek valamilyen egész kitevőjű hatványa kell, hogy legyen. A minta hosszúsága $T = (N-1)\cdot\Delta t$. A transzformáció eredményeként a

$$C_k = \sum_{j=0,N-1} c_j \cdot \mathrm{e}^{2\cdot\pi\cdot i\cdot j\cdot k/N}, \quad k = 0,1,\ldots N-1 \tag{7.44}$$

komplex számokat kapjuk, amelyekkel a spektrális energiasűrűséget a következőképpen közelíthetjük meg:

$$\begin{aligned} S(0) &= \frac{1}{N^2}\cdot|C_0|^2, \\ S(f_k) &= \frac{1}{N^2}\cdot\left[|C_k|^2 + |C_{N-k}|^2\right], \quad k = 1,2,\ldots\left(\frac{N}{2}-1\right), \\ S(f_c) &= \frac{1}{N^2}\cdot|C_{N/2}|^2, \end{aligned} \tag{7.45}$$

ahol

$$f_k = \frac{k}{N\cdot\Delta t}, \tag{7.46}$$

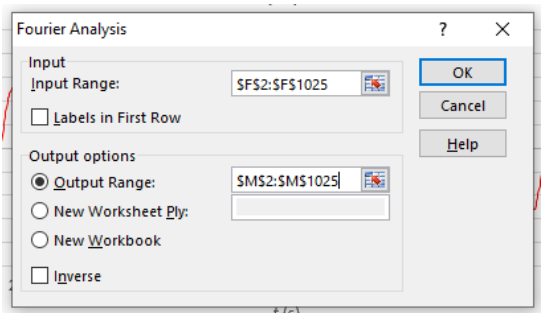
és

$$f_c = \frac{1}{2\cdot\Delta t}. \tag{7.47}$$

Ez utóbbi a Nyquist-frekvencia. A spektrumnak ilyen formán a pozitív felét kapjuk, $1+N/2$ pontban. Mivel a spektrális teljesítménysűrűség páros függvénye a

frekvenciának, a negatív felét (ha arra szükség van) tükrözéssel tudjuk előállítani: $S(-f_k)=S(f_k)$, $S(-f_c)=S(f_c)$.

Ha az elvégzendő számításoknak nem is a legalkalmasabb eszköze, de Excelben az Analysis ToolPak-bővítményben elérhető Fourier-analízishez folyamodhatunk. Az ablakában (7.11. ábra) meg kell adnunk a bemeneti adatok sorát, valamint ki kell jelölnünk egy üres oszlopot a kiszámított diszkrét transzformált értékek tárolására. A bemeneti adatsor pontjainak a száma 2^n kell, hogy legyen, ahol n legalább 1 (legkevesebb két bemeneti adat) és legtöbb 12 (maximum 4096 adat). Ha a bemeneti adatsor elemeinek száma kettőnek nem egész számú hatványa, akkor két dolgot tehetünk: a pontok közötti lineáris interpolációval újra-mintavételezzük az adatokat (ami komplikáltabb), vagy pedig nulla értékeket teszünk a sor végére, hogy azt a legközelebbi egész hatványig töltsük fel (ez a könnyebb megoldás).



7.11. ábra. *Fourier-transzformáció az Analysis ToolPak-ben*

A példánk folytatásaként bemeneti adatsorként megadjuk az *F* oszlopnak 1024 kiszámított w_i értékeket tartalmazó részét (*F2:F1025*), a kimeneti adatok tárolására pedig az *M* oszlopban foglalunk megfelelő nagyságú helyet, amely az 1024 transzformált érték tárolásához elegendő (*M2:M1025*). A számítások elvégzéséhez sok idő szükséges, még egy 1024 elemű adatsor esetében is türelmeseknek kell lennünk, eredményképpen pedig az *M* oszlopban lefoglalt hely $a+b\cdot i$ algebrai formában felírt komplex számokkal telik meg (7.12. ábra).

