Valószínűségszámítás és matematikai statisztika

Segédanyag a Matematika 3. tantárgy második blokkjához

Szalay Krisztina

Bevezetés

Ez a jegyzet a Széchenyi István Egyetem mérnöki BSc-hallgatói számára készült a Matematika 3. tárgy második blokkjához. Bevezetést ad a matematikai statisztikába és áttekinti a szükséges valószínűségszámítási ismereteket. A jegyzet az elméleti ismeretek rövid ismertetése mellett feladatok megoldásán keresztül mutatja be a fogalmakat és a megoldási módszereket. Az egyes fejezetek az elmélet áttekintése mellett tartalmaznak példákat, amelyek a fogalmak megértését szolgálják, illetve kidolgozott feladatokat, amelyek az elméleti ismeretek alkalmazását mutatják be. A egyes fejezetek egymásra épülnek, javasolt azokat a megadott sorrendben feldolgozni.

A fejezeteken belül az egyes alfejezetek (pl. 7.1) egy önálló tanulási egységet jelentenek, célszerű ezeket egyben áttekinteni.

Szalay Krisztina

Tartalomjegyzék

1. Statisztikai alapfogalmak

1.1. Sokaság

A statisztika tárgya véletlen tömegjelenségek megfigyelése, viselkedésének leírása. A **statisztikai sokaság** a megfigyelés eredményeinek összessége. A sokaság lehet diszkrét, illetve folytonos.
Diszkrét sokaság esetén a sokaság konkrét értékeket tartalmaz, ilyen például az egy kereszteződésben áthaladó gépjárművek száma.
Folytonos sokaság értékei valamely intervallum(ok)ban tetszőleges értéket felvehetnek, ilyen például az adott kereszteződésben áthaladó autók sebessége, egy adott munkafolyamat elvégézéséhez szükséges idő stb. Utóbbiakat csak bizonyos mérési pontossággal tudjuk megadni.

Példa A Matematika3 tárgy zh eredményeit vizsgáljuk, ehhez 10 véletlenszerűen kiválasztott hallgató pontszámát tekintjük:

8	7	10	12	4	10	8	8	4	6

(A zh-n elérhető pontszám 0 és 12 közötti egész szám.)
Ebben a példában diszkrét sokaságot látunk, amelyben a pontszámokat csak felsoroltuk, de semmilyen szempont szerint nem rendeztük. Látható azonban, hogy a felsorolásban több azonos pontszám is szerepel. Szemléletesebbé tehetjük az adatainkat, ha az egyes pontszámokat az előfordulásuk száma (**gyakoriságuk**) szerint csoportosítjuk:

Pontszám	4	6	7	8	10	12	
Gyakoriság	2	1	1	3	2	1	Összesen:10

Diszkrét sokaság esetén egy másik lehetőség az adatok rendezésére, hogy úgynevezett **osztályokba soroljuk** őket a megfigyelésünk szempontja(i) szerint. (Fontos, hogy úgy adjuk meg az osztályokat, hogy a sokaság minden egyes elemét egyértelműen be tudjuk sorolni egy osztályba.) Ha például az érdekel bennünket, hogy sikerült-e teljesíteni a hallgatónak az aláírás feltételét (azaz legalább 6 pontot elérni a zh-n), akkor az alábbi csoportosítás célszerű:

Pontszám	0-5	6-12	
Gyakoriság	2	8	Összesen:10

Szokás az eredményt az alábbi alakban is megadni:

Pontszám	0-6	6-13	
Gyakoriság	2	8	Összesen:10

Ebben az esetben a '0-6' osztályba azok a pontszámok tartoznak, amelyekre $0 \leq \text{pontszám} < 6$ teljesül, a '6-13' osztályban pedig azok, amelyekre $6 \leq \text{pontszám} < 13$.

Érdemes megemlíteni azonban, hogy az osztályokba sorolás ebben az esetben információ vesztést eredményez(het) az eredeti adatokhoz képest.

Példa 10 hallgató testmagasságát tartalmazza az alábbi táblázat (centiméterben):

182	177	158	164	167	188	178	169	175	192

Most egy folytonos sokaságot látunk, hiszen a mérési eredmények nyilvánvalóan nem egész értékek, ebben az esetben kerekítéssel kerültek a táblázatba. (A kerekítés szabályai szerint pl. akinek a testmagassága legalább 181,5 cm, de kevesebb, mint 182,5 cm, annak a magassága 182 cm-rel került az adatok közé.) Természetesen most is besorolhatjuk az adatainkat osztályokba. Az egyes osztályokhoz tartozó intervallumok hossza lehet azonos (**egyenközű beosztás**), de akár eltérő hosszúságú intervallumokat is megadhatunk. Ezt azonban formálisan különböző felírással tehetjük meg. Az alábbi táblázatban egy egyenközű beosztást látunk, ugyanazokat az osztályokat három különböző módon megadva.
Jelölés: m jelölje a testmagasságot cm-ben mérve.

I.	$157,5 \leq m < 166,5$	$166,5 \leq m < 175,5$	$175,5 \leq m < 184,5$	$184,5 \leq m < 193,5$
II.	157,5-166,5	166,5-175,5	175,5-184,5	184,5-193,5
III.	158-166	167-175	176-184	185-193

Újra kiemeljük, hogy a három megadási mód ugyanazokat az osztályokat jelöli, más-más módon formalizálva!

Ha meghatároztuk az oszályokat, akkor már csak annyi a dolgunk, hogy összeszámoljuk, a sokaság hány eleme sorolható az adott osztályba. (A besorolást az I. típusú felírás könnyíti meg leginkább, de ezt ritkán használjuk.)

Testmagasság	158-166	167-175	176-184	185-193	Összesen
Gyakoriság	2	3	3	2	10

Ugyanezen sokaságot nem egyenközű osztályokba is sorolhatjuk, pl. az alábbi módon:

Testmagaság (cm)	155-180	181-190	191-200	
Gyakoriság	7	2	1	Összesen:10

Gyakran az adatokat már osztályokba sorolva kapjuk meg, ez különösen akkor praktikus, ha egyébként nagyon nagy elemű a sokaság.

1.2. Hisztogramok

1.2.1. Gyakorisági hisztogram

A sokaságot többek között hisztogramokon jeleníthetjük meg. Ehhez az osztályoknak megfelelő intervallumok fölé a gyakorisággal arányos területű téglalapot rajzolunk. A **hisztogram**ok legfontosabb jellemzője tehát, hogy a **téglalapok területe mindig arányos az adott intervallumhoz tartozó gyakorisággal**. A következő példákon megnézzük, hogy hogyan sorolhatjuk osztályokba az adatainkat, és milyen típusú hisztogramokat rajzolhatunk.

Példa Egy telefonos ügyfélszolgálatnál a bejövő hívások hosszát vizsgálták. Ehhez 20 egymást követő hívás hosszát mérték meg. A sokaság tehát $n = 20$ elemet tartalmaz. A kapott adatok (másodpercben):

48	142	90	78	168	88	55	24	49	71
110	92	73	34	53	65	78	114	88	183

1. Első lépésben növekvő sorrendbe rendezzük az adatainkat (nem feltétlenül szükséges, de megkönnyíti a dolgunkat):

24	34	48	49	53	55	65	71	73	78
78	88	88	90	92	110	114	142	168	183

2. Második lépésben létrehozzuk az **intervallumbeosztás**t. Ehhez a rendezett sokaság elemeit belefoglaljuk egy olyan $[a, b)$ intervallumba, amelyben a értéke legfeljebb akkora, mint a sokaság legkisebb eleme, b értéke pedig nagyobb, mint a sokaság legnagyobb eleme. Majd ezt az $[a, b)$ intervallumot felosztjuk r részintervallumra (osztályra) úgy, hogy az egyes osztályok között nincs átfedés, de együtt lefedik az $[a, b)$ intervallumot.

Egy sokaság esetén a hisztogram alakja nagymértékben függ az egyes osztályok hosszától és a részintervallumok számától. Az osztályok számát a feladatnak megfelelően célszerű meghatározni. Ha n a sokaság elemeinek száma, akkor kisebb sokaság esetén $\sqrt{n}$-hez, nagyobb sokaság esetén $\log_2(n)$-hez közeli értéket javasolt választani az osztályok számának. (Ha túl kevés osztályt választunk, akkor a hisztogram durva, elnagyolt lesz, ha túl sokat, akkor a hisztogram 'ugrálni' fog.) Természetesen ettől el lehet térni.

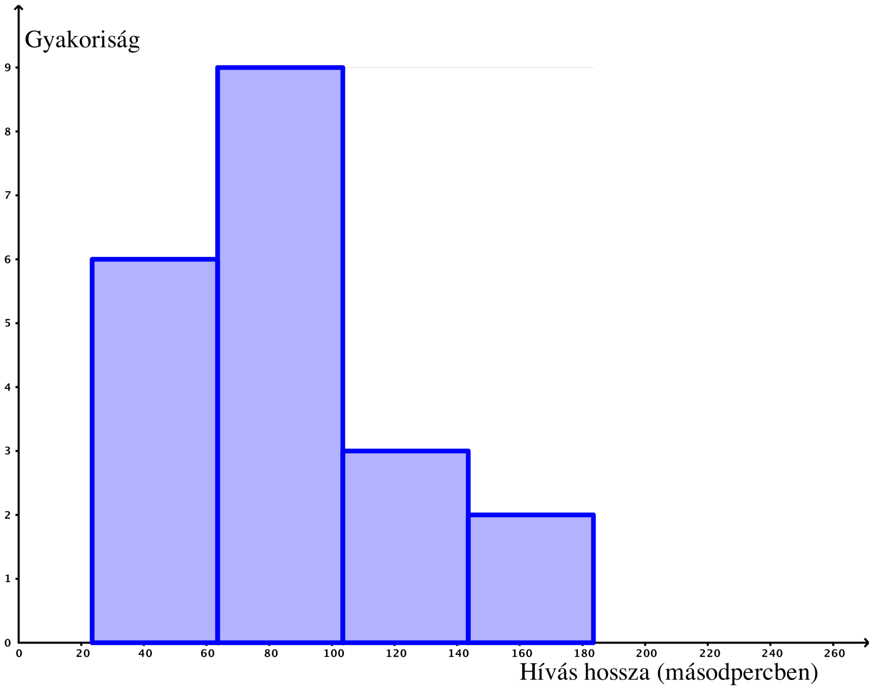
Példánkban $n = 20$, $\sqrt{20} = 4,4721$, ezért 4 vagy 5 osztályt érdemes meghatározni.
Legyen az $[a, b)$ intervallum baloldali határa a sokaság legkisebb eleme $a = 23,5$. Itt figyelembe vesszük, hogy a sokaság folytonos (híváshossz), a táblázat kerekített értékeket tartalmaz.
Jobboldali határnak a legnagyobb elemnél nagyobb számot kell választani, ezért legyen $b = 183,5$.
Az osztályok száma legyen $r = 4$, a részintervallumok hossza pedig azonos, $\dfrac{183,5 - 23,5}{4} = 40$.
Egyenközű beosztásunk van, az osztályhatárok: $23,5$, $63,5$, $103,5$, $143,5$, $183,5$.
Jelölje h a hívás hosszát.
A gyakorisági eloszlás tehát (a hívás hosszát másodpercben mérve):

Hívás hossza	$23,5 \leq h < 63,5$	$63,5 \leq h < 103,5$	$103,5 \leq h < 143,5$	$143,5 \leq h < 183,5$
Gyakoriság	6	9	3	2

Általánosan használt alternatív felírási mód:

Hívás hossza	24-63	64-103	104-143	144-183	
Gyakoriság	6	9	3	2	Összesen:20

3. A harmadik lépésben ábrázoljuk a sokaságot. Erre egy szemléletes lehetőség, hogy a vízszintes tengelyen az osztályhatárokat ábrázoljuk, a függőleges tengelyen pedig (egyenközű beosztás esetén!) a gyakoriságokat.



1. ábra.

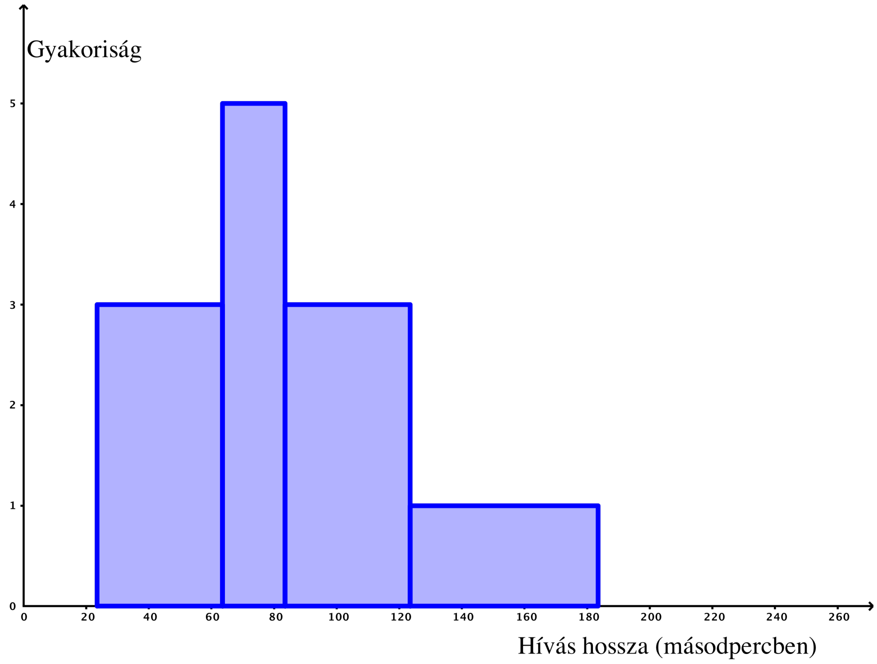
(Ellenőrizzük, hogy az egyes osztályokhoz tartozó téglalapok területe arányos a hozzájuk tartozó gyakorisággal!)
Vegyük észre, hogy ha a hisztogramon kiszámoljuk a téglalapok területének összegét, akkor az $40 \cdot 6 + 40 \cdot 9 + 40 \cdot 3 + 40 \cdot 2 = 800$.

Tekintsük a példánkban megadott sokaságot, de nem egyenközű beosztással:

A hisztogramoknál nagyon fontos, hogy a téglalapok területe arányos legyen a gyakorisággal. Ahhoz, hogy ez teljesüljön, ha nem egyenközű a beosztás, akkor a téglalapok magasságát arányosítani kell az osztályok hosszával.
Az osztályhatárok legyenek: $23,5$, $63,5$, $83,5$, $123,5$, $183,5$. Mivel a $[23,5; 63,5)$ intervallum hossza kétszerese a (referenciának tekintett) $[63,5; 83,5)$ intervallum hosszának, ezért az ábrázolásnál a téglalap magassága az osztályhoz tartozó gyakoriság fele, $\frac{6}{2} = 3$. A $[123,5; 183,5)$ intervallum hossza háromszorosa a $[63,5; 83,5)$ intervallum hosszának, a téglalap magassága $\frac{3}{3} = 1$ A gyakorisági eloszlás ekkor:

Hívás hossza (mp)	24-63	64-83	84-123	124-183	
Gyakoriság	6	5	6	3	Összesen:20
A téglalap magassága	3	5	3	1	

A gyakoriságokat hisztogramon ábrázoljuk:



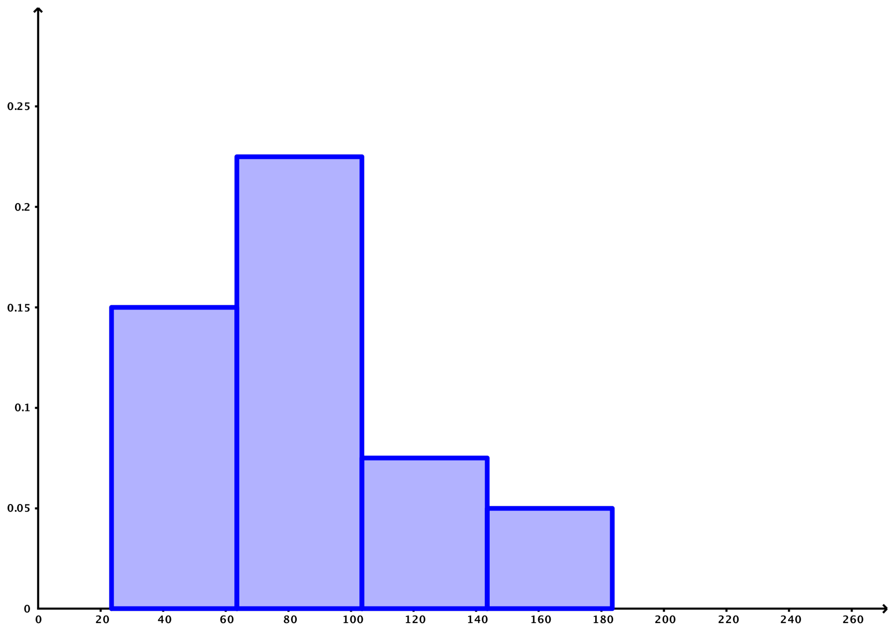
2. ábra.

(A téglalapok összterülete $40 \cdot 3 + 20 \cdot 5 + 40 \cdot 6 + 60 \cdot 1 = 400$.)

Látható, hogy ugyan a hisztogramokon a területek arányosak a hozzájuk tartozó gyakoriságokkal, de a téglalapok összterülete függ az osztálybeosztástól. Hogyan lehet ezt elkerülni?
Ha a hisztogramot úgy ábrázoljuk, hogy a gyakoriságokat elosztjuk a hozzájuk tartozó intervallumok hosszával, akkor **gyakorisági hisztogram**ot kapunk. Ekkor a nem egyenközű beosztás esetén nem kell arányosítani a téglalapok magasságának meghatározásához, elég a gyakoriságokat elosztani az intervallum hosszával.