Ezután az *N* oszlopban kiszámítjuk a komplex transzformáltak $|\sqrt{a^2+b^2}|$ moduluszát, amire célszerűen az Excel IMABS() függvényét használhatjuk.

	K	L	M	N	O	P	Q
1	*k*	*f (Hz)*	*Fourier-transzformált (a + i·b)*	*FT modulusz*	*A*	*B*	*S(f)*
2	0	0	-9.95042303656185	9.95042304	9.950423037	0	9.44242E-05
3	1	0.097656	12.2198278484895+0.732030153637328i	12.2417344	12.24173439	12.24173439	0.000285835
4	2	0.195313	-13.4708399612316+7.27236610453006i	15.3085217	15.30852175	15.30852175	0.000446989
5	3	0.292969	9.18045370829178-11.3161754039212i	14.5717726	14.57177258	14.57177258	0.000405
6	4	0.390625	-17.5262964278304+14.2520620118297i	22.5896511	22.58965113	22.58965113	0.000973305
7	5	0.488281	2.54290888219356+7.92834210556576i	8.32616323	8.32616323	8.32616323	0.000132227
8	6	0.585938	-24.4777140075417+4.92956858203707i	24.9691636	24.96916357	24.96916357	0.001189154
9	7	0.683594	3.63364378407681+1.151179518970003i	3.81163763	3.811637632	3.811637632	2.77111E-05

7.12. ábra. *A spektrális teljesítménysűrűség kiszámítása*

A spektrális teljesítménysűrűséget a 7.45. összefüggések általánosításával kapott

$$S(f_k)=\frac{1}{N^2}\cdot\left[|C_k|^2+|C_{N-k}|^2\right] \quad \rightarrow \quad S_k=\frac{1}{N^2}\cdot\left[A^2+B^2\right] \qquad (7.48)$$

képlettel lehet kiszámítani($k\in[0,N/2]$), ahol a C_k, C_{N-k} moduluszokat A-val és B-vel jelöltük. Az első, $S(0)$ érték kiszámításához A a C_0 értéknek felel meg, míg az utolsó, $S(f_c)$ értékhez A-nak $C_{N/2}$-t kell vegyünk; mindkét esetben $B=0$.

A könnyebb tájékozódás érdekében a k index értékeit a *K* oszlopban tüntettük fel.

Összesen $N/2+1$ spektrális értéket számítunk ki, ezek közül az első a *2*, az utolsó pedig az *514* sorban van, mivel a példánkban $N=1024$. Az A tényezőket az *O*, a B tényezőket pedig a *P* oszlopban tároljuk.

Az A paraméterek meghatározása egyszerű, mert azok minden esetben az illető sorban levő komplex szám moduluszával azonosak.

A B értékeket a legfelső ($k=0$) és a legalsó ($k=N/2$) cellákban nullázzuk, a kettő között pedig azt az

=INDEX($N$1:$N$1025,$A$2+2-K3) (*7.49*)

függvény segítségével keressük ki a moduluszok *N* oszlopából. Képletünk szerint, a k-adik sorban levő B-nek az $N-k$-adik modulusz felel meg, így az INDEX() függvénnyel a moduluszok *N1*:*N1025* tartományából alulról a k-adik elemet kell megkeresnünk.

A legfelső cellában szereplő spektrális érték tehát végső soron az

=(O2*O2+P2*P2)/($A$2*$A$2) (*7.50*)

képlettel kerül kiszámításra; az *A2*-es cellában N értéke áll.