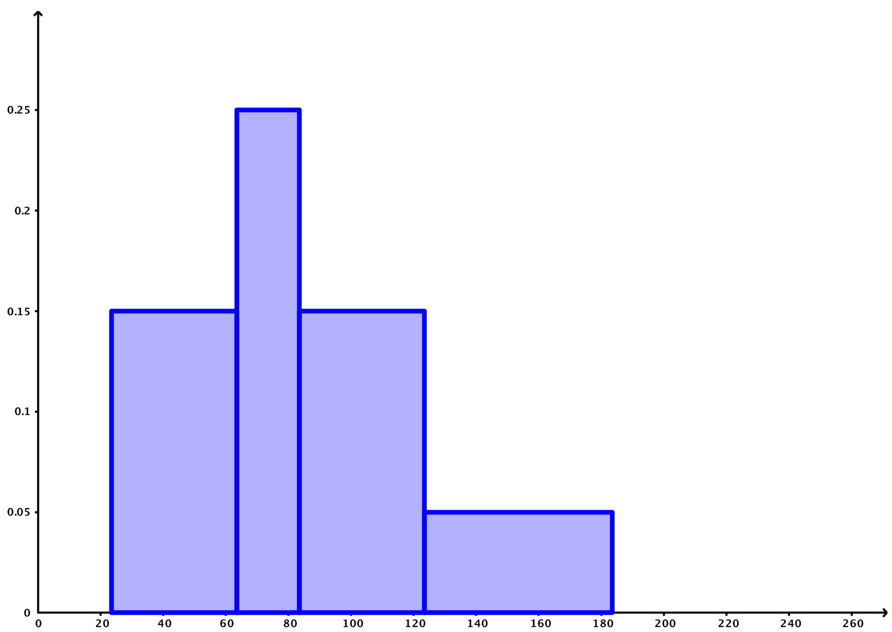
Gyakorisági hisztogram esetén a téglalapok területe továbbra is arányos a megfelelő gyakoriságokkal, de az összterületük megegyezik a minta elemszámával.

A fenti példa egyenközű beosztásánál a gyakorisági hisztogram ábrázolásához nem kell mást tennünk, mint a gyakoriságokat minden osztályban osztani 40-nel. Ez azt jelenti, hogy a hisztogramon mindegyik téglalap magassága arányosan csökken. A következő ábrán látható a gyakorisági hisztogram:



3. ábra. Gyakorisági hisztogram egyenközű beosztáshoz

A fenti példa nem egyenközű beosztásához a gyakorisági hisztogram:



4. ábra. Gyakorisági hisztogram nem egyenközű beosztáshoz

A téglalapok területének összege mindkét esetben a sokaság $n = 20$ elemszámával egyezik meg. Azt is észrevehetjük, hogy a gyakorisági hisztogramok alakja nem tér el lényegesen az eredeti hisztogramokétól, mindössze az y-tengely mértékegysége változott.

1.2.2. Sűrűséghisztogram

A gyakorisági hisztogramon a téglalapok összterülete a sokaság elemeinek számával egyezik meg. Most azt szeretnénk elérni, hogy a téglalapok területének összege független legyen a sokaság elemeinek számától.

A **sűrűséghisztogram** esetében nem a gyakoriságokat, hanem a **relatív gyakoriság**okat, azaz az adott osztályba tartozó elemek sokaságbeli előfordulási arányát osztjuk a megfelelő intervallum hosszával.
Ez egyben azt is jelenti, hogy a sűrűséghisztogramon a téglalapok területének összege szükségszerűen 1. A sűrűséghisztogram ábrázolásában csak y-tengely beosztásában tér el a gyakorisági hisztogramtól, emiatt gyakran előforul, hogy ugyanabban a koordináta-rendszertben ábrázolják őket, jobb oldalon jelölve az eltérő beosztású y-tengelyt.
Ennél többet is mondhatunk: az egyes téglalapok területe a sűrűséghisztogramon megegyezik az adott osztályba tartozó elemek előfordulási arányával (relatív gyakoriságával) a sokaságban.

1. feladat Egy rövidtávú parkolóban egy adott napon feljegyezték az összes ott megálló autó parkolási idejét (percben). Ezt az alábbi táblázat tartalmazza.

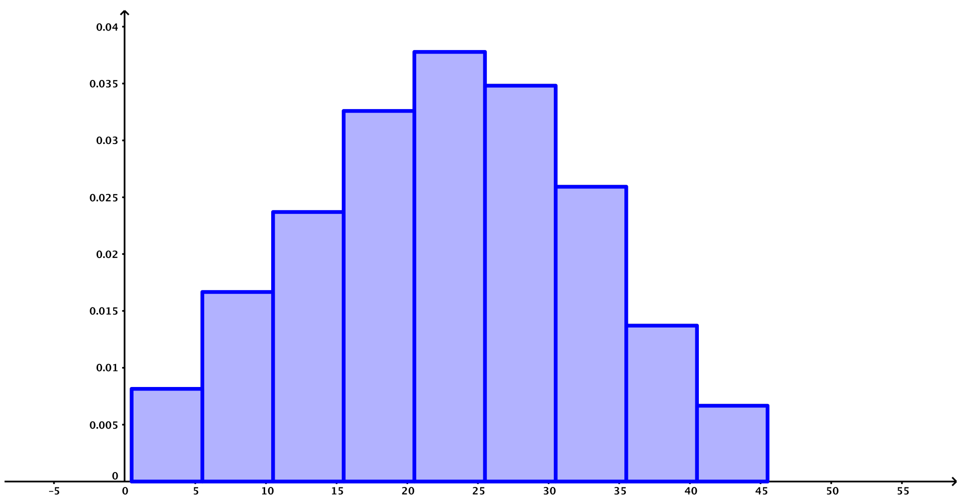
Parkolási idő	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	
Gyakoriság	22	45	64	88	102	94	70	37	18	Össz: 540

Rajzoljuk fel a sűrűséghisztogramot!

Megoldás: Először kiszámoljuk a relatív gyakoriságokat, majd minden relatív gyakoriságot osztunk 540-nel az ábrázoláshoz:

Parkolási idő	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	
Gyakoriság	22	45	64	88	102	94	70	37	18	Össz: 540
Rel. gyakoriság	0,0407	0,0833	0,1185	0,1630	0,1889	0,1741	0,1296	0,0685	0,0333	Össz: 1

A sűrűséghisztogram:



5. ábra.

1.2.3. Gyakorisági és relatív gyakorisági poligon

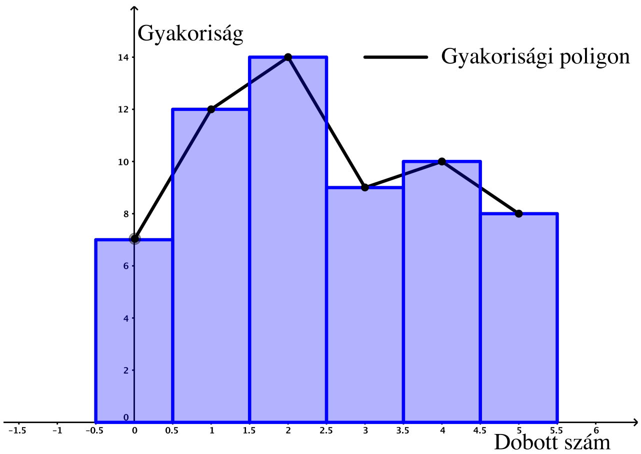
A hisztogramokon az oszlopok felső oldalának felezőpontjait egyenes szakaszokkal összekötve egy **poligont** kapunk. Ha a gyakorisági hisztogramra alkalmazzuk, akkor **gyakorisági poligon**t, ha a sűrűséghisztogramra, akkor **relatív gyakorisági poligon**t kapunk.

2. feladat Egy dobókockát 60-szor feldobva az alábbi eredmények adódtak:

Dobott szám	1	2	3	4	5	6	
Gyakoriság	7	12	14	9	10	8	Összesen:60

Ábrázoljuk a gyakorisági hisztogramot és a gyakorisági poligont!

Megoldás: Az egyes osztályok szélessége most 1, így a gyakorisági hisztogram és a gyakorisági poligon:



6. ábra. Gyakorisági poligon

3. feladat Egy cégnél a dolgozók életkor szerinti eloszlását vizsgálták. A 120 dolgozó életkorából alábbi adatokat kapták:

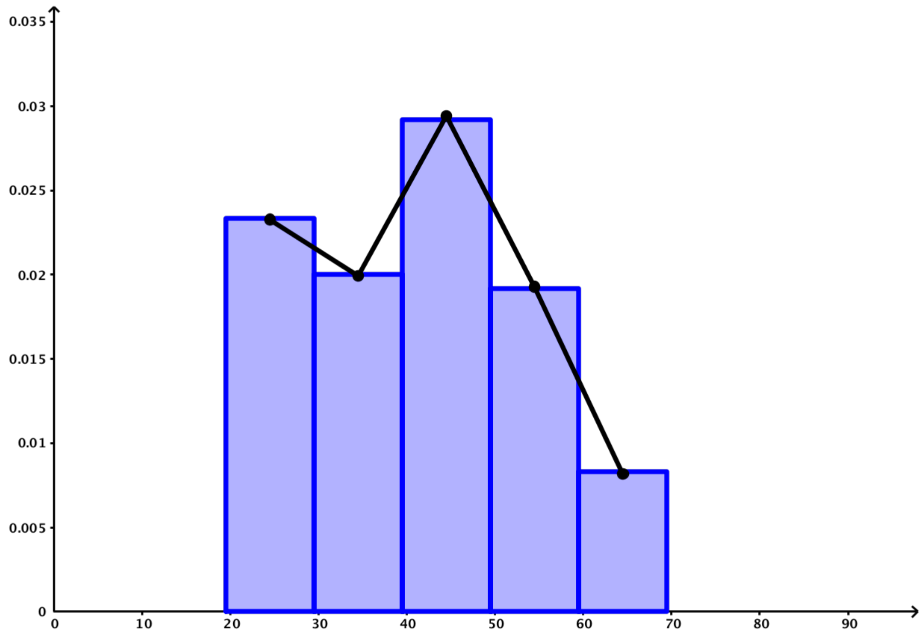
Életkor (év)	20-29	30-39	40-49	50-59	60-69
Dolgozók száma	28	24	35	23	10

Ábrázoljuk a sűrűséghisztogramot és a relatív gyakorisági poligont!

Megoldás: Az $n = 120$ adatot egyenközű osztályokba sorolva adja meg a feladat. Először határozzuk meg a relatív gyakoriságokat az egyes osztályokban. Ehhez annyit kell tennünk, hogy a gyakoriságokat rendre osztjuk $n = 120$-szal. Ezután a sűrűséghisztogramon az oszlopok magasságát megkapjuk, ha a relatív gyakoriságokat elosztjuk az osztály szélességével (ami a példában minden osztály esetén 10).

Életkor (év)	20-29	30-39	40-49	50-59	60-69
Gyakoriság	28	24	35	23	10
Relatív gyakoriság	0,2333	0,2	0,2917	0,1917	0,0833
Oszlop magassága	0,02333	0,02	0,02917	0,01917	0,00833

A sűrűséghisztogram és a relatív gyakorisági poligon:



7. ábra.

Gyakran nem (csak) a gyakoriságra (vagyis arra, hogy egy adat hányszor fordul elő a sokaságban) vagyunk kíváncsiak, hanem arra is, hogy egy adott értéknél kisebb (nagyobb) adat hányszor található meg a sokaságban. Ehhez pedig nem kell más tennünk, mint összeadni az adott értéknél kisebb (nagyobb) adatokhoz tartozó gyakoriságokat. Ezzel el is jutottunk a **kumulált gyakoriság** fogalmához. Diszkrét esetben a kumulált gyakoriság azt adja meg, hogy az adott értéknél hány nem nagyobb (tehát vele megegyező vagy kisebb) érték fordul elő a sokaságban, folytonos esetben pedig azt, hogy az adott osztályköz felső határánál hány nem nagyobb érték fordul elő a sokaságban.
Az 2. feladatban megadott diszkrét sokaság esetén a kumulált gyakoriság:

Dobott szám	Kumulált gyakoriság
≤ 1	7
≤ 2	$7 + 12 = 19$
≤ 3	$7 + 12 + 14 = 33$
≤ 4	$7 + 12 + 14 + 9 = 42$
≤ 5	$7 + 12 + 14 + 9 + 10 = 52$
≤ 6	$7 + 12 + 14 + 9 + 10 + 8 = 60$

Vegyük észre, hogy a kumulált gyakoriság oszlopának utolsó értéke szükségszerűen meg kell egyezzen a dobások számával, mivel minden dobás értéke 6 vagy annál kisebb.

4. feladat Kertészetben nordman fenyők magasságát vizsgálták. Az ültetés utáni 8. évben 40 fenyő csemete magasságát megmérve az alábbi méreteket kapták (cm-ben):

Csemeték magassága	Gyakoriság
121-125	2
126-130	4
131-135	7
136-140	14
141-145	9
146-150	3
151-155	1

A kumulált gyakorisági táblázat felírásához adjuk meg az osztályok felső határait: 125,5, 130,5, 135,5, 140,5, 145,5, 150,5 155,5. Az első osztály alsó határa 120,5. Ennek segítségével a kumulált gyakorisági táblázat:

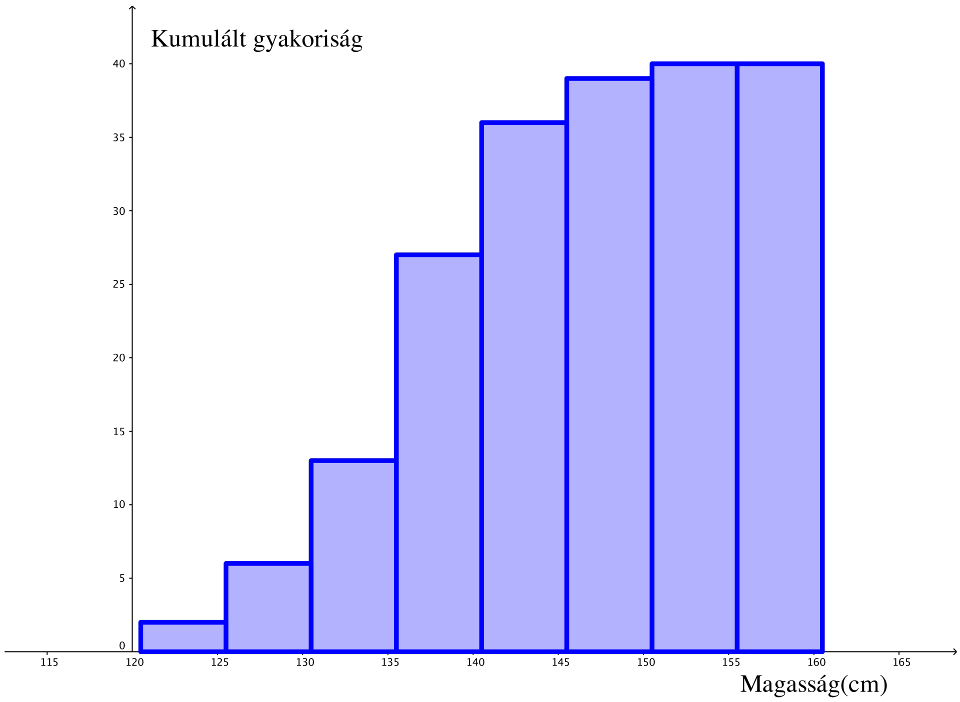
Csemeték magassága	Gyakoriság
$<$ 120,5	0
$<$ 125,5	$0+2$
$<$ 130,5	$0+2+4=6$
$<$ 135,5	$0+2+4+7=13$
$<$ 140,5	$0+2+4+7+14=27$
$<$ 145,5	$0+2+4+7+14+9=36$
$<$ 150,5	$0+2+4+7+14+9+3=39$
$<$ 155,5	$0+2+4+7+14+9+3+1=40$

Hagyjuk el a táblázatból a számításokat, így a kumulált gyakorisági táblázat:

Csemeték magassága	Gyakoriság
$<$ 120,5	0
$<$ 125,5	2
$<$ 130,5	6
$<$ 135,5	13
$<$ 140,5	27
$<$ 145,5	36
$<$ 150,5	39
$<$ 155,5	40

A táblázatból leolvashatjuk például, hogy 140,5 cm-nél 27 fenyő magassága kisebb, legalább 145,5 cm magas (azaz nem alacsonyabb, mint 145,5 cm) $40-36=4$ csemete, sőt, azt is meg tudjuk adni, hogy pl. a legalább 130.5, de kevesebb, mint 145,5 cm magas facsemeték száma $36-6=30$. (Ez utóbbihoz a 145,5 cm-nél alacsonyabb fák számából kivontuk a 130,5 cm-nél alacsonyabb fák számát.)

A kumulált gyakoriságot oszlopdiagramon ábrázolhatjuk:



8. ábra.

Ha az oszlopok tetejének felezőpontjait összekötjük, akkor egy poligont kapunk (szokás *ogivának* is nevezni). Ha a felosztást sűrítjük, akkor egy *kumulált gyakorisági görbét* kapunk.

A kumulált gyakorisági táblázat segítségével olyan gyakoriságokat is egyszerűen tudunk becsülni(!), amelyek a táblázatból nehézkesen olvashatók le.

1.3. Mérőszámok

A korábbiakban megnéztük, hogyan tudjuk ábrázolni a sokaságot. Most nézzük, milyen mérőszámokkal jellemezhetjük.