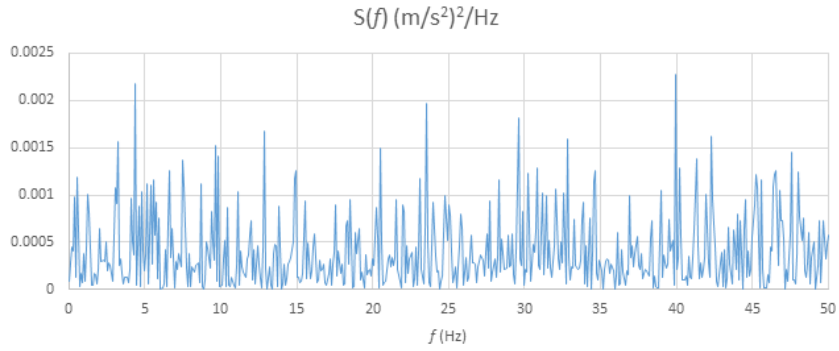
Az ezekhez az értékekhez tartozó frekvenciát az $f_k = k/(N \cdot \Delta t)$ képletnek megfelelően az *L* oszlopban a

=K2/($A$2*$A$11) (*7.51*)

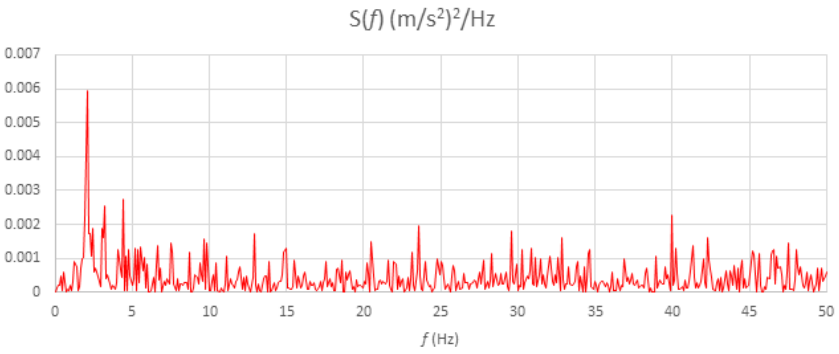
formulával adhatjuk meg (ez a legfelső cellában érvényes képlet), ahol k -t osztjuk az *A2* cellában levő N és az *A11* cellában levő Δt szorzatával. Ez érvényes a legfelső, nulla frekvenciájú spektrális értékre is, meg a legalsó, f_c Nyquist-frekvenciájú pontra is.

Soronként lefelé haladva ugyanígy számítjuk ki a többi f_k és S_k mennyiségeket.

A spektrális értékeket a frekvencia függvényében ábrázolva a 7.13. ábrán látható grafikonhoz jutunk. Mivel ez a bemeneti „fehér" zaj, az ekképpen megközelített spektruma lapos és aránylag egyenletes.

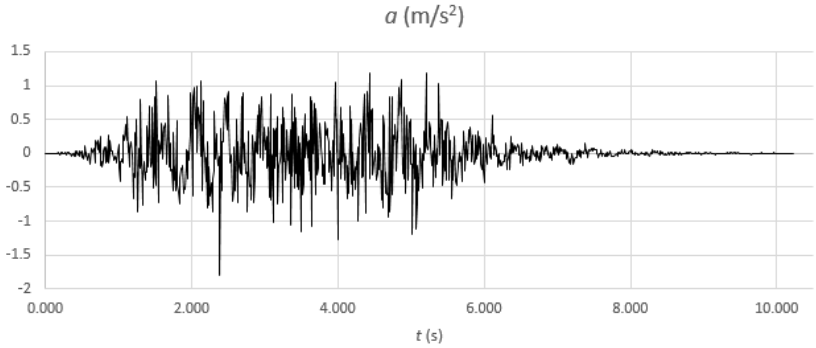


7.13. ábra. *A bemeneti fehér zaj spektrális teljesítménysűrűsége*



7.14. ábra. *A kimeneti szintetikus gyorsulásgörbe spektrális teljesítménysűrűsége*

Ugyanígy számíthatjuk ki a szintetikus gyorsulásgörbe spektrumát is (a bemeneti adatokat a *G* oszlopból kell vennünk). Ebben az esetben a spektrumnak a szűrő $f_0 = \omega_0 /(2\cdot\pi) = 12/(2\cdot\pi) \approx 1.9\ \text{Hz}$ sajátfrekvenciájának környékén egy kimagasló csúcsa van (7.14. ábra).



7.15. ábra. *Nem-stacionárius szintetikus gyorsulásgörbe*

Végül, ha a gyorsulásgörbének nem-stacionárius jelleget szeretnénk adni, akkor a 7.5. ábrán látható intenzitásgörbe alkalmazásával, ahol $t_1 = 1.5\,\text{s}$, $t_2 = 5\,\text{s}$, $c = 1$, a 7.15. ábrán látható szeizmogramhoz jutunk. Ehhez a szűrő kimenetén kapott, a *G* oszlopban tárolt stacionárius $a(t_i)$ gyorsulásokat meg kell szorozni az intenzitásgörbe kiszámított $I(t_i)$ értékeivel.

IRODALOMJEGYZÉK

A valószínűségszámítás és a statisztika szakirodalma magyar nyelven is bőséges, és az interneten is számos, e témakörben megírt tankönyv, jegyzet, dolgozat érhető el. E művek egy része kimondottan a matematikai alapokat részletezi, más részük pedig alkalmazásszintű leírásokat tartalmaz. A ritkábban használt fogalmak tisztázására azonban magyar nyelven nem mindig találunk elegendő információt, az alkalmazásokhoz pedig példát, emiatt a hazai viszonylatokhoz mérten az esetleg könnyebben érthető román, illetve az angol nyelvű művek tanulmányozása is elkerülhetetlen lesz. Az alábbi felsorolás a teljesség igénye nélkül készült, és inkább a közérthetőbb, alapvető műveket tartalmazza.

[1] Cseke Vilmos, *A valószínűségszámítás és gyakorlati alkalmazásai*, Kolozsvár: Dacia Könyvkiadó, 1982.

[2] Dan Lungu, Dan Ghiocel, *Metode probabilistice în calculul construcţiilor*, Bucureşti: Editura Tehnică, 1982.

[3] Denkinger Géza, *Valószínűségszámítás*, Budapest: Tankönyvkiadó, 1978.

[4] Denkinger Géza, *Valószínűségszámítási gyakorlatok*, Budapest: Tankönyvkiadó, 2000.

[5] Edwin.T. Jaynes, *Probability Theory: The Logic Of Science*, Cambridge University Press, 2004.

[6] Granino A. Korn, Theresa M. Korn, *Matematikai kézikönyv műszakiaknak*, Budapest: Műszaki Könyvkiadó, 1975.

[7] Gyéresi Ştefan, *Statistică matematică*, Bucureşti: Editura Paco, 2003.

[8] OctavOnicescu, Vasile Ştefanescu, *Elemente de statistică informaţională cu aplicaţii*, Bucureşti: EdituraTehnică, 1979.

[9] R. Lowell Wine, *Statistics for Scientists and Engineers*, Englewood Cliffs, N. J, New Jersey: Prentice Hall, 1990

[10] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes in Fortran 77: The Art of Scientific Computing, Second Edition*, Cambridge University Press, 1997

[11] Reimann József, *Valószínűségelmélet és matematikai statisztika mérnököknek*, Budapest: Tankönyvkiadó, 1992.

[12] T. T. Soong, *Fundamentals of Probability and Statistics for Engineers*, eBook, 2004.