Tekintsünk egy n elemű sokaságot, melynek elemeit jelölje rendre $x_1, x_2, \ldots, x_n$. Láttuk, hogy hasznos lehet, ha a sokaság elemeit nagyság szerint növekvő (nem feltétlenül *szigorúan* növekvő!) sorrendbe rendezzük. A sokaság legkisebb elemét jelölje x_1^*, a nagyság szerint utána következőt x_2^* és így tovább. A rendezett sokaság tehát az $x_1^*, x_2^*, \ldots, x_n^*$ elemekből áll.

a) A **medián** a rendezett sokaság 'középső eleme'. Jelölés: *med*. Ez a gyakorlatban azt jelenti, hogy ha a sokaság elemszáma páratlan, $n = 2k + 1$, akkor a rendezés után a $(k + 1)$-edik elem a medián, azaz x_{k+1}^*. Ha a sokaság elemszáma páros, $n = 2k$, akkor a medián a k-adik és a $(k+1)$-edik elemek számtani közepe, $\frac{x_k^* + x_{k+1}^*}{2}$. Azaz:

$$med = \begin{cases} \dfrac{x_k^* + x_{k+1}^*}{2} & \text{ha } n = 2k \\[2ex] x_{k+1}^* & \text{ha } n = 2k+1 \end{cases}$$

A mediánnál kisebb és nagyobb elemek száma azonos a sokaságban. Előnye, hogy a szélsőséges adatok nem befolyásolják az értékét.

b) A **módusz** a legtöbbször előforduló elem a sokaságban. Akár több is lehet belőle, de ekkor nem túl hasznos mérőszám. Könnyen meghatározható, előnye, hogy nem veszi figyelembe a szélsőséges értékeket.
A gyakorisági hisztogramról egyszerűen leolvasható az értéke: a legnagyobb gyakoriságú osztály osztályközepével (az intervallum felezőpontjával) egyezik meg.

c) A **számtani átlag** (továbbiakban **átlag**) kiszámításához a sokaság elemeinek összegét elosztjuk a sokaság elemszámával. Kiszámítása:

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Ha gyakorisági eloszlás van megadva, akkor minden osztály osztályközepe felel meg az osztályba sorolt értékeknek.
A medián, a módusz és az átlag a sokaság középértékének jellemzésére szolgáló mérőszámok.

A sokaságnak a középértéke mellett a szóródás is fontos jellemzője. Ezt többek között a terjedelemmel és a szórással jellemezhetjük:

d) A minta **terjedelme** nem más, mint a legnagyobb és legkisebb elem különbsége: $x_n^* - x_1^*$. Könnyen számolható, ugyanakkor egyetlen szélsőséges adat is jelentősen befolyásolja az értékét.

e) A **szórás** a sokaság elemeinek az átlagtól való eltérésnégyzeteinek összegének n-edrészének gyöke. A formula felírása egyszerűbb, ha a **szórásnégyzet**et számoljuk ki, ami:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n}$$

Megjegyezzük, hogy a számítás kerekítésekből adódó pontatlanságának javításához a szórásnégyzet kiszámítására az alábbi formulát használhatjuk:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i \right)^2$$

5. feladat Adott az alábbi sokaság:

1,98	1,96	1,99	2,00	2,01	1,95	1,97	1,96	1,97	2,02

a) Határozzuk meg a sokaság mediánját!

b) Határozzuk meg a sokaság moduszát!

c) Határozzuk meg a sokaság átlagát!

d) Határozzuk meg a sokaság szórását!

Megoldás
A sokaság n=10 elemű. Mielőtt elkezdjük a számításokat, rendezzük a sokaság elemeit.

1,95	1,96	1,96	1,97	1,97	1,98	1,99	2,00	2,01	2,02

a) A medián kiszámításához a rendezett sokaságot tekintsük. Mivel n=10, azaz páros elemünk van, a két középső elem, $x_5^* = 1,97$ és $x_6^* = 1,98$. A medián a számtani közepük, azaz $\frac{x_5^* + x_6^*}{2} = \frac{1,97 + 1,98}{2} = 1,975$.

b) Két olyan érték van a sokaságban, amely kétszer fordul elő, így két módusz van, az $1,96$ és az $1,97$.

c) Az átlag kiszámításához nincs szükség a rendezésre, mindössze a korábban megadott formulába kell helyettesítenünk.

$$\overline{x} = \frac{1}{10} \cdot \sum_{i=1} 10x_i = \frac{x_1 + x_2 + \ldots + x_{10}}{10} = \frac{1,98 + 1,96 + \ldots + 2,02}{10} = \frac{19,81}{8} = 1,981$$

d) A szórás kiszámításához felhasználjuk az átlag értékét is.

$$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{10} (x_i - \overline{x})^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_{10} - \overline{x})^2}{n}} =$$

$$= \sqrt{\frac{(1,98 - 1,981)^2 + (1,96 - 1,981)^2 + \ldots + (2,02 - 1,981)^2}{10}} = \sqrt{\frac{0,00489}{10}} = 0,0221$$

6. feladat Egy egyetemen a diplomaosztó után megkérdeztek 50 friss diplomást arról, hogy hány nyelvet beszélnek (legalább társalgási szinten). Az eredményeket az alábbi táblázat tartalmazza.

Beszélt nyelvek száma	1	2	3	4	5	
Gyakoriság	26	17	4	2	1	Összesen:50

a) Határozzuk meg a sokaság mediánját!

b) Határozzuk meg a sokaság moduszát!

c) Határozzuk meg a sokaság átlagát!

d) Határozzuk meg a sokaság szórását!

Megoldás

a) Mivel összesen 50 adatunk van, a 25. és 26. elem átlaga a medián. A táblázatból látszik, hogy az első 26 érték azonos, így a 25. és 26. értéke is 1. Az átlaguk, ezzel a medián $\frac{1+1}{2} = 1$.

b) A módusz leolvasása nem okozhat nehézséget, a leggyakorabban (26-szor) előforduló érték az 1.

c) Az n=50 elemű sokaságban sok ismétlődő elem van, ezért az átlag és a szórás kiszámításánál kihasználhatjuk, hogy az összegben bizonyos tagok többször is előfordulnak. Ezeket figyelembe véve az alábbiak szerint rövidíthetünk a számításokon. Az átlag:

$$\overline{x} = \frac{26 \cdot 1 + 17 \cdot 2 + 4 \cdot 3 + 2 \cdot 4 + 1 \cdot 5}{50} = \frac{85}{50} = 1,7$$

d) A szórás:

$$\sqrt{\frac{26 \cdot (1-1,7)^2 + 17 \cdot (2-1,7)^2 + 4 \cdot (3-1,7)^2 + 2 \cdot (4-1,7)^2 + 1 \cdot (5-1,7)^2}{50}} = \sqrt{\frac{42,5}{50}} = 0,9220$$

7. feladat Egy légitársaság két város közötti napi járatának repülési idejét mérte meg 30 egymást követő napon. Az eredményeket percben, a kerekítési szabályoknak megfelelően az alábbi táblázat tartalmazza.

Repülési idő (perc)	101-110	111-120	121-130	131-140	141-150	150-160
Járatok száma	2	3	8	14	2	1

a) Határozzuk meg a sokaság átlagát!

b) Határozzuk meg a sokaság szórását!

Megoldás Ebben a feladatban az adatokat már 6 osztályba sorolva adták meg. Ez azt jelenti számunkra, hogy nem tudjuk az $n = 30$ mérés értékét felsorolni, mindössze következtetni tudunk azokra (becsülni tudjuk az értéküket). Hogyan becsüljünk? Legegyszerűbben az egyes osztályok középpontjával (osztályközepével) becsülhetünk az alábbiak szerint:

Osztály	Osztályközép:	Gyakoriság:k_i	Relatív gyakoriság:$\frac{k_i}{n}$
101-110	105,5	2	$\frac{2}{30}$
111-120	115,5	3	$\frac{3}{30}$
121-130	125,5	8	$\frac{8}{30}$
131-140	135,5	14	$\frac{14}{30}$
141-150	145,5	2	$\frac{2}{30}$
151-160	155,5	1	$\frac{1}{30}$
		$\sum_{i=1}^{6} k_i = 30$	$\sum_{i=1}^{6} \frac{k_i}{n} = 1$

Ez azt jelenti, hogy azoknak a gépeknek a menetidejét, amelyeknél a valós menetidő a $[101, 110)$ intervallumba esik (tehát legalább $100,5$ perc, de kevesebb, mint $110,5$ perc), egységesen $105,5$ percnek tekintjük, mintha mindkét repülőgép 105,5 perc alatt tette volna meg az utat. (Természetesen ezzel az eredeti sokaság elemeit csak becsüljük!)

a) Az átlag:

$$\overline{x} = \frac{2 \cdot 105,5 + 3 \cdot 115,5 + 8 \cdot 125,5 + 14 \cdot +14 \cdot 135,5 + 2 \cdot 145,5 + 1 \cdot 155,5}{30} = \frac{3905}{30} = 130,1667$$

b) A szórás:

$$\sqrt{\frac{2 \cdot (105,5 - 130,1667)^2 + 3 \cdot (115,5 - 130,1667)^2 + \ldots + 1 \cdot (155,5 - 130,1667)^2}{30}} =$$

$$= \sqrt{\frac{3546,6667}{30}} = 10,8730$$

2. A valószínűség fogalma, statisztikai háttere, a valószínűségi változó fogalma

2.1. A valószínűség fogalma, statisztikai háttere

2.1.1. A valószínűség fogalma klasszikus esetben

Dobjunk fel egy szabályos pénzérmét. Majd ismételjük meg ezt a kísérletet egymás után többször. Minden dobás után jegyezzük fel a dobás értékét (F:fej, I:írás), illetve azt, hogy a kísérlet n-edik ismétléséig hány fejet dobtunk összesen. Ezt az értéket, tehát azt, hogy az n kísérletből hány fejet dobtunk (másképpen a fejdobások **gyakoriságát**) jelölje k. Ekkor a k/n értéket a fejdobások **relatív gyakoriságá**nak nevezzük.

	n. dobás értéke	dobott fejek száma (k)	relatív gyakoriság ($\frac{k}{n}$)
$n = 1$	I	0	0
$n = 2$	F	1	0,5
$n = 3$	I	1	0,3333
$n = 4$	F	2	0,5
$n = 5$	F	3	0,6
$n = 6$	F	4	0,6666
$n = 7$	I	4	0,5417
$n = 8$	I	4	0,5
$n = 9$	I	4	0,4444
$n = 10$	I	4	0,4
$n = 11$	F	5	0,4545
$n = 12$	F	6	0,5
$n = 13$	I	6	0,4615
$n = 14$	F	7	0,5
$n = 15$	F	8	0,5333

Az eredményeket megnézve azt látjuk, hogy a kísérletek számát, azaz n-et növelve a relatív gyakoriság (azaz a dobott fejek arányának) értékének ingadozása csökken.

Egy kísérletnek (véletlen tömegjelenségnek) számos kimenetele lehet.
A kísérlet egyes kimeneteleit **elemi esemény**eknek, az összes elemi esemény halmazát pedig **eseménytér**nek (jele: Ω) nevezzük. Az eseménytér bármely részhalmazát **esemény**nek nevezzük, jele általában valamely nagybetű (A,B,...). Az alábbi példákon nézzük meg, mit jelentenek ezek a fogalmak:

- Egy szabályos pénzérmét egyszer feldobva kétféle eredményt kaphatunk, a dobás lehet fej(F) vagy írás(I). Az eseménytér tehát: $\Omega = \{F, I\}$.
- Ha két szabályos pénzérmét dobunk fel, akkor a kísérletnek már négyféle kimenetele lehet, az eseménytér ekkor: $\Omega = \{FF, FI, IF, II\}$. Mindkét példánkban véges halmaz az eseménytér. Az Ω eseménytérből kiválaszthatunk tetszés szerint részhalmazokat: pl. az $\{FI, IF\}$ részhalmaz által meghatározott A esemény az jelenti, hogy 'pontosan egy fejet dobunk'. Megfordítva: jelölje B a 'legalább egy fej van a dobások között' eseményt, ekkor $B = \{FF, FI, IF\}$, ami egy 3 elemi eseményt tartalmazó részhalmazt jelent. Ha a 3 elemi esemény bármelyike bekövetkezik a kísérlet során, akkor azt mondjuk, hogy a B esemény bekövetkezik, B bekövetkezésekor a kedvező kimenetelek száma 3.
- Ha addig dobunk egy szabályos dobókockával, míg fejet nem dobunk, akkor a lehetséges kimenetelek száma végtelen, az eseménytér: $\Omega = \{F, IF, IIF, IIIF,...\}$ már (megszámlálhatóan) végtelen halmaz.

A valószínűség fogalma a klasszikus megközelítés alapján a következő: Ha az Ω eseményteret véges sok (n), azonos valószínűséggel bekövetkező elemi esemény alkotja, és az A esemény bekövetkezésekor a kedvező kimenetelek száma k, akkor az A **esemény** $P(A)$-val jelölt **valószínűsége**:

$$P(A) = \frac{k}{n}$$

A fenti példánkban az A esemény valószínűsége: $P(A) = \dfrac{2}{4} = 0,5$, a B esemény valószínűsége pedig: $P(B) = \dfrac{3}{4} = 0,75$.

Megjegyzés: Természetesen az, hogy az eseménytér véges sok és azonos valószínűségű elemi eseményből áll, jelentősen leszűkíti a klasszikus kiszámítási mód alkalmazhatóságát, ugyanakkor a módszer más feladatokra is általánosítható.

Az n elemű Ω eseménytérben tekintsük a k-féleképpen bekövetkező A eseményt. A $P(A)$ valószínűségről az alábbiakat mondhatjuk:

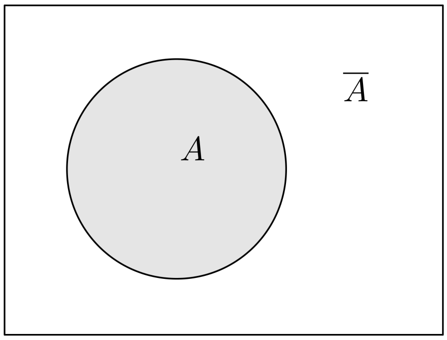
- Véges halmaz esetén a részhalmaz elemszáma legalább 0 és legfeljebb az eredeti halmaz elemszáma, ezért
$$0 \leq k \leq n$$
Ebből adódik, hogy
$$0 \leq \frac{k}{n} \leq 1$$
Tehát
$$0 \leq P(A) \leq 1$$
Azaz a valószínűség 0 és 1 közötti szám.

  Ha egy esemény nem következhet be egy kísérlet során, akkor **lehetetlen esemény**nek, ha biztosan bekövetkezik, akkor **biztos esemény**nek nevezzük. Klasszikus esetben teljesül, hogy ha $P(A) = 0$, akkor A lehetetlen esemény, ha $P(A) = 1$, akkor A biztos esemény.

  (*Megjegyzés:* Később látni fogjuk, hogy ha az eseménytér nem véges halmaz, akkor ez nem feltétlenül igaz.)

- Jelölje $\overline{A}$ (A **ellentett**) azt az eseményt, amikor A nem következik be. Az $\overline{A}$ esemény ekkor $(n-k)$-féleképpen következhet be, valószínűsége tehát:
$$P(\overline{A}) = \frac{n-k}{n} = 1 - \frac{k}{n} = 1 - P(A)$$
Rendezés után
$$P(A) + P(\overline{A}) = 1$$
Azaz esemény és ellentettjének valószínűségének összege 1.

  (*Ötlet:* Gyakran egy esemény valószínűsége helyett egyszerűbb első lépésben az ellentett esemény valószínűségét kiszámolni. Különösen hasznos lehet az ellentett eseménnyel számolni abban az esetben, ha 'legalább', 'legfeljebb' típusú kérdéseink vannak.)



9. ábra. Esemény és ellentettje

1. feladat Két szabályos dobókockát feldobuk.

a) Mi a valószínűsége annak, hogy pontosan egy hatost dobunk?
b) Mi a valószínűsége annak, hogy legfeljebb egy hatost dobunk?

Megoldás: A lehetséges kimenetelek:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	**(1,6)**
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	**(2,6)**
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	**(3,6)**
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	**(4,6)**
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	**(5,6)**
(6,1)	**(6,2)**	**(6,3)**	**(6,4)**	**(6,5)**	(6,6)

$\Omega = \big\{(1,1),(1,2),(1,3),\ldots,(6,5),(6,6)\big\}$, az eseménytér tehát olyan rendezett számpárokból áll, ahol mindkét komponens 1 és 6 közötti egész szám. Összesen $6 \cdot 6 = 36$ elemű az eseménytér, melyben minden elemi esemény azonos valószínűséggel következik be (szabályos mindkét dobókocka).

a) Jelölje A azt az eseményt, hogy 'pontosan egy hatost dobunk'. A táblázatban vastag vonallal emeltük ki, hogy mely elemi események jelentik az A bekövetkezését. Ez egy 10 elemű részhalmaz, vagyis
$$P(A) = \frac{10}{36}$$

b) Jelölje B azt az eseményt, hogy 'legfeljebb egy hatost (azaz 0 vagy 1 hatost) dobunk' . A táblázatból könnyen leolvashatjuk, hogy 35 ennek megfelelő elemi esemény van. Tehát
$$P(B) = \frac{35}{36}$$

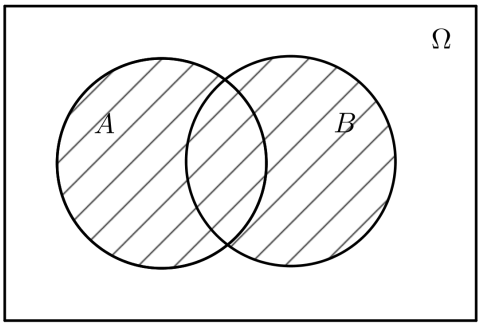
Vegyük azonban észre, hogy ha nem írjuk fel a táblázatot, akkor egyszerűbben számolhatunk az ellentett segítségével. $\overline{B}$-en azt az eseményt értjük, hogy 'nem igaz, hogy legfeljebb egy hatost dobunk', azaz 'több, mint egy hatost dobunk'. Ez a kísérletünkben azzal egyenlő, hogy 'pontosan két hatost dobunk', ez pedig táblázat nélkül is egyszerű, a $(6,6)$ elemi eseményt jelenti. Felhasználva az ellentett esemény valószínűségére vonatkozó szabályunkat:

$$P(B) = 1 - P(\overline{B}) = 1 - \frac{1}{36} = \frac{35}{36}$$

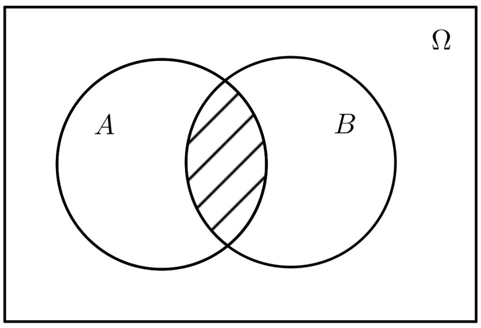
- Legyen A és B két olyan esemény az Ω eseménytérben, amelyeknek nem nulla a valószínűsége. Ekkor

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

ahol $A \cup B$, két esemény **összege** azt az eseményt jelenti, amikor 'A **vagy** B' bekövetkezik, $A \cap B$, két esemény **szorzata** pedig azt az eseményt jelenti, amikor 'A **és** B', tehát mindkettő bekövetkezik. Fontos megjegyezni, hogy ez nem kizáró 'vagy', tehát 'A vagy B' azt jelenti, hogy A vagy B vagy mindkét esemény bekövetkezik. (Egyszerűbben megfogalmazva: legalább az egyik esemény bekövetkezik.)



10. ábra. A és B esemény összege $(A \cup B)$



11. ábra. A és B esemény szorzata $(A \cap B)$

2. feladat Egy osztályban a 30 tanulóból 14 tanuló felvételizik. A felvételizők közül 10-nek van középfokú nyelvvizsgája, a többiek közül 8-nak. Véletlenszerűen kiválasztva egy tanulót, mi a valószínűsége annak, hogy felvételizik vagy van középfokú nyelvvizsgája?

Megoldás: Jelölje A azt az eseményt, hogy 'a kiválasztott felvételizik', B pedig azt az eseményt, hogy 'a kiválasztottnak van középfokú nyelvvizsgája'. Tudjuk, hogy $P(A) = \frac{14}{30}$, $P(B) = \frac{18}{30}$, $P(A \cap B) = \frac{10}{30}$. Így

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{14}{30} + \frac{18}{30} - \frac{10}{30} = \frac{22}{30} = 0,733\dot{3}$$

Tehát $0,733\dot{3}$ annak a valószínűsége, hogy a kiválasztott felvételizik vagy van középfokú nyelvvizsgája.

3. feladat Egy pakli magyar kártyából kihúzunk egy lapot.

a) Mi a valószínűsége annak, hogy a lap piros vagy ász?

b) Mi a valószínűsége annak, hogy a lap piros vagy zöld?

Megoldás: Az eseménytér 32 elemű, mivel a pakliban 32 különböző lap van. Legyen A 'a kihúzott lap piros', B 'a kihúzott lap ász', C 'a kihúzott lap zöld' esemény. A pakliban 8 piros lap, 4 ász és 8 zöld lap van.

a) Az $A \cap B$ esemény azt jelenti, hogy 'a kihúzott lap piros is és ász is', ilyen lap egy van a pakliban, a piros ász. Így

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{8}{32} + \frac{4}{32} - \frac{1}{32} = \frac{11}{32} = 0,3438$$

b) Az $A \cap C$ esemény azt jelenti, hogy 'a kihúzott lap piros is és zöld is', ilyen lap viszont nincs a pakliban. Így

$$P(A \cup B) = P(A) + P(C) - P(A \cap C) = \frac{8}{32} + \frac{8}{32} - \frac{0}{32} = \frac{16}{32} = 0,5$$

A példában azt láttuk, hogy az A és C események nem következhetnek be együtt a kísérlet során, így a számításunk egyszerűsödött.
Ha két esemény nem következhet be együtt, azt mondjuk, hogy a két esemény **kizáró**, azaz $A \cap B = \emptyset$.

Ha az A és B események kizárják egymást, akkor

$$P(A \cup B) = P(A) + P(B)$$

és

$$P(A \cap B) = \emptyset$$

A fenti összefüggést több eseményre is általánosíthatjuk: legyenek $A_1, A_2, \ldots, A_n$ olyan események, amelyek közül bármely kettő kizárja egymást (**páronként kizáróak**). Ekkor

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

2.1.2. Feltételes valószínűség

Definíció: Legyenek az Ω eseménytérben A és B olyan események, melyeknek nem nulla a valószínűsége (azaz $P(A) \neq 0$ és $P(B) \neq 0$). Ekkor annak a valószínűsége, hogy **A** bekövetkezik, **feltéve** hogy **B** már bekövetkezett (jelölés: $P(A \mid B)$, olvasva: A feltéve B valószínűsége):

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

Ebből

$$P(A \cap B) = P(A \mid B) \cdot P(B) \quad \left[= P(B \mid A) \cdot P(A)\right]$$

Megjegyzés: Könnyen belátható, hogy ha A és B kizáró események, akkor $P(A \mid B) = 0$.

4. feladat Két szabályos dobókockát feldobunk.

a) Mi a valószínűsége annak, hogy az első dobás hatos, feltéve, hogy a dobások összege 11?

b) Mi a valószínűsége annak, hogy a dobások összege 11, feltéve, hogy az első dobás hatos?

Megoldás: Az eseménytérben 36 elemi esemény van. Legyen az A esemény, hogy 'az első dobás hatos', a B esemény pedig, hogy 'a dobások összege 11'.

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	**(5,6)**
(6,1)	(6,2)	(6,3)	(6,4)	**(6,5)**	(6,6)

A táblázat utolsó sorában felírt elemi események az A esemény bekövetkezését, a vastag vonallal kiemelt elemi események pedig a B esemény bekövetkezését jelentik.

$$P(A) = \frac{6}{36}$$

$$P(B) = \frac{2}{36}$$

$$P(A \cap B) = \frac{1}{36}$$

a)

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

b)

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

Megjegyzés: Az eredményeket egyszerűbben is megkaphatjuk. A 36 elemű eseményteret szűkítsük le

a feltétel eseményre. A b) esetben ez azt jelenti, hogy a leszűkített eseménytér a $\{(6,1),(6,2),(6,3),(6,4),(6,5),$ elemi eseményből áll (mivel tudjuk, hogy az első dobás hatos, fölösleges a többi elemi eseményt figyelembe venni, hiszen ők nem következtek be a kísérletben). Így az összes lehetőség 6, amelyből mindössze egy felel meg a B esemény bekövetkezésének, a $(6,5)$. Így a keresett valószínűség $\frac{1}{6}$.
Ugyanezen gondolatmenettel az a) esetben a leszűkített eseménytér az $\{(5,6),(6,5)\}$ elemi eseményekből áll, 2 elemű, melyből pontosan az egyik esetben lesz az első dobás hatos (teljesül az A esemény). A keresett valószínűség ebben az esetben $\frac{1}{2}$.

Legyenek A és B tetszőleges események, és tekintsük az $A \cap B$ és az $A \cap \overline{B}$ eseményeket. Ezekre teljesül, hogy $(A \cap B) \cap (A \cap \overline{B}) = \emptyset$ (azaz kizáróak), és $(A \cap B) \cup (A \cap \overline{B}) = A$. Kihasználva, hogy a két esemény kizáró,

$$P(A) = P\big[(A \cap B) \cup (A \cap \overline{B})\big] = P(A \cap B) + P(A \cap \overline{B}) = P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B})$$

A kapott összefüggés tehát:

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

5. feladat Egy dobozban 30 alkatrész van, melyekből 8 hibás. Kihúzunk egy alkatrészt, amelyet félreteszünk, majd újabbat húzunk. Mi a valószínűsége annak, hogy a második alkatrész hibás?

Megoldás: Jelölje A a kérdésben szereplő 'második alkatrész hibás' eseményt, B az 'első alkatrész hibás' eseményt (amelyről nem tudjuk, hogy bekövetkezett-e). Ekkor $\overline{B}$ 'az első alkatrész nem hibás' eseményt jelenti. Amit tudunk a feladatból: $P(B) = \frac{8}{30}$, $P(\overline{B}) = \frac{22}{30}$, illetve $P(A \mid B) = \frac{7}{29}$ és $P(A \mid \overline{B}) = \frac{8}{29}$. (Emlékeztetőül: $A \mid B$ azt jelenti, hogy 'ha az első alkatrész hibás, akkor a második is hibás'. Ebben az esetben az első húzás után félretettünk egy hibás alkatrészt, tehát a második húzáskor összesen 29 alkatrész van a dobozban, közülük már csak 7 hibás.) Keressük $P(A)$-t:

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B}) = \frac{7}{29} \cdot \frac{8}{30} + \frac{8}{29} \cdot \frac{22}{30} = 0,266\dot{6}$$

2.1.3. Függetlenség

Ha a B esemény bekövetkezése vagy be nem következése semmilyen hatással nincs az A bekövetkezésének valószínűségére, akkor azt mondjuk, hogy A és B események **függetlenek**. A feltételes valószínűséget felírva a két független eseményre, ez azt jelenti, hogy $P(A \mid B) = P(A \mid \overline{B}) = P(A)$. Ez megfordítva is igaz, $P(B \mid A) = P(B)$.

Definíció: Az A és B események függetlenek, ha

$$P(A \cap B) = P(A) \cdot P(B)$$

Felhívjuk a figyelmet arra, hogy ha két esemény független, az nem jelenti azt, hogy kizáróak is!

Megjegyzés: Megmutatható, hogy ha A és B események függetlenek, akkor $\overline{A}$ és B események, A és $\overline{B}$ események és $\overline{A}$ és $\overline{B}$ események is függetlenek.
Az összefüggés több eseményre általánosítva: az $A_1, A_2, \ldots, A_n$ események **függetlenek**, ha

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$$

6. feladat Tudjuk, hogy annak a valószínűsége, hogy egy irodában valamely napon meghibásodik a fénymásológép, 0,02. (A fénymásolót csak másnapra javítják meg.)

a) Mi a valószínűsége annak, hogy két egymást követő napon meghibásodik a fénymásoló?

b) Mi a valószínűsége annak, hogy két egymást követő napból pontosan egyszer hibásodik meg a fénymásoló?

Megoldás: Jelölje A esemény azt, hogy 'az első napon meghibásodik a fénymásoló', B pedig azt, hogy 'a második napon meghibásodik'. Feltételezzük, hogy az egyes napokon a meghibásodások egymástól függetlenül következnek be, így a két esemény független.

a) A két egymást követő napon történő meghibásodás azt jelenti, hogy a A esemény és a B esemény is bekövetkezik (azaz $A \cap B$ esemény bekövetkezik). Mivel a két esemény független, ennek valószínűsége:
$$P(A \cap B) = P(A) \cdot P(B) = 0,02 \cdot 0,02 = 0,0004$$

b) A két napból most csak az egyik napon történik meghibásodás. Vagy az első napon történik és a másodikon nem: $A \cap \overline{B}$ következik be, vagy az első napon nem történik meghibásodás de a másodikon igen: $\overline{A} \cap B$ következik be. Ezek egymást kizárják, így
$$P\big((A \cap \overline{B}) \cup (\overline{A} \cap B)\big) = P(A \cap \overline{B}) + P(\overline{A} \cap B) = P(A) \cdot P(\overline{B})) + P(\overline{A}) \cdot P(B) =$$
$$= P(A) \cdot \big[1 - P(B)\big] + \big[1 - P(A)\big] \cdot P(B) = 0,02 \cdot 0,98 + 0,98 \cdot 0,02 = 0,0392$$

Kihasználtuk, hogy az A és $\overline{B}$ események és az $\overline{A}$ és B események is függetlenek.

2.1.4. Néhány hasznos számítási módszer

Az 1. és 3. feladatban láttuk, hogy az egyes eseményeknek megfeleltetett (véges) halmazok elemszámát táblázat segítségével határoztuk meg, ehhez először felsoroltuk az összes elemi eseményt. Könnyen belátható, hogy nagyobb eseményterek esetében ez nem járható út. Segítségül hívhatjuk azonban a középiskolában tanult kombinatorikai módszereket, melyek közül azokat tekintjük át, amelyeket a későbbiekben hasznosak lehetnek.

- Az **ismétlés nélküli permutáció** az mondja meg, hogy n különböző elemet hányféleképpen rendezhetünk sorba úgy, hogy minden elemet pontosan egyszer használunk fel. A lehetőségek száma $n!$ ('n faktoriális').

  Megjegyzés: $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$, és $0! = 1$.

  7. feladat Tekintsük az 1,2,3,4,5,6 számjegyeket. Hány különböző 6-jegyű számot hozhatunk létre belőlük úgy, hogy minden számjegyet pontosan egyszer használunk fel?

  Megoldás: Az első számjegy kiválasztására 6 lehetőségünk van (bármelyiket választhatjuk). Ezután a második számjegyet már csak 5-ből választhatjuk (az első számjegy már nincs a lehetőségek között), a harmadikat már csak négyből és így tovább. Az összes lehetőség tehát a hat számjegy különböző módon való sorbarendezésére: $6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 6! = 720$.

- Az **ismétlés nélküli kombináció** azt mondja meg, hogy n különböző elemből k különbözőt hányféleképpen választhatunk ki úgy, hogy a kiválasztás **sorrend**je **nem számít**. A lehetőségek száma $\binom{n}{k}$ ('n alatt a k').

  Megjegyzés: $\binom{n}{k} = \dfrac{n!}{k! \cdot (n-k)!}$.

  8. feladat Vizsgán 52 tételből 4-et kell kihúzni. Hányféleképpen tehetjük ezt meg? (A tételek húzásának sorrendje nem számít.)

  Megoldás: 52 különböző elemünk van, ebből 4 különbözőt húzunk, de az, hogy milyen sorrendben húzzuk a tételeket, nem számít. Ezt

  $$\binom{52}{4} = \frac{52!}{4! \cdot (52-4)!} = \frac{52!}{4! \cdot (28!)} = 270\,725$$

  -féleképpen tehetjük meg.

- Az **ismétlés nélküli variáció** esetében n különböző elemből k különbözőt választunk ki úgy, hogy a kiválasztás sorrendje számít. Ezt $\dfrac{n!}{(n-k)!}$ féleképpen tehetjük meg.

  Megjegyzés: $\dfrac{n!}{(n-k)!} = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)$.

  9. feladat Tekintsük az 1,2,3,4,5,6 számjegyeket. Hány különböző 3-jegyű számot hozhatunk létre belőlük úgy, hogy minden számjegyet legfeljebb egyszer használunk fel?
  Megoldás: Kövessük a 7. feladatban látott gondolatmentet: az első számjegy kiválasztására hat lehetőségünk van, a másodikra öt, a harmadikra négy. Több számjegyre nincs szükség, tehát a lehetőségek száma:

  $$6 \cdot 5 \cdot 4 = 120$$

- **Ismétléses variáció** esetén n különböző elemből k nem feltétlenül különbözőt választunk ki (vagyis a kiválasztás során ugyanazt az elemet akár többször is kiválaszthatjuk) úgy, hogy a kiválasztás sorrendje számít. Ezt n^k -féleképpen tehetjük meg.

  10. feladat Továbbra is tekintsük az 1,2,3,4,5,6 számjegyeket. Hány különböző 4-jegyű számot hozhatunk létre belőlük úgy, hogy minden számjegyet többször is felhasználhatunk?

  Megoldás: Az első számjegyet hatból választhatjuk, a másodikat szintén hatból (az első számjegy most újra kiválasztható), majd a harmadikat és negyediket is hatból. A megoldás így:

  $$6 \cdot 6 \cdot 6 \cdot 6 = 6^4 = 1296$$

Alkalmazzuk a következő feladatokon az eddig tanultakat!

11. feladat Egy ládában 10 hibás és 40 hibátlan izzó van. Kihúzunk a ládából 5 izzót *visszatevés nélkül* (azaz a kihúzott izzót mindig félretesszük). Mi a valószínűsége annak, hogy a kihúzottak között

a) pontosan 2 hibás van?

b) van hibás?

c) legalább 2 hibás van?

Megoldás:

a) Arra vonatkozóan nincs kikötés a feladatban, hogy a húzások sorrendje számít-e vagy nem. Megmutatható, hogy ugyan az eseménytér a két esetben nem ugyanakkora, de a valószínűség mindkét esetben megegyezik. Ezért számításoknál egyszerűbb, ha a sorrendet nem vesszük figyelembe (ekkor ugyanis ismétlés nélküli kombinációra vezet a feladat.)