AZ ANGOL ÉS A MAGYAR NYELVŰ FÜGGVÉNYEK MEGFELELTETÉSE

Angol	Magyar	Angol	Magyar
AND	ÉS	LOGNORM.INV	LOGNORM.INVERZ
AVERAGE	ÁTLAG	MAX	MAX
BINOM.DIST	BINOM.ELOSZL	MIN	MIN
BINOM.INV	BINOM.INVERZ	MOD	MARADÉK
CHISQ.DIST	KHINÉGYZET.ELOSZLÁS	NORM.INV	NORM.INVERZ
CHISQ.DIST.RT	KHINÉGYZET.ELOSZLÁS.JOBB	NORM.S.DIST	NORM.S.ELOSZLÁS
CHISQ.INV	KHINÉGYZET.INVERZ	NORM.S.INV	NORM.S.INVERZ
CHISQ.INV.RT	KHINÉGYZET.INVERZ.JOBB	NORMD.IST	NORM.ELOSZLÁS
CHISQ.TEST	KHINÉGYZET.PRÓBA	NOT	NEM
CONFIDENCE.NORM	MEGBÍZHATÓSÁG.NORM	OR	VAGY
CONFIDENCE.T	MEGBÍZHATÓSÁG.T	PI	PI
CORREL	KORREL	POISSON.DIST	POISSON.ELOSZLÁS
COS	COS	POWER	HATVÁNY
COUNTIF	DARABTELI	RAND	VÉL
EXP	KITEVŐ	RANK.AVG	RANG.ÁTL
EXPON.DIST	EXP.ELOSZL	SIN	SIN
F.DIST	F.ELOSZL	SQRT	GYÖK
F.DIST.RT	F.ELOSZLÁS.JOBB	STDEV.P	SZÓR.S
F.INV	F.INVERZ	STDEV.S	SZÓR.M
F.INV.RT	F.INVERZ.JOBB	SUM	SZUM
GAMMA	GAMMA	T.DIST	T.ELOSZLÁS
HYPGEOM.DIST	HIPGEOM.ELOSZLÁS	T.DIST.2T	T.ELOSZLÁS.2SZ
IF	HA	T.DIST.RT	T.ELOSZLÁS.JOBB
IMABS	KÉPZ.ABSZ	T.INV	T.INVERZ
INDEX	INDEX	T.INV.2T	T.INVERZ.2SZ
INT	INT	T.TEST	T.PRÓB
LOG	LOG	WEIBULL.DIST	WEIBULL.ELOSZLÁS
LOGNORM.DIST	LOGNORM.ELOSZLÁS	Z.TEST	Z.PRÓB

PROBABILITY AND STATISTICS IN TECHNICAL ENGINEERING

ABSTRACT

The book introduces the essential elements of the probability theory and those of statistics. The main goal is to clarify the basic concepts, approached with emphasis on the application side and less on mathematical theory. Those presented in the book are addressed primarily, but not exclusively, to students with technical specializations, laying the practical foundations of their training, but it also can serve as starting point for in-depth theoretical studies in this field.

The first chapter of the book defines the event and presents the operations with events, these concepts being clarified by the examples of dice throwing and that of shooting at a target.

The second chapter introduces the concept of random event and that of the probability. The third chapter is focused on the random variables and their characterization. This chapter introduces some fundamental terms such the probability density function and the cumulative distribution function, expected value and variance.

The fourth chapter presents the practically most important distributions, these being illustrated by easy-to-follow examples done in Excel. Among these distributions are included the most important discrete and continuous distributions, such as the binomial distribution, the hypergeometric distribution and the normal one. There are also included those used to model the distribution of the extrema of various random phenomena.

The fifth chapter introduces the multidimensional distributions. The theoretical concepts are applied in practice through some simple examples. Thus, the case of independent variables is exemplified by the practical problem of load-strength relation in structure's design. The aspects and concepts regarding the case of dependent variables are illustrated by an example using a not too complicated bivariate distribution function.

The thematic of the sixth chapter is the statistical processing of experimental data. The theoretical concepts are applied in some examples taking some simulated random data as the input values in this calculations. The average and the empirical variance are computed, the confidence intervals are established, after then various statistical tests are performed.

In the seventh chapter the stochastic processes are introduced and their main statistical characteristics are described. Markov chains, which underlie the theory of these processes, are presented by a simple discrete example. Markov processes with the continuous variable are not detailed, because they cannot be treated at an elementary level.

CONTENTS

WAHRSCHEINLICHKEIT UND STATISTIK IN DER TECHNIK

ZUSAMMENFASSUNG

Im vorliegenden Buch werden die Grundlagen der Wahrscheinlichkeitsrechnung und Statistik behandelt. Zielsetzung des Buches ist die Klärung der Grundbegriffe mit Schwerpunkt auf die Anwendungsbereichen. Die mathematischen Theorien werden knapp dargestellt.