Ha nem számít a sorrend, akkor az eseménytérben $\binom{50}{5}$ elemi esemény van. A 10 hibásból 2 izzót $\binom{10}{2}$ féleképpen, és a 40 hibátlanból 3 izzót $\binom{40}{3}$ féleképpen választhatunk ki, a kedvező kimenetelek (2 hibás és 3 hibátlan izzót húzunk) száma így $\binom{10}{2} \cdot \binom{40}{3}$. A keresett valószínűség: $\dfrac{\binom{10}{2} \cdot \binom{40}{3}}{\binom{50}{5}} = 0,2098$.

b) Az, hogy van hibás a kihúzott izzók között, az jelenti, hogy legalább 1 hibás van, azaz 1 vagy 2 vagy 3 vagy 4 vagy 5 hibásat húzunk. A 'van hibás' esemény ellentettje viszont az, hogy 'nincs hibás', ennek a valószínűségét pedig egyszerűen számolhatjuk: $\dfrac{\binom{10}{0} \cdot \binom{40}{5}}{\binom{50}{5}} = 0,3106$. A keresett valószínűség: $1 - 0,3106 = 0,6894$.

c) A legalább két hibás izzó a kihúzottak között azt jelenti, hogy 2 vagy 3 vagy 4 vagy 5 hibás izzót húzunk. Mivel ez túl sok eset, érdemes megnézni az ellentettjét. A legalább két hibásat húzunk esemény ellentettje, hogy 2-nél kevesebb hibásat húzunk, ez pedig azt jelenti, hogy 0 vagy 1 hibásat húzunk. Jelölje A_0 azt az eseményt, hogy '0 hibásat húzunk', A_1 pedig azt, hogy '1 hibásat húzunk'. A két esemény kizárja egymást, így annak valószínűsége, hogy 0 vagy 1 hibásat húzunk:

$$P(A_0 \cup A_1) = P(A_0) + P(A_1) = \frac{\binom{10}{0} \cdot \binom{40}{5}}{\binom{50}{5}} + \frac{\binom{10}{1} \cdot \binom{40}{4}}{\binom{50}{5}} = 0,7419$$

A legalább 2 hibás húzásának valószínűsége az ellentett esemény valószínűségére vonatkozó összefüggéssel: $1 - 0,7419 = 0,2581$.

12. feladat Egy ládában 10 hibás és 40 hibátlan izzó van. Kihúzunk a ládából 5 izzót *visszatevéssel* (azaz a kihúzott izzót a húzás után azonnal visszatesszük a ládába). Mi a valószínűsége annak, hogy a kihúzottak között

a) pontosan 2 hibás van?

b) van hibás?

c) legalább 2 hibás van?

Megoldás:

a) Ugyanazt az izzót most akár többször is kihúzhatjuk a ládából, és a húzás sorrendje számít. Az eseménytér most $50 \cdot 50 \cdot 50 \cdot 50 \cdot 50 = 50^5 =$ elemi eseményből áll. 2 hibás és 3 hibátlan izzót akarunk húzni. Az első két izzó hibás és a következő három hibátlan $10 \cdot 10 \cdot 40 \cdot 40 \cdot 40 = 10^2 \cdot 40^3$ féleképpen lehet, de nekünk az 5 húzásból nem csak az első kettő lehet hibás, hanem bármelyik kettő. Az összes lehetőséget úgy kapjuk meg, hogy a megmondjuk, a két hibásat hányféleképpen helyezhetjük el az 5 húzás között: $\binom{5}{2}$ féleképpen. A kedvező kimenetelek száma így $10^2 \cdot 40^3 \cdot \binom{5}{2}$. A kérdésre adott válasz:

$$\frac{10^2 \cdot 40^3 \cdot \binom{5}{2}}{50^5} = 0,2048$$

b) Az előző feladathoz hasonlóan az ellentett esemény valószínűségével számolunk:

$$1 - \frac{10^0 \cdot 40^5 \cdot \binom{5}{0}}{50^5} = 0,6723$$

c) Szintén az előző feladat c) megoldása szerinti gondolatmenet alkalmazzuk a visszatevéses mintavételre:

$$1 - \left[\frac{10^0 \cdot 40^5 \cdot \binom{5}{0}}{50^5} + \frac{10^1 \cdot 40^4 \cdot \binom{5}{1}}{50^5}\right] = 0,2627$$

13. feladat Egy családban $0,5 - 0,5$ valószínűséggel születik fiú, illetve lány gyermek.

a) Mi a valószínűsége annak, hogy egy családban mindkét gyermek lány?

b) Feltéve, hogy az egyik gyermek lány, mi a valószínűsége, hogy a másik fiú?

Megoldás: Jelölje A_1 eseményt azt, hogy 'az első gyermek lány', A_2 pedig azt, hogy 'a második lány'. Jelölje F a fiú, L a lány gyermeket. Az eseménytér: $\{(F,F),(F,L),(L,F),(L,L)\}$.

a) A 4 elemű eseménytérben egyetlen kedvező elemi esemény van, a (L,L), így a keresett valószínűség:

$$\frac{1}{4} = 0,25$$

Kihasználhatjuk ugyanakkor az események (feltételezett) függetlenségét is, ugyanarra az eredményre jutunk: A_1 és A_2 egymástól függetlenek, így $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = 0,5 \cdot 0,5 = 0,25$.

b) Szűkítsük le az eseményteret a feltételnek megfelelően azokra az elemi eseményekre, mikor az egyik gyermek lány: $\{(F,L),(L,F),(L,L)\}$. A leszűkített eseménytérben már csak 3 elemi esemény van, ebből kettő esetén van egy fiú közöttük: $\{(F,L),(L,F)\}$. A megoldás tehát

$$\frac{2}{3} = 0,6667$$

14. feladat Tudjuk, hogy a férfiak 5%-a színvak, a nőknél 0,5% ez az arány. Egy 30 férfiból és 40 nőből álló csoportból véletlenszerűen kiválasztunk egy embert. Mi a valószínűsége annak, hogy színvakot választunk?

Megoldás: Jelölje A azt az eseményt, hogy 'a kiválasztott színvak'. Ennek az eseménynek a valószínűsége attól függ, hogy férfi vagy nő a kiválasztott, ezért jelölje B azt az eseményt, hogy 'férfi a kiválasztott', ekkor $\overline{B}$ pedig az az eseményt, hogy 'nő a kiválasztott'. Amit a feladat szövegéből tudunk:

$$P(A \mid B) = 0,05$$
$$P(A \mid \overline{B}) = 0,005$$
$$P(B) = \frac{30}{70}$$
$$P(\overline{B}) = \frac{40}{70}$$

Így

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B}) = 0,05 \cdot \frac{30}{70} + 0,005 \cdot \frac{40}{70} = 0,0243$$

15. feladat Egy céltáblára 8 lövést adunk le. Tudjuk, hogy a találat valószínűsége minden egyes lövésnél (egymástól függetlenül) $0,2$.

a) Feltéve, hogy az első lövés talál, mi a valószínűsége annak, hogy a második is talál?

b) Mi a valószínűsége annak, hogy az első lövés talál és a második is talál?

c) Mi a valószínűsége annak, hogy minden lövés talál?

d) Mi a valószínűsége annak, hogy a lövések közül egy sem talál?

Megoldás: Jelölje A_i azt az eseményt, hogy 'az i-edik lövés talál' ($i = 1, 2, \ldots, 8$). Ezek az események függetlenek, valamint $P(A_i) = 0,2$, $P(\overline{A_i}) = 0,8$.

a) Használjuk ki A_1 és A_2 függetlenségét a feltételes valószínűségre:

$$P(A_2 \mid A_1) = P(A_2) = 0,2$$

b)

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = 0,2 \cdot 0,2 = 0,04$$

c)

$$P(A_1 \cap A_2 \cap \ldots \cap A_8) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_8) = 0,2 \cdot 0,2 \cdot \ldots \cdot 0,2 = 0,2^8 = 0,00000256$$

d) Kihasználjuk, hogy az ellentettek 'öröklik' a függetlenséget:

$$P(\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_8}) = P(\overline{A_1}) \cdot P(\overline{A_2}) \cdot \ldots \cdot P(\overline{A_8}) = 0,8 \cdot 0,8 \cdot \ldots \cdot 0,8 = 0,8^8 = 0,1678$$

2.2. Valószínűségi változó

2.2.1. A diszkrét valószínűségi változó és jellemzői

Definíció: Legyen X olyan függvény, amely az eseményteret leképezi a valós számok halmazára.

$$X\colon \Omega \longmapsto \mathbf{R}$$

Ekkor X-et **valószínűségi változó**nak nevezzük.

Egy X valószínűségi változó tehát minden elemi eseményhez egyértelműen hozzárendel egy valós számot. Ezeket a valós számokat a valószínűségi változó **lehetséges értékei**nek nevezzük.

Az X-et **diszkrét valószínűségi változó**nak nevezzük, ha lehetsége értékei megszámlálhatóak. Tehát egy diszkrét valószínűségi változónak véges sok, vagy megszámlálhatóan végtelen sok lehetséges értéke van.
A valószínűségi változó jelölésére általában nagy betűket használunk (X, Y, X_1, X_2, stb.), a lehetséges (felvett) értékek jelölésére kis betűket (x_1, x_2, stb.).

Az X diszkrét valószínűségi változó lehetséges értékeit jelölje x_i $(i = 1, 2, \ldots)$. Ekkor $\{X = x_i\}$ azon elemi eseményeket jelenti, amelyekre X éppen x_i értéket vesz fel. Mivel ezen elemi események halmaza egyben esemény is, így $\{X = x_i\}$ egy eseményt jelent. Az x_i- hez $(i = 1, 2, ...)$ tartozó valószínűségeket jelölje rendre $p_1, p_2, \ldots$. (Azaz $P(X = x_i) = p_i$, $i = 1, 2, \ldots$-re. Pl. annak valószínűsége, hogy az X értéke éppen x_1 lesz, p_1.)

A lecke elején láttuk azt a példát, mikor 2 szabályos pénzérmét feldobunk. Az eseménytér: $\Omega = \{FF, FI, IF, II\}$. Ezen az eseménytéren akarunk értelmezni egy valószínűségi változót. Ehhez a megfigyelésünk szempontját érdemes figyelembe venni. Ha a dobott fejek száma fontos, akkor legyen X a dobott fejek száma. A dobott fejek száma 0,1,2 lehet. Ezért X lehetséges értékei: $x_1 = 0$, $x_2 = 1$, $x_3 = 2$. $\{X = x_2\}$ például azt az eseményt jelenti, hogy a dobott fejek száma $x_2 = 1$. (Ennek p_2 valószínűsége pedig könnyedén számolható.) A lehetséges értékekhez tartozó valószínűségek:

$$P(X = x_1) = P(X = 0) = \frac{1}{4} = p_1$$
$$P(X = x_2) = P(X = 1) = \frac{2}{4} = p_2$$
$$P(X = x_3) = P(X = 2) = \frac{1}{4} = p_3$$

A diszkrét valószínűségi változó jellemzői:

- **Definíció:** Az X diszkrét valószínűségi változó **eloszlásá**n az x_i lehetséges értékeket és a hozzájuk tartozó p_i valószínűségeket értjük:

  $$X : \begin{cases} x_1 & x_2 & \ldots \\ p_1 & p_2 & \ldots \end{cases}$$

  A p_i valószínűség kiszámításához gyakran egy (diszkrét helyeken értelmezett, azaz egyetlen pontban sem folytonos) függvény áll rendelkezésünkre.

  Diszkrét valószínűségi változó eloszlásában a valószínűségek összege mindig 1. Azaz

  $$\sum_i p_i = 1$$

- **Definíció:** Az X diszkrét valószínűségi változó **várható értékén** azt a $E(X)$-szel jelölt valós számot értjük, melyre

$$E(X) = \sum_i x_i \cdot p_i$$

(ha létezik és véges az összeg.) Megszámlálhatóan végtelen lehetséges érték esetén nem biztos, hogy az összeg létezik és véges.
A gyakorisági hisztogramoknál láttuk már a következő példát:

Példa Egy szabályos dobókockát 60-szor feldobva az alábbi eredmények adódtak:

Dobott szám	1	2	3	4	5	6	
Gyakoriság	7	12	14	9	10	8	Összesen: 60

A megfigyelésünk eredményét felhasználva a dobott számok átlagát a tanult módon számolhatjuk:

$$\overline{x} = \frac{1 \cdot 7 + 2 \cdot 12 + 3 \cdot 14 + 4 \cdot 9 + 5 \cdot 10 + 6 \cdot 8}{60} = \frac{207}{60} = 3,45$$

A várható érték kiszámításához nem kell nagyszámú megfigyelés, nem dobunk 60-szor. Legyen most az X valószínűségi változó egy szabályos dobókockát feldobva a dobott szám értéke. A dobás értéke 6 különböző szám lehet, így X-nek 6 lehetséges értéke van: $1, 2, 3, 4, 5, 6$. Mindegyikhez ugyanakkkora, $\frac{1}{6}$ valószínűség tartozik, tehát X eloszlása:

$$X : \begin{cases} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{cases}$$

Az (elméleti) várható érték:

$$E(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + x_3 \cdot p_3 + x_4 \cdot p_4 + x_5 \cdot p_5 + x_6 \cdot p_6 =$$
$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3,5$$

Kérdés, hogy hogyan számolhatunk, ha pl. a dobott szám 3-szorosát tekintjük a valószínűségi változónak, esetleg a dobott szám négyzete a valószínűségi változó. Könnyen belátható, hogy ezekben az esetekben az eloszlásban csak a lehetséges értékek változnak, a valószínűségek nem (pl. a dobott szám négyzete 9 csak akkor lehet, ha 3-ast dobunk, ennek a valószínűsége pedig $\frac{1}{6}$). Ha az új valószínűségi változó $3 \cdot X$, akkor eloszlása:

$$3 \cdot X : \begin{cases} 3 & 6 & 9 & 12 & 15 & 18 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{cases}$$

$3 \cdot X$ várható értéke:

$$E(3 \cdot X) = 3 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} + 12 \cdot \frac{1}{6} + 15 \cdot \frac{1}{6} + 18 \cdot \frac{1}{6} = \frac{63}{6} = 10,5 = 3 \cdot E(X)$$

Állítás: Ha $E(X^2)$ és $E(X)$ létezik, és a, b, c tetszőleges valós számok, akkor $E(a \cdot X^2 + b \cdot X + c)$ is létezik és

$$E(a \cdot X^2 + b \cdot X + c) = a \cdot E(X^2) + b \cdot E(X) + c$$

Megjegyzés: A várható érték akár negatív szám is lehet, mivel a valószínűségi változó negatív értéket is felvehet. Gondoljunk arra, mikor egy játékban a nyeremény összege a valószínűségi változó. Ha nyerünk, a valószínűségi változó értéke pozitív, viszont ha veszítünk, akkor az elvesztett összeg negatív előjellel lesz a valószínűségi változó értéke.

- **Definíció:** Az X diszkrét valószínűségi változó **szórása** az a $D(X)$-szel jelölt szám amelyre:

$$D(X) = \sqrt{E\big((X - E(X))^2\big)}$$

azaz a szórás a valószínűségi változónak saját várható értékétől való eltérés négyzetének várható értékéből vont négyzetgyök, ha $E(X)$ és $E(X - E(X))$ létezik.

(Emlékezzünk vissza, hogy sokaság esetében a tapasztalati szórás értéke a sokaság elemeinek az átlagtól való eltérésének négyzetösszege átlagának gyöke volt.)

A formulát egyszerűbb alakban is felírhatjuk:

$$D(X) = \sqrt{E\big((X - E(X))^2\big)} = \sqrt{E\big(X^2 - 2 \cdot X \cdot E(X) + E^2(X)\big)} =$$

$$= \sqrt{E(X^2) - 2 \cdot E^2(X) + E^2(X)} = \sqrt{E(X^2) - E^2(X)}$$

Számításoknál érdemes ezt az alakot használni:

$$D(X) = \sqrt{E(X^2) - E^2(X)}$$

ahol $E(X^2) = \sum_i x_i^2 \cdot p_i$. $E(X^2)$-et az X valószínűségi változó második momentumának is szokás nevezni.

Megjegyzés: A szórás (a várható értéktől eltérően) nem lehet negatív szám!

Állítás: Ha az X valószínűségi változó $D(X)$ szórása létezik és a, b tetszőleges valós számok, akkor:

$$D(a \cdot X + b) =| a | \cdot D(X)$$

(Felhasználtuk, hogy ha X mindössze egyetlen b értéket vehet fel, akkor $D(b) = 0$.)

16. feladat Adott az X valószínűségi változó eloszlása:

$$X : \begin{cases} 0 & 1 & 2 & 3 \\ \dfrac{125}{216} & \dfrac{75}{216} & \dfrac{15}{216} & \dfrac{1}{6} \end{cases}$$

a) $D(X) =?$

b) $D(4 \cdot X - 3)$=?

Megoldás:

a) A szórás kiszámításához első lépésben az $E(X)$ és $E(X^2)$ értékére van szükség:

$$E(X) = 0 \cdot \frac{125}{216} + 1 \cdot \frac{75}{216} + 2 \cdot \frac{15}{216} + 3 \cdot \frac{1}{216} = \frac{108}{216} = \frac{1}{2}$$

$$E(X^2) = 0^2 \cdot \frac{125}{216} + 1^2 \cdot \frac{75}{216} + 2^2 \cdot \frac{15}{216} + 3^2 \cdot \frac{1}{216} = \frac{144}{216} = \frac{2}{3}$$

Ezután a szórás:

$$D(X) = \sqrt{E(X^2) - E^2(X)} = \sqrt{\frac{2}{3} - \left(\frac{1}{2}\right)^2} = \sqrt{\frac{5}{12}} = 0,6455$$

b)
$$D(4 \cdot X - 3) = \mid 4 \mid \cdot D(X) = 4 \cdot D(X) = 4 \cdot 0,6455 = 2,5820$$

- Az előző leckében láttuk, hogy ha egy sokasághoz tartozó gyakorisági eloszlás adott, akkor a kumulatív gyakoriságok kiszámításához csak összegezni kellett az adott értéknél kisebb gyakoriságokat. Hasonló módon járhatunk el egy diszkrét X valószínűségi változó esetén a valószínűségekkel. Az X diszkrét valószínűségi változó lehetséges értékei legyenek $x_1, x_2, \ldots$, a hozzájuk tartozó valószínűségek rendre $p_1, p_2, \ldots$. Ekkor X **eloszlásfüggvénye** az a $F(x)$-szel jelölt függvény, amelyre
$$F(x) = P(X < x) = \sum_{i:x_i<x} p_i$$
minden x valós szám esetén.