Der Band richtet sich vor allem an Ingenieurstudierende an Universitäten und trägt zum Aufbau der Praxiskenntnisse bei. Gleichzeitig kann er auch als ein Ausgangspunkt für vertiefte theoretische Auseinandersetzungen der Themen dienen.

Das erste Kapitel befasst sich mit den Ereignissen und Operationen über Ereignismengen. Die Begriffe werden anhand des Beispiels des Würfelwurfes genommen und zielgerichtet erklärt.

Das zweite Kapitel behandelt die Grundbegriffe der zufälligen Ereignisse und Wahrscheinlichkeit.

Das dritte Kapitel fokussiert auf die Beschreibung von Zufallsvariablen und deren Eigenschaften. Auch in diesem Kapitel werden Grundbegriffe wie Wahrscheinlichkeitsdichtefunktion, Verteilungsfunktion, Mittelwert und Varianz vorgestellt.

Das vierte Kapitel stellt aus der praktischer Sicht wichtige Wahrscheinlichkeitsverteilungen durch einfache Excel-Anwendungen vor. Zu diesen Verteilungen zählen die wichtigsten diskreten und stetigen Verteilungen, wie z. B. die Binomial-, die Normalverteilung, die hypergeometrische Verteilung. Auch Verteilungen, die zum Modellieren der Wahrscheinlichkeit von Extrema dienen, werden in diesem Abschnitt erklärt.

Das fünfte Kapitel behandelt die mehrdimensionalen Verteilungen. Die theoretischen Konzepte werden durch Beispiele erläutert. Unabhängige Variablen lassen sich anhand einer Strukturbemessung erklärt, abhängige Variablen werden anhand einer einfachen bivariaten Verteilungsfunktion illustriert.

Die Thematik des sechsten Kapitels ist die statistische Verarbeitung experimenteller Datenreihen. Die praktische Umsetzung der Theorie wird durch die Bearbeitung einer simulierten Datenreihen dargestellt. Hier werden der

empirische Median und die empirische Varianz berechnet, Konfidenzintervalle bestimmt, verschiedene statistische Proben durchgeführt.

Das siebte Kapitel bietet eine Einführung in die stochastischen Prozesse und stellt deren wichtigste statistische Merkmale vor. Markov-Ketten werden mittels eines einfachen Beispiels dargelegt. Da die Markov-Prozesse mit kontinuierlichen Variablen nicht auf elementarem Niveau behandelt werden können, werden sie nicht detailliert.

INHALTSVERZEICHNIS

TEORIA PROBABILITĂȚILOR ŞI STATISTICA MATEMATICĂ APLICATE ÎN INGINERIE

REZUMAT

Cartea introduce elementele esenţiale ale calculului probabilităţilor şi statisticii matematice. Scopul principal urmărit este clarificarea concepţiilor de bază, abordat cu accentul pe latura aplicativă şi mai puţin pe teoria matematică. Cele prezentate în carte se adresează în primul rând, dar nu exclusiv, studenţilor cu specializări tehnice, punând bazele practice ale pregătirii lor, dar totodată pot servi ca punct de plecare în studiile teoretice aprofundate în acest domeniu.

Primul capitol al cărţii defineşte evenimentul şi prezintă operaţiile cu evenimente, conceptele fiind clarificate prin exemplele aruncării cu zaruri şi tirului asupra ţintei.

Capitolul al doilea introduce conceptul evenimentului aleatoriu şi cel al probabilităţii, iar capitolul trei focusează asupra variabilele aleatoare şi asupra caracterizării lor. În acest capitol sunt introduse termeni fundamentali ca densitatea de repartiţie şi funcţia repartiţiei, valoarea medie şi dispersia.

Capitolul patru prezintă repartiţiile importante din punctul de vedere practic, acestea fiind ilustrate prin exemple uşor de urmărit realizate în Excel. Între acestea regăsim cele mai importante repartiţii discrete şi continue, cum ar fi repartiţia binomială, cea hipergeometrică şi normală, dar totodată sunt detaliate şi repartiţiile folosite mai frecvent în modelarea probabilităţilor extremelor.