  Példa A 16. feladatban látott X valószínűségi változó eloszlásfüggvénye a következő:
  X eloszlása:
$$X : \begin{cases} 0 & 1 & 2 & 3 \\ \frac{125}{216} & \frac{75}{216} & \frac{15}{216} & \frac{1}{6} \end{cases}$$
  X eloszlásfüggvénye:
$$F(x) = \begin{cases} 0 & \text{ha } x \leq 0 \\ \frac{125}{216} & \text{ha } 0 < x \leq 1 \\ \frac{125}{216} + \frac{75}{216} = \frac{200}{216} & \text{ha } 1 < x \leq 2 \\ \frac{125}{216} + \frac{75}{216} + \frac{15}{216} = \frac{215}{216} & \text{ha } 2 < x \leq 3 \\ \frac{125}{216} + \frac{75}{216} + \frac{15}{216} + \frac{1}{206} = \frac{216}{216} = 1 & \text{ha } x > 3 \end{cases}$$
  Azaz
$$F(x) = \begin{cases} 0 & \text{ha } x \leq 0 \\ \frac{125}{216} & \text{ha } 0 < x \leq 1 \\ \frac{200}{216} & \text{ha } 1 < x \leq 2 \\ \frac{215}{216} & \text{ha } 2 < x \leq 3 \\ 1 & \text{ha } x > 3 \end{cases}$$

Megjegyzés: Az eloszlásfüggvényt diszkrét valószínűségi változó esetén ritkán használjuk számítási feladatokhoz.

17. feladat Egy pénzérmét egymás után 3-szor feldobunk. Az X valószínűségi változó legyen a dobott fejek száma.

a) Írjuk fel X eloszlását!

b) $E(X)$=?

c) $D(X)$=?

Megoldás: Egyetlen dobás esetén a fejdobás valószínűsége $\frac{1}{2}$, a dobások egymástól függetlenek.

a) A kísérlet során a dobott fejek száma lehet 0, 1, 2, 3. X lehetséges értékei tehát:

$$x_1 = 0 \text{ (0 fejet dobunk)}$$

$$x_2 = 1 \text{ (1 fejet dobunk)}$$

$$x_3 = 2 \text{ (2 fejet dobunk)}$$

$$x_4 = 3 \text{ (3 fejet dobunk)}$$

A hozzájuk tartozó valószínűségek:

$$P(x_1 = 0) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

$$P(x_2 = 1) = \left(\frac{1}{2}\right)^3 \cdot \binom{3}{1} = \frac{3}{8}$$

$$P(x_3 = 2) = \left(\frac{1}{2}\right)^3 \cdot \binom{3}{2} = \frac{3}{8}$$

$$P(x_4 = 3) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

Az eloszlás tehát:

$$X : \left\{ \begin{array}{cccc} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array} \right.$$

b)

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1,5$$

c)

$$E(X^2) = 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{8} = \frac{24}{8} = 3$$

$$D(X) = \sqrt{3 - 1,5^2} = 0,8660$$

2.2.2. Folytonos valószínűségi változó és jellemzői

A folytonos valószínűségi változó az olyan, előző leckében látott folytonos sokaságok leírására szolgál, mint pl. a hossz (ld. nordman fenyős példa), idő (ld. telefonhívások hossza) stb. Ekkor a valószínűségi változó értékeit nem tudjuk felsorolni (elméletben tetszőleges mérési pontosságot tekintve), hanem értékeire intervallumo(ka)t adhatunk meg. Például egy termék hossza csak nemnegatív szám lehet, de termék hosszának a mérethibája már tetszőleges valós szám lehet.

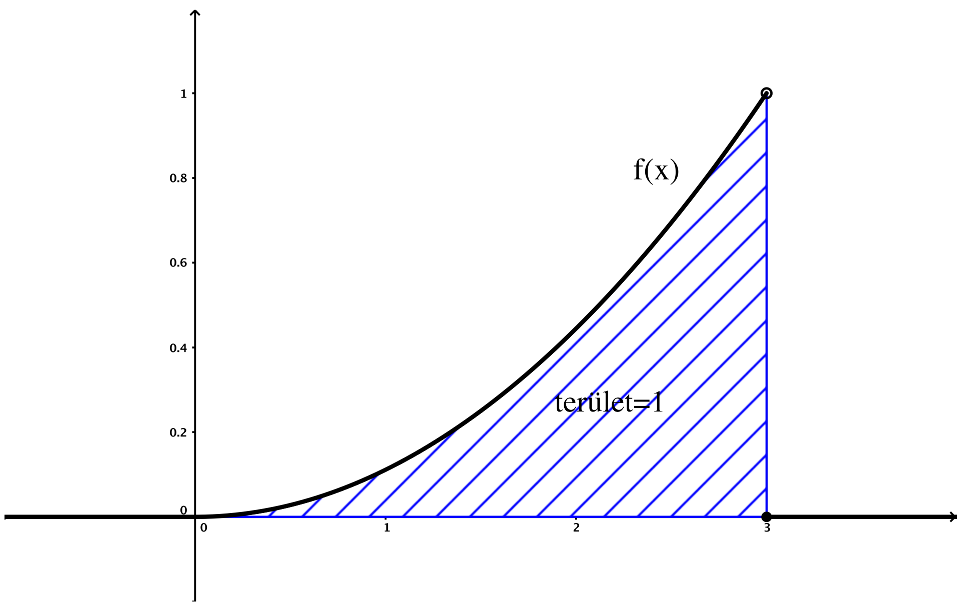
A folytonos valószínűségi változó jellemzői:

- A folytonos valószínűségi változót az $f(x)$-szel jelölt **sűrűségfüggvénnyel** jellemezzük. A sűrűségfüggvény **főbb tulajdonságai** a következők:

  a) A sűrűségfüggvény **nem vehet fel negatív értéket**: $f(x) \geq 0$. (Nulla azonban lehet az értéke.)

  b) Egy sűrűségfüggvény **teljes görbe alatti területe 1**.

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$



12. ábra. Sűrűségfüggvény görbe alatti területe

  c) A $P(a \leq x < b)$ valószínűség a sűrűségfüggvény [a,b) intervallumra vett görbe alatti területe.

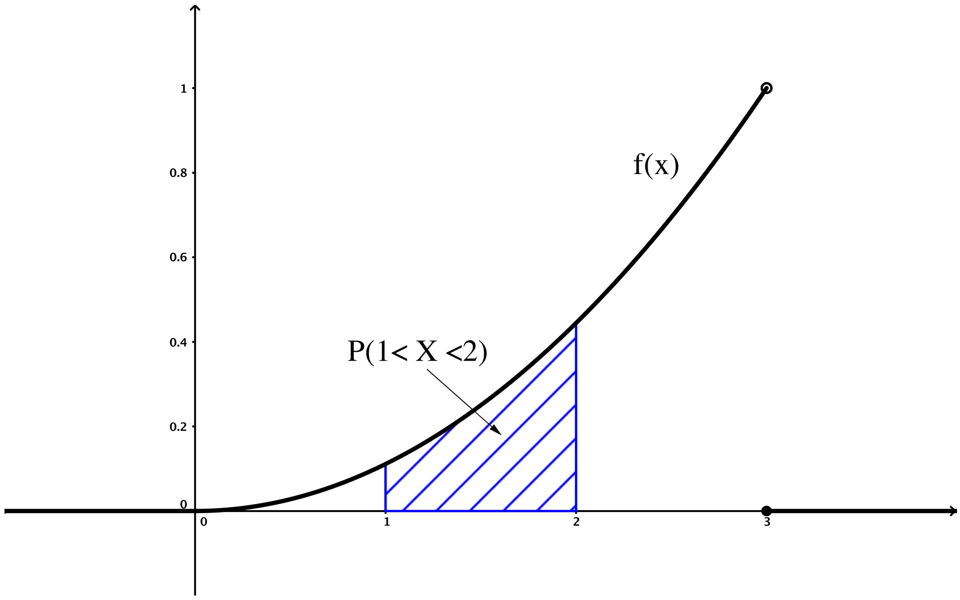
$$P(a \leq x < b) = \int_a^b f(x)\mathrm{d}x$$

  Az integrál tulajdonságai miatt

$$P(a \leq x < b) = P(a < x < b) = \int_a^b f(x)\mathrm{d}x$$

A valószínűségek kiszámításának ez az integrálos módja alkalmat ad nekünk arra, hogy a valószínűséget a sűrűségfüggvény grafikonján területekkel ábrázoljuk. Ezzel némi hasonlóságot vélhetünk felfedezni az előző fejezet sűrűségi hisztogramjaival, ahol a téglalapok összterülete 1, és az egyes téglalapok területe az adott érték sokaságbeli előfordulási arányát adja meg.

Megjegyzés: Az(oka)t az intervallumo(ka)t, amelyekben a sűrűségfüggvény értéke nem nulla, a továbbiakban alapintervallum(ok)nak nevezzük.



13. ábra. Valószínűség ábrázolása sűrűségfüggvény grafikonján

- **Definíció:** Az X folytonos valószínűségi változó **várható értéke**:

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

amennyiben az integrál létezik és véges.

- **Definíció:** Ha az X folytonos valószínűségi változó esetén létezik $E(X)$ és $E(X^2)$, akkor X **szórása**:

$$D(X) = \sqrt{E(X^2) - E^2(X)}$$

ahol folytonos valószínűségi változó esetén $E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) \mathrm{d}x$.

A várható értékre és a szórásra vonatkozó, diszkrét valószínűségi változónál látott összefüggések folytonos valószínűségi változóra is fennállnak:

Állítás: Ha az X valószínűségi változónak létezik $E(X)$ várható értéke és $E(X^2)$ második momentuma és a, b, c tetszőleges valós számok, akkor

$$E(a \cdot X^2 + b \cdot X + c) = a \cdot E(X^2) + b \cdot E(X) + c$$

Állítás: Ha az X valószínűségi változónak létezik $D(X)$ szórása és a, b tetszőleges valós számok, akkor

$$D(a \cdot X + b) = \mid a \mid \cdot D(X)$$

teljesül.
(Megjegyezzük, hogy természetesen a fenti összefüggések akkor teljesülnek, ha X-nek létezik várható értéke és szórása, illetve az első összefüggésnél X^2-nek létezik várható értéke.)

18. feladat Az X valószínűségi változó sűrűségfüggvénye legyen $f(x) = \begin{cases} \frac{24}{x^4} & \text{ha } 2 < x \\ 0 & \text{különben} \end{cases}$.

a) Ellenőrizzük, hogy valóban sűrűségfüggvény-e $f(x)$!
b) Határozzuk meg a $P(3 \leq X < 4)$ valószínűséget!
c) $P(1 \leq X < 4)$=?
d) $E(X) =$?
e) $D(X) =$?
f) $P\big(\mid X - E(X) \mid < D(X)\big)$=?

Megoldás:

a) Ellenőriznünk kell, hogy $f(x)$ nemnegatív értékű, illetve a görbe alatti területe 1. A hozzárendelési utasításból látszik, hogy $f(x)$ nemnegatív értékű (a tört számlálója és nevezője is pozitív szám). Az integrál kiszámításánál vegyük figyelembe, hogy ahol $f(x)$ értéke 0, ott az integrálja is nulla, tehát a $(-\infty, 2)$ intevallumon az integrál nulla, számolnunk csak a $(2, \infty)$ alapintervallumon kell:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_{2}^{\infty} \frac{24}{x^4}\mathrm{d}x = \Big[-\frac{8}{x^3} \Big]_2^{\infty} = 0 - \Big(-\frac{8}{2^3} \Big) = 1$$

Tehát $f(x)$ valóban sűrűségfüggvény.

b)

$$P(3 \leq x < 4) = \int_3^4 f(x)\mathrm{d}x = \int_3^4 \frac{24}{x^4}\mathrm{d}x = \Big[-\frac{8}{x^3} \Big]_3^4 = -\frac{8}{4^3} - \Big(-\frac{8}{3^3} \Big) = 0,1713$$

c) Vegyük észre, hogy az [1,4) integrálási tartományt két részre kell bontanunk: az [1,2) intervallumon a $f(x)$ értéke nulla, így ott az integrál is nulla, számolnunk csak a [2,4) intervallumon kell.

$$P(1 \leq X < 4) = \int_1^4 f(x)\mathrm{d}x = \int_2^4 \frac{24}{x^4}\mathrm{d}x = \Big[-\frac{8}{x^3} \Big]_2^4 = -\frac{8}{4^3} - \Big(-\frac{8}{2^3} \Big) = 0,8750$$

d) Az a) részhez hasonlóan ahol $f(x)$ nulla, ott $x \cdot f(x)$ is nulla, tehát az alapintervallumra szűkül az integrálási tartomány.

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x = \int_2^{\infty} x \cdot \frac{24}{x^4}\mathrm{d}x = \int_2^{\infty} \frac{24}{x^3}\mathrm{d}x = \Big[-\frac{12}{x^2} \Big]_2^{\infty} = 0 - \Big(-\frac{12}{2^2} \Big) = 3$$

e) Először $E(X^2)$-et számoljuk ki, figyelembe véve, hogy csak az alapintervallumon kell integrálni.

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\mathrm{d}x = \int_2^{\infty} x^2 \cdot \frac{24}{x^4}\mathrm{d}x = \int_2^{\infty} \frac{24}{x^2}\mathrm{d}x = \Big[-\frac{24}{x} \Big]_2^{\infty} = 0 - \Big(-\frac{24}{2} \Big) = 12$$

Ezt felhasználva

$$D(X) = \sqrt{E(X^2) - E^2(X)} = \sqrt{12 - 3^2} = \sqrt{3} = 1,7321$$

f) A kérdés szavakkal úgy is megfogalmazható, hogy mi a valószínűsége annak, hogy az X valószínűségi változó a várható értékétől a szórásánál kevesebbel tér el.

$$P\big(\mid X - E(X) \mid < D(X)\big) = P\big(-D(X) < X - E(X) < D(X)\big) =$$

$$= P\big(E(X) - D(X) < X < E(X) + D(X)\big) =$$

$$= P(3 - 1,7321 < X < 3 + 1,7321) = P(1,2679 < X < 4,7321) =$$

$$= \int_{1,2679}^{4,7321} f(x)\mathrm{d}x = \int_{2}^{4,7321} \frac{24}{x^4}\mathrm{d}x = \Big[-\frac{8}{x^3}\Big]_2^{4,7321} = -\frac{8}{4,7321^3} - \Big(-\frac{8}{2^3}\Big) = 0,9245$$

(Az integrálási tartományt itt is az alapintervallumhoz kellett igazítani.)

Megjegyzés: A sűrűségfüggvény maximumhelye a módusz. A grafikon ábrázolásával a módusz értéke nagyságrendileg meghatározható, bizonyos esetekben pedig felhasználhatjuk a függvények szélsőérték számításánál tanultakat a maximumhely pontos meghatározására (természetesen az alapintervallumon belül).

Azt láttuk, hogy a sűrűségfüggvény remek eszköz a valószínűségek szemléltetésére, a számításoknál remekül használható, de két hátránya van: egyrészt csak folytonos valószínűségi változó esetén létezik, másrészt a számításokhoz minden esetben integráni kell. Létezik-e olyan eszköz, amellyel folytonos valószínűségi változó esetén (is) egyszerűen tudunk valószínűségeket számolni?

- A diszkrét valószínűségi változóhoz hasonlóan értelmezhetjük az **eloszlásfüggvény**t folytonos valószínűségi változóra is:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

minden $x \in \mathbf{R}$ esetén.

Az eloszlásfüggvény **főbb tulajdonságai**:

a)

$$0 \leq F(x) \leq 1$$

Értéke 0 és 1 közé esik, mivel értékkészlete események valószínűségeiből áll.

b)

$$\lim_{x\to\infty} F(x) = 1, \quad \lim_{x\to-\infty} F(x) = 0$$

(Határértéke $-\infty$-ben nulla, ∞-ben 1.)

c)

$$P(a \leq X < b) = F(b) - F(a)$$

mivel $P(X < b) = F(b)$ és $P(X < a) = F(a)$.

d)

$$f(x) = F'(x)$$

A sűrűségfüggvény értéke egy x_0 helyen az eloszlásfüggvény meredekségével egyezik meg az adott x_0 helyen. Ha az eloszlásfüggvény nem differenciálható egy x_0 helyen, akkor a sűrűségfüggvény értékét x_0-ban nullának tekintjük.

e) **A medián a sűrűségfüggvény görbe alatti területét két azonos nagyságú részre osztja**. A teljes görbe alatti terület 1, a medián ezt felezi, így a medián *med* értékére igaz, hogy

$$0,5 = \int_{-\infty}^{med} f(x)\mathrm{d}x$$

így

$$F(med) = 0.5$$

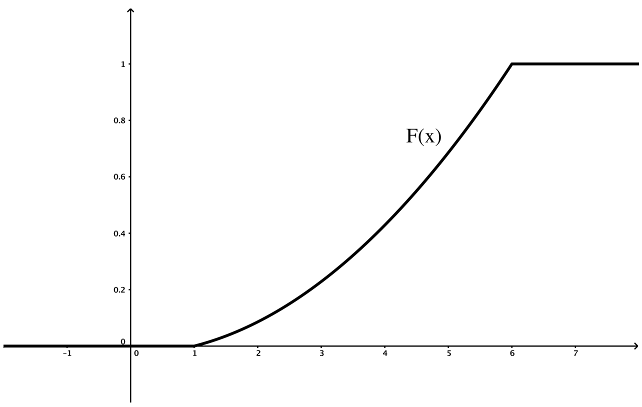
Természetesen a medián mindig beleesik az alapintervallumba.
Ha a sűrűségfüggvény valamely $x = x_0$ egyenesre szimmetrikus (ilyen esetben szimmetrikus eloszlásról beszélhetünk), akkor a medián és a várható érték értéke megyegyezik.