Capitolul cinci este consacrat repartiţiilor multidimensionale, elementele teoretice fundamentale fiind aplicate în practică în cadrul unor exemple simple. Astfel cazul variabilelor independente este exemplificat prin problema de ordin practic al dimensionării structurilor, iar aspectele şi concepţiile referitoare cazului variabilelor dependente sunt ilustrate printr-un exemplu cu funcţia de repartiţie bivariată nu prea complicată.

Tematica capitolului şase este prelucrarea statistică a datelor experimentale, unde aplicarea teoriei se face prin calcule făcute pe un şir de date obţinut prin simulare. Se calculează media şi dispersia empirică, se stabilesc intervalele de confidenţă, apoi se efectuează diverse probe statistice.

În al şaptelea capitol sunt introduse procesele stohastice şi sunt stabilite principalele caracteristici statistice ale lor. Lanţurile Markov, ce stau la baza teoriei acestor procese, sunt prezentate printr-un exemplu discret simplu.

Procesele Markov cu variabila continuă nu sunt detaliate, fiindcă acestea nu pot fi tratate la un nivel elementar.

CUPRINS

A sorozat eddig megjelent kötetei

1. Jodál Endre: *Számítástechnika az ezredforduló küszöbén*. 1992.
2. Pálfalvi Attila: *Porkohászat*. 1993.
3. Bagyinszki Gyula – Bitay Enikő: *Bevezetés az anyagtechnológiák informatikájába.* 2007.
4. Bitay Enikő: *Lézeres felületkezelés és modellezés.* 2007.
5. Bagyinszki Gyula – Bitay Enikő: *Felületkezelés*. 2009.
6. Forgó Zoltán: *Bevezetés a mechatronikába*. 2009.
7. Tolvaly-Roşca Ferenc: *A számítógépes tervezés alapjai. AutoLisp és Autodesk Inventor alapismeretek*. 2009.
8. Bagyinszki Gyula – Bitay Enikő: *Hegesztéstechnika I. Eljárások és gépesítés.* 2010.
9. Bagyinszki Gyula – Bitay Enikő: *Hegesztéstechnika II. Berendezések és mérések.* 2010.
10. Máté Márton: *Műszaki mechanika – kinematika*. 2010.
11. Bitay Enikő: *Anyagtudományi laboratórium I. Tulajdonságminősítő vizsgálatok.* 2011.
12. Máté Márton: *Hengeres fogaskerekek gyártószerszámai.* 2016.
13. Tolvaly-Roşca Ferenc: *Gépelemek*. 2019.
14. Papp István: *Mechanizmusok optimális kiegyensúlyozásának elmélete*. 2020.

Előkészületben

Bitay Enikő: *Anyagtudományi laboratórium II. Anyagszerkezettani vizsgálatok*.

Gergely Attila: *Bevezetés a polimer anyagok feldolgozásába*.

Máté Márton: *Forgácsoló szerszámok tervezése*.

.

Könyvünkben a valószínűségszámítás és a statisztika elemeivel ismerkedhetünk meg. A könyvben bemutatottak elsősorban a mérnöki szakokat hallgató diákok alkalmazásszintű ismereteinek megalapozását segítik elő, de ugyanakkor kiindulási pontot jelenthetnek a mélyebb elméleti fejtegetésekhez is. Az elsődleges cél a témakör alapvető fogalmainak a tisztázása, amelyet az alkalmazásuk szemszögéből és nem annyira a matematikai oldalról közelítünk meg. Ekképpen az eseményekkel, a valószínűségekkel, a valószínűség-eloszlásokkal, a sztochasztikus folyamatokkal és a kísérleti adatok feldolgozásával kapcsolatos alapok bemutatását könnyen követhető példákon keresztül illusztráltuk, a szükséges számításokat pedig Excel-alkalmazások segítségével végeztük el.

ISBN 978-606-739-181-7