19. feladat A X valószínűségi változó eloszlásfüggvénye legyen $F(x) = \begin{cases} 0 & \text{ha } x \leq 1 \\ \frac{x^2-1}{35} & \text{ha } 1 < x \leq 6 \\ 1 & \text{ha } 6 < x \end{cases}$

a) Ábrázoljuk $F(x)$-et!
b) $P(X < 4)$=?
c) $P(2 \leq X < 4)$=?
d) $P(0 \leq X < 4)$=?
e) $P(3 \leq X)$=?
f) $f(x)$=?
g) Mekkora a medián értéke?

Megoldás:

a)



14. ábra.

b)

$$P(X < 4) = F(4) = \frac{4^2 - 1}{35} = \frac{15}{35} = 0,4286$$

c)

$$P(2 \leq X < 4) = F(4) - F(2) = \frac{4^2 - 1}{35} - \frac{2^2 - 1}{35} = \frac{15}{35} - \frac{3}{35} = \frac{12}{35} = 0,3429$$

d) Vegyük észre, hogy az eloszlásfüggvényünk értéke 0-ban nulla (a hozzárendelési utasítás első sorát tekintsük)! Így

$$P(0 \leq X < 4) = F(4) - F(0) = \frac{4^2 - 1}{35} - 0 = \frac{15}{35} - 0 = \frac{15}{35} = 0,4286$$

e) Kihasználjuk az ellentett esemény valószínűségére korábban felírt összefüggést, és azt, hogy a $\{3 \leq X\}$ esemény ellentettje az $\{X < 3\}$ esemény.

$$P(3 \leq X) = 1 - P(X < 3) = 1 - F(3) = 1 - \frac{3^2 - 1}{35} = 1 - \frac{8}{35} = 0,7714$$

f) A grafikonra pillantva látható, hogy $f(x)$ csak az $(1,6)$ intervallumon fog 0-tól különbözni. (A $-\infty, 1]$ és a $(6, \infty)$ intervallumokon a derivált nulla, az $x = 6$ helyen pedig nem differenciálható $F(x)$, ezért a sűrűségfüggvény értéke itt is 0.) Így

$$f(x) = \begin{cases} \frac{2x}{35} & \text{ha } 1 < x < 6 \\ 0 & \text{különben} \end{cases}$$

g) A mediánra(med) vonatkozó összefüggést felhasználva:

$$0,5 = F(med) = \frac{med^2 - 1}{35}$$

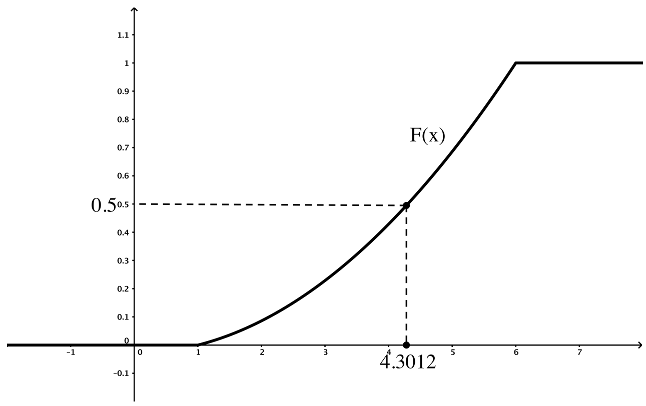
rendezés után:

$$med^2 = 0,5 \cdot 35 + 1 = 18,5$$

így a medián értéke

$$med = \sqrt{18,5} = 4,3012$$

A medián 1 és 6 közé kell essen, így a negatív gyök nem jöhet szóba megoldásként.



15. ábra. A 19. feladat eloszlásfüggvényének grafikonja, rajta a mediánnal

20. feladat A X valószínűségi változó eloszlásfüggvénye legyen $F(x) = \begin{cases} 1 - \frac{27}{x^3} & \text{ha } 3 < x \\ 0 & \text{különben} \end{cases}$.

a) $P(X < 5)$=?

b) $P(4 < X < 6)$=?

c) $P(4 < X)$=?

d) $E(X)$=?

Megoldás:

a) $P(X < 5) = F(5) = 1 - \frac{27}{5^3} = \frac{98}{125} = 0,7840.$

b) $P(4 < X < 6) = F(6) - F(4) = \left(1 - \dfrac{27}{6^3}\right) - \left(1 - \dfrac{27}{4^3}\right) = 0,2969.$

c) $P(4 < X) = 1 - P(X \leq 4) = 1 - F(4) = 1 - \left(1 - \dfrac{27}{4^3}\right) = \dfrac{27}{125} = 0,216.$

d) A várható értékhez a sűrűségfüggvényre van szükségünk, amit $F(x)$ deriválásával kapunk:

$$f(x) = \begin{cases} \dfrac{81}{x^4} & \text{ha } 3 < x \\ 0 & \text{különben} \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x = \int_3^{\infty} x \cdot \frac{81}{x^4}\mathrm{d}x = \int_3^{\infty} \frac{81}{x^3}\mathrm{d}x = \left[-\frac{81}{2x^2}\right]_3^{\infty} = 0 - \left(-\frac{81}{2 \cdot 3^2}\right) = 4,5.$$

21. feladat Az X valószínűségi változó sűrűségfüggvénye $f(x) = \begin{cases} \dfrac{3}{16}x^2 & \text{ha } -2 < x < 2 \\ 0 & \text{különben} \end{cases}$.

a) $P(X < 1)$=?

b) $P(0 < X)$=?

c) Ábrázoljuk $f(x)$-et!

d) $E(X)$=?

e) $D(X)$=?

f) Mi a valószínűsége annak, hogy az X valószínűségi változó a várható értékétől a szórásánál kevesebbel tér el?

Megoldás:

a) A folytonos valószínűségi változóhoz kapcsolódó valószínűségeket az eloszlásfüggvénnyel a legegyszerűbb kiszámítani. Most az eloszlásfüggvény nincs megadva. Ilyen esetekben nem érdemes az eloszlásfüggvényt kiszámolni a sűrűségfüggvényből, mivel a sűrűségfüggvény ugyancsak alkalmas a valószínűségek kiszámítására. (Az alapintervallumra figyeljünk, ennél a függvénynél ez a $(-2,2)$ intervallum!)

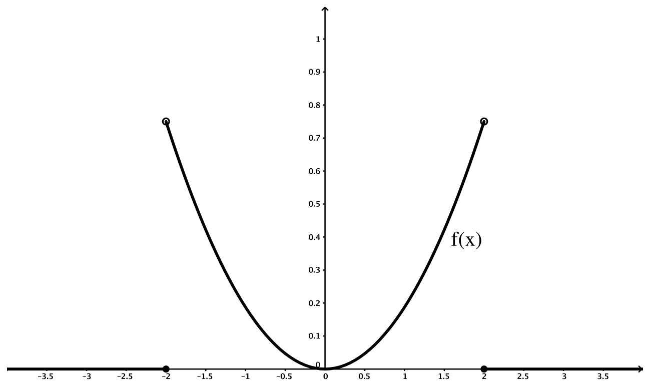
$$P(X < 1) = \int_{-\infty}^{1} f(x)\mathrm{d}x = \int_{-2}^{1} \frac{3}{16}x^2\mathrm{d}x = \left[\frac{x^3}{16}\right]_{-2}^{1} = \frac{1^3}{16} - \frac{(-2)^3}{16} = \frac{9}{16} = 0,5625$$

b)

$$P(0 < X) = \int_0^{\infty} f(x)\mathrm{d}x = \int_0^{2} \frac{3}{16}x^2\mathrm{d}x = \left[\frac{x^3}{16}\right]_0^{2} = \frac{2^3}{16} - \frac{0^3}{16} = \frac{8}{16} = 0,5$$

(A c) kérdésnél a sűrűségfüggvény ábrázolása után térjünk vissza ehhez a feladatrészhez. Vegyük észre, hogy ezt a valószínűséget a sűrűségfüggvény grafikonjáról számolás nélkül leolvashattuk volna, kihasználva a sűrűségfüggvény szimmetriáját.)

c)



16. ábra.

d)

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x = \int_{-2}^{2} \frac{3}{16}x^2 \cdot x\mathrm{d}x = \int_{-2}^{2} \frac{3}{16}x^3\mathrm{d}x = \left[\frac{3x^4}{64}\right]_{-2}^{2} = \frac{3 \cdot 2^4}{64} - \frac{3 \cdot (-2)^4}{64} = 0$$

Ha a sűrűségfüggvény grafikonjára pillantunk, látjuk, hogy $f(x)$ szimmetrikus függvény, $x = 0$ (y-tengely) a szimmetriatengelye. A várható érték is 0, sőt, ha a (b) rész eredményére tekintünk, azt látjuk, hogy a medián értéke is 0 (felezi a görbe alatti területet). Korábban már említettük, hogy szimmetrikus sűrűségfüggvények esetén mindig elmondható, hogy a szimmetriatengely megadja a várható értéket, amely megegyezik a mediánnal. Nemszimmetrikus sűrűségfüggvények esetén a medián és a várható érték nem kell, hogy megegyezzen.

e)

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\mathrm{d}x = \int_{-2}^{2} \frac{3}{16}x^2 \cdot x^2\mathrm{d}x = \int_{-2}^{2} \frac{3}{16}x^4\mathrm{d}x = \left[\frac{3x^5}{80}\right]_{-2}^{2} = \frac{3 \cdot 2^5}{80} - \frac{3 \cdot (-2)^5}{80} =$$

$$= \frac{192}{80} = 2,4$$

$$D(X) = \sqrt{E(X^2) - E^2(X)} = \sqrt{2,4 - 0^2} = \sqrt{2,4} = 1,5492$$

f)

$$P\big(\mid X - E(X) \mid < D(X)\big) = P\big(-D(X) < X - E(X) < D(X)\big) =$$

$$= P\big(E(X) - D(X) < X < E(X) + D(X)\big) =$$

$$= P(0 - 1,5492 < X < 0 + 1,5492) = P(-1,5492 < X < 1,5492) =$$

$$= \int_{-1,5492}^{1,5492} f(x)\mathrm{d}x = \int_{-1,5492}^{1,5492} \frac{3}{16}x^2\mathrm{d}x = \left[\frac{x^3}{16}\right]_{-1,5492}^{1,5492} = \frac{1,5492^3}{16} - \frac{(-1,5492)^3}{16} = 0,4648$$

3. Nevezetes eloszlások I.

A valószínűségszámítás klasszikus kiszámítási módjánál a számítási igény gyorsan megnő. Gondoljunk arra, ha egy mintavételnél nem 10-szer, hanem 100-szor húzunk, akkor a kihúzott hibás alkatrészek számának várható értéke 101 tagú összegre vezet. Ugyanakkor a megfigyelések során gyakran találkozunk olyan feladatokkal, amelyeket hasonló módon oldunk meg, csak más paraméterekkel. A következő fejezetben megnézzük, hogy bizonyos feladattípusoknál hogyan használhatjuk ki az eloszlásból adódó törvényszerűségeket.

3.1. Nevezetes diszkrét eloszlások

Diszkrét valószínűségi változó esetében az eloszlásra (x_i, p_i) van szükségünk a számoláshoz. Bizonyos gyakran használt eloszlásokat megkülönböztetünk, ezek esetében jelentősen tudunk egyszerűsíteni a számításokon.

3.1.1. Binomiális eloszlás

Egy megfigyelésnek gyakran a megfigyelési szempont szerint kétféle kimenetele fontos: 'sikeres', illetve 'sikertelen', azaz a várt kimenetelt kapjuk vagy az ellentettjét. Ha az érdekel bennünket, hogy n független megfigyelés során hányszor következett be a 'sikeres' kimenetel, akkor **binomiális eloszlás**sal modellezhetünk. Például:

- Egy játékban az nyer, aki 10 dobásból több hatost dob. Ekkor egy dobást a játékos akkor tekint sikeresnek, ha hatost dob, minden más érték sikertelen kimenetelnek számít a játékos szemszögéből.
- Egy termék gyártója a minőségellenőrzés során véletlenszerűen kiválasztott termékeket vizsgál meg. Természetesen minél kevesebb selejtet szeretne, ezért sikeres a kimenetel, ha a termék hibátlan, ellenkező esetben sikertelen.

1. feladat Egy ládában az alkatrészek 10%-a selejt. Visszatevéssel húzunk 4 alkatrészt. (Minden húzás után visszatesszük a kihúzottat.) Mekkora annak a valószínűsége, hogy pontosan 2 selejt van a kihúzottak között?

Megoldás: Az előző fejezetben láttunk hasonló példát. A visszatevéses mintavétel jellemzője, hogy az egyes húzások függetlenek egymástól, amit a valószínűségek kiszámításánál használhatunk ki. Jelölje S, hogy a kihúzott alkatrész selejt, $\overline{S}$ pedig, hogy nem selejt. A selejt húzás valószínűsége minden esetben $P(S) = 0,1$, a nem selejt húzás valószínűsége pedig $P(\overline{S}) = 0,9$. Két selejtet húzhatunk többféleképpen. Lehet az első két húzás selejt, ennek valószínűsége: $P(S\,S\,\overline{S}\,\overline{S}) = 0,1^2 \cdot 0,9^2$. Emellett azt, hogy a két selejt helye hányféle lehet a 4 húzás között, $\binom{4}{2}$ adja meg. Így

$$P(\text{pontosan 2 selejt}) = 0,1^2 \cdot 0,9^2 \cdot \binom{4}{2} = 0,0486.$$

Egyszerűsíthetünk a jelölésünkön, ha az X valószínűségi változót a kihúzott selejtek számának választjuk. (A kérdésből: a kihúzott selejtek számához kapcsolódott a kérdés, így azt választottunk valószínűségi változónak.) Ezzel a megoldás

$$P(X = 2) = 0,1^2 \cdot 0,9^2 \cdot \binom{4}{2} = 0,0486$$

alakba írható.

Általánosíthatjuk ezt a megoldást: Legyen p annak a valószínűsége, hogy selejtet húzunk a ládából (azaz a selejtarány), ami tulajdonképpen nem más, mint az A: 'selejtet húzunk' (egyetlen húzásból) esemény valószínűsége. Tehát $p = P(A)$. Ekkor

$$P(\mathrm{X}{=}2){=}p^2 \cdot (1-p)^2 \cdot \binom{4}{2}$$

Az is könnyen belátható, hogy a kihúzott selejtek száma (A bekövetkezési száma) lehet $0, 1, 2, 3, 4$, azaz X lehetséges értéke 0-tól a húzások számáig bármely egész szám lehet.

Definíció: Ha egy megfigyelés során n független kísérletet hajtunk végre és a sikeres kimenetel (vagyis az A esemény) bekövetkezésének valószínűsége minden kísérletnél $p = P(A)$, akkor a sikeres kimenetelek (A esemény bekövetkezéseinek) száma **binomiális eloszlású valószínűségi változó**, melyre

$$P(X = k) = p^k \cdot (1-p)^{n-k} \cdot \binom{n}{k}$$

ahol $k = 0, 1, \ldots, n$.

A binomiális eloszlású valószínűségi változót akkor tekintjük adottnak ha ismerjük a megfigyelések számát (n) és a sikeres kimenetel valószínűségét (p). Ezt a két értéket a binomiális eloszlás paramétereinek nevezzük. (Tehát a binomiális eloszlás két paraméteres eloszlás.)
Jelölés: X:Bin(n, p).

Az X binomiális eloszlású valószínűségi változó lehetséges értékei rendre $0, 1, \ldots, n$, és bármelyik lehetséges értékhez tartozó valószínűséget a megadott formulával számolhatjuk. Az eloszlást a korábban megszokott (bár itt hosszadalmas) módon is felírhatjuk:

$$X : \begin{cases} 0 & 1 & 2 & \ldots & n \\ (1-p)^n & p^1 \cdot (1-p)^{n-1} \cdot \binom{n}{1} & p^2 \cdot (1-p)^{n-2} \cdot \binom{n}{2} & \ldots & p^n \end{cases}$$

Diszkrét valószínűségi változó várható értékét és szórását az eloszlásból számoltuk ki. Ez a fentiek alapján binomiális eloszlásnál:

$$E(X) = \sum_{k=0}^{n} k \cdot p^k \cdot (1-p)^{n-k} \cdot \binom{n}{k}$$

amelyről megmutatható, hogy éppen np-vel egyenlő. A szórás hasonlóan egyszerű alakban írható fel, így az X:Bin(n, p) valószínűségi változó **várható értéke**

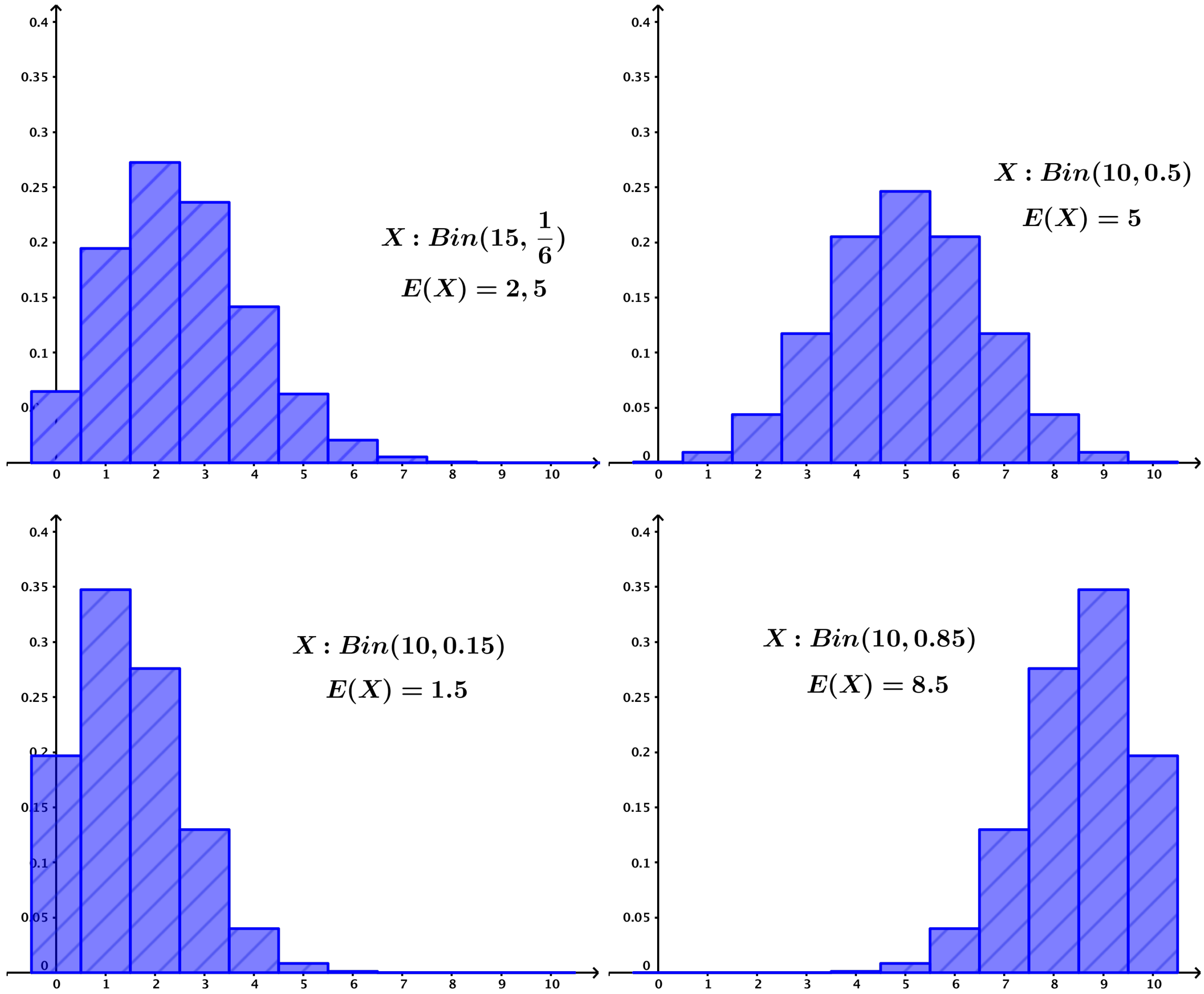
$$E(X) = np$$

és **szórása**

$$D(X) = \sqrt{np(1-p)}$$

A következő ábrákon különböző paraméterű, binomiális eloszlású valószínűségi változók eloszlását láthatjuk, a hisztogramokon jelölve a valószínűségi változó várható értékét is.

Ötlet: Nézzük meg a hisztogramok alakjának és a paramétereknek a kapcsolatát! Milyen kapcsolat van a hisztogram és a várható érték között?



17. ábra. Binomiális eloszlás különböző paraméterekkel

A hisztogramokon látszik, hogy ha a várható érték egész szám, akkor az $X = E(X)$ lehetséges értékhez tartozik a legnagyobb valószínűség, így ő a módusz is egyben, ha a várható érték nem egész, akkor valamelyik szomszédos lehetséges értékhez tartozik a legnagyobb valószínűség.

Azt is észrevehetjük, hogy ha a p valószínűséget $(1 - p)$-re cseréljük ugyanolyan n kísérletszám mellett, akkor a két hisztogram gyakorlatilag egymás 'tükörképe'. Könnyen ellenőrizhető a binomiális együttható definíciójából, hogy ha $X : Bin(n, p)$ és $Y : Bin(n, 1 - p)$, akkor

$$P(X = k) = p^k \cdot (1 - p)^{n-k} \cdot \binom{n}{k} = p^k \cdot (1 - p)^{n-k} \cdot \binom{n}{n-k} = P(Y = n - k)$$

Gondoljunk bele: egy 2% selejtet tartalmazó ládából 10-szer húzva ugyanakkora valószínűséggel húzunk pontosan 3 selejtet, mint pontosan 7 nem selejtet.

2. feladat Egy mobiltelefon szervizben a beadott készülékek 12%-nál kell kijelzőt cserélni. Hétfőn 15 telefont adnak be.

a) Mi a valószínűsége annak, hogy ebből 5 kijelzőcserés?

b) Mi a valószínűsége annak, hogy legalább 3-nál kell kijelzőt cserélni?

c) Mennyi a kijelzőcserés mobilok számának várható értéke?

d) Mi a valószínűsége annak, hogy a kijelzőcserés mobilok száma megegyezik a várható értékkel?

e) Határozzuk meg a kijelzőcserére szoruló mobilok számának szórását!

Megoldás: Feltételezzük, hogy a beadott mobilok hibái függetlenek egymástól. A szerviz 15 mobilt vizsgál meg (megfigyelések száma, $n = 15$), egy vizsgálat a megfigyelési szempontunk szerint akkor sikeres, ha kijelzőcserére szorul a telefon. (A megfigyelt A esemény: 'kijelzőcserére szorul a készülék', $p = P(A) = 0,12$.) A kérdések A esemény bekövetkezési számához kapcsolódnak, legyen tehát X: a kijelzőcserés mobilok száma. Ekkor X:Bin$(15; 0,12)$.

a)
$$P(X = 5) = 0,12^5 \cdot (1 - 0,12)^{15-5} \cdot \binom{15}{5} = 0,12^5 \cdot 0,88^{10} \cdot \binom{15}{5} = = 0,0208$$

b)
$$P(X \geq 3) = 1 - P(X < 3) = 1 - \big[P(X = 0) + P(X = 1) + P(X = 2)\big] =$$
$$= 1 - \left[0,12^0 \cdot 0,88^{15} \cdot \binom{15}{0} + 0,12^1 \cdot 0,88^{14} \cdot \binom{15}{1} + 0,12^2 \cdot 0,88^{13} \cdot \binom{15}{2}\right] = 0,2654$$

c)
$$E(X) = np = 15 \cdot 0,12 = 1,8$$

d)
$$P\big(X = E(X)\big) = P(X = 1,8) = 0$$

A kijelzőcserés mobilok száma csak egész lehet!

e)
$$D(X) = \sqrt{np(1 - p)} = \sqrt{15 \cdot 0,12 \cdot 0,88} = \sqrt{1,584} = 1,2586$$

3. feladat Létezik-e olyan binomiális eloszlás, melynek

a) várható értéke 10, szórása 2?

b) várható értéke 10, szórása 3?

Megoldás: Tudjuk, hogy egy binomiális eloszlású valószínűségi változó várható értéke $E(X) = np$, szórása $D(X) = \sqrt{np(1 - p)}$.

a) Ennek alapján meg kell oldanunk az
$$10 = np$$
és
$$2 = \sqrt{np(1 - p)}$$

egyenletekből álló egyenletrendszert. A második egyenletből:

$$4 = np(1-p)$$

Helyettesítsünk np helyére 10-et:

$$4 = 10 \cdot (1-p)$$

Innen:

$$1-p = \frac{4}{10} = 0,4$$
$$p = 1 - 0,4 = 0,6$$

Az első egyenletbe visszahelyettesítve $p = 0,6$-ot:

$$10 = n \cdot 0,6$$

$$n = \frac{10}{0,6} = 16,6667$$

Tudjuk, hogy a binomiális eloszlás n paramétere csak egész szám lehet, ezért nem létezik a feladatban megadott várható értékű és szórású binomiális eloszlású valószínűségi változó.

b) Hasonlóan számolunk, mint az előző részben: A megoldandó egyenletrendszer:

$$10 = np$$

és

$$3 = \sqrt{np(1-p)}$$

A második egyenletből:

$$9 = np(1-p)$$

np helyére 10-et helyettesítünk az első egyenletből:

$$9 = 10 \cdot (1-p)$$

Rendezve:

$$1-p = \frac{9}{10} = 0,9$$
$$p = 1 - 0,9 = 0,1$$

Az első egyenletbe visszahelyettesítve $p = 0,1$-ot:

$$10 = n \cdot 0,1$$

$$n = \frac{10}{0,1} = 100$$

Ez egész szám, így létezik a feltételeknek megfelelő binomiális eloszlás: $n = 100$ és $p = 0,1$ paraméterekkel.

Érdekes kérdés, hogy ha adott egy esemény valószínűsége, akkor hány kísérletet kell elvégeznünk a megfigyelésére ahhoz, hogy nagy valószínűséggel legalább egyszer bekövetkezzen az esemény? A következő példában megnézzük, hogy ennek a kiszámolása egy egyenlet megoldására vezet.

4. feladat Hányszor kell feldobnunk egy szabályos dobókockát ahhoz, hogy legalább 0,95 valószínűséggel legalább egy hatos legyen a dobások között?

Megoldás: X legyen a dobott hatosok száma (továbbra is a kérdésből célszerű meghatározni: aminek a számához kapcsolódik a kérdés, azt érdemes valószínűségi változónak választani). A kísérletek száma, n most ismeretlen, a hatos dobás (megfigyelt A esemény) valószínűsége $p = P(A) = \frac{1}{6}$ minden egyes dobásnál, a dobások eredménye független egymástól. X:Bin$(0, \frac{1}{6})$.

Ha hiányzik az eloszlás valamely paramétere, akkor általában egy valószínűséget várunk helyette, amiből kiszámolhatjuk a paramétert. Most adott:

$$P(X \geq 1) \geq 0,95$$

A binomiális eloszlásból az egyenlőtlenség bal oldala:

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \left(\frac{1}{6}\right)^0 \cdot \left(1 - \frac{1}{6}\right)^n \cdot \binom{n}{0} = 1 - \left(\frac{5}{6}\right)^n$$

A megoldandó egyenlőtlenség így:

$$1 - \left(\frac{5}{6}\right)^n \geq 0,95$$

$$0,05 \geq \left(\frac{5}{6}\right)^n$$

Vesszük mindkét oldal 10-es alapú logaritmusát és rendezzük az egyenlőtlenséget:

$$\lg 0,05 \geq n \cdot \lg\left(\frac{5}{6}\right)$$

$$-1,3010 \geq n \cdot (-0,0792)$$

$$n \geq 16,4268$$

A legkisebb egész szám, amely megfelel az eredményünknek, a 17. Ez azt jelenti, hogy legalább 17-szer fel kell dobnunk egy szabályos dobókockát, hogy legalább $0,95$ valószínűséggel legyen hatos a dobások között.

5. feladat Egy populációban az emberek 8%-a visel szemüveget. Véletlenszerűen kiválasztunk 20 embert. Mi a valószínűsége annak, hogy közülük

a) a várható értéknél többen viselnek szemüveget?

b) a szemüvegesek száma a várható értéktől a szórásnál kevesebbel tér el?

Megoldás: $n = 20$ ember megfigyelése során az A: 'szemüveges a kiválasztott' esemény bekövetkezési számához kapcsolódnak a kérdéseink. (A véletlenszerű kiválasztás miatt a függetlenséget feltételezhetjük.) Ezért legyen X a kiválasztott szemüveges emberek száma. Ez binomiális eloszlást követ, és tudjuk, hogy $p = P(A) = 0,08$. Ezért

$$E(X) = np = 20 \cdot 0,08 = 1,6$$

és

$$D(X) = \sqrt{np(1-p)} = \sqrt{20 \cdot 0,08 \cdot 0,92} = \sqrt{1,472} = 1,2133$$

a)

$$P(X > E(X)) = P(X > 1,6) = 1 - P(X \leq 1,6) = 1 - \Big[P(X = 0) + P(X = 1)\Big] =$$

$$= 1 - \left[0,08^0 \cdot 0,92^{20} \cdot \binom{20}{0} + 0,08^1 \cdot 0,92^{19} \cdot \binom{20}{1}\right] = 0,4831$$

b) Korábban láttunk hasonló kérdést a folytonos eloszlásnál:

$$P\big(\mid X - E(X) \mid < D(X)) = P(-D(X) < X - E(X) < D(X)\big) = P\big(E(X) - D(X) < X < E(X) + D(X)\big) =$$

$$= P(1,6 - 1,2133 < X < 1,6 + 1,2133) = P(0,3867 < X < 2,8133) = \Big[P(X = 1) + P(X = 2)\Big] =$$

$$= \left[0,08^1 \cdot 0,92^{19} \cdot \binom{20}{1} + 0,08^2 \cdot 0,92^{18} \cdot \binom{20}{2}\right] = 0,5993$$

Tehát $0,5993$ annak a valószínűsége, hogy a várható értéktől a szórásnál kevesebbel tér el a kiválasztottak között a szemüvegesek száma.

3.1.2. Geometriai eloszlás

Képzeljünk el egy olyan kísérletet, mikor addig dobáljuk a szabályos dobókockát, míg hatost nem dobunk. Továbbra is független kísérleteket hajtunk végre a dobásokkal, de most nem tudjuk előre, hogy hány kísérletre lesz szükségünk, csupán azt, hogy egy A esemény (hatos dobás) bekövetkezése állítja meg a dobássorozatot.

Definíció: A p valószínűségű A esemény megfigyelésére független kísérleteket hajtunk végre. Az A esemény első bekövetkezéséig végrehajtott kísérletek száma **geometriai eloszlású** valószínűségi változó, melynek eloszlása:

$$P(X = k) = (1 - p)^{k-1} \cdot p$$

ahol $k = 1, 2, 3, \ldots$.

A geometriai eloszlásnál a lehetséges értékek száma megszámlálhatóan végtelen (fel tudjuk sorolni a lehetséges értékeket, de számuk nem véges). Egy paraméteres eloszlás, elegendő p-t ismernünk.
Jelölés: X:Geo(p).

Az eloszlás felírható:

$$X : \begin{cases} 1 & 2 & 3 & 4 & \ldots \\ p & (1-p) \cdot p & (1-p)^2 \cdot p & (1-p)^3 \cdot p & \ldots \end{cases}$$

alakban is.

A várható érték:

$$E(X) = \sum_{k=1}^{\infty} k \cdot (1 - p)^{k-1} \cdot p$$

amelyről a gemetriai sor összegképletét felhasználva megmutatható, hogy értéke $\frac{1}{p}$. A **szórás** is hasonlóan egyszerű alakba írható. Ha $X : Geo(p)$, akkor **várható értéke**

$$E(X) = \frac{1}{p}$$

és **szórása**

$$D(X) = \sqrt{\frac{1 - p}{p^2}}$$

alakba írható.

6. feladat Egy izzó minden egyes bekapcsoláskor $0,004$ valószínűséggel kiég. Addig használjuk, míg kiég.

a) Mi a valószínűsége annak, hogy a századik bekapcsoláskor ég ki?

b) Mennyi a bekapcsolások számának várható értéke?

c) Mi a valószínűsége annak, hogy 150-nél többször használhatjuk?

d) Mi a valószínűsége annak, hogy a várható értéknél tovább használhatjuk?

e) Mekkora az a bekapcsolás szám, amelynél tovább $0,95$ valószínűséggel nem használhatjuk az izzót?

Megoldás: Az X valószínűségi változó a bekapcsolások száma (a kiégésig). Az A esemény tehát az, hogy az izzó kiég. $P(A) = p = 0,004$, így X:Geo($0,004$).

a) Felejtsük el az eloszlás formulát, gondolkodjunk 'józan paraszti ésszel'! Az, hogy a 100. bekapcsoláskor ég ki az izzó, pontosan azt jelenti, hogy az első 99 bekapcsoláskor nem ég ki ($\overline{A}$ következik be), és a 100-nál kiég. Így

$$P(X = 100) = 0,996^{99} \cdot 0,004 = 0,0027$$

(Természetesen ugyanezt az eredményt kapjuk a képletbe helyettesítéssel is.)

b) Itt már kihasználjuk a geometriai eloszlást:

$$E(X) = \frac{1}{p} = \frac{1}{0,004} = 250$$

A bekapcsolások számának várható értéke 250.

c) Elsőként az alábbi megoldás jut eszünkbe:

$$P(X > 150) = 1 - P(X \le 150) = 1 - \big[P(X = 1) + P(X = 2) + P(X = 3) + \ldots + P(X = 150)\big] =$$

$$= 1 - \big[0,996^0 \cdot 0,004 + 0,996^1 \cdot 0,004 + 0,996^2 \cdot 0,004 \ldots + 0,996^{149} \cdot 0,004\big] =$$

$$= 1 - \big[0,996^0 + 0,996^1 + 0,996^2 \ldots + 0,996^{149}\big] \cdot 0,004 = 1 - \frac{1 - 0,996^{150}}{1 - 0,996} \cdot 0,004 =$$

$$= 0,996^{150} = 0,5482$$

Kihasználtuk, hogy a zárójelben egy geometriai sorozat első 150 tagjának összege áll, melyre $a_0 = 1$ és $q = 0,996$. (Egy geometriai sorozat első n elemének összege pedig $a_0 \cdot \frac{1 - q^n}{1 - q}$.).
Ez a számolás azonban elég hosszadalmas volt, legalábbis az eredmény egyszerűségéhez képest. Gondoljuk át, mit jelent a kérdezett esemény: 150-nél többször pontosan akkor kapcsolhatjuk be az izzót, ha az első 150 bekapcsolás egyikénél sem ég ki. Ezt felhasználva a fenti számolás helyett azonnal adódik, hogy:

$$P(X > 150) = 0,996^{150} = 0,5482$$

d)

$$P\big(X > E(X)\big) = P(X > 250) = 0,996^{250} = 0,3671$$

Itt a végeredmény érdekes, ami tulajdonképpen azt jelenti,hogy a bekapcsolások száma a kiégésig, vagyis a használatok száma csak az izzók alig több, mint 36%-ánál haladja meg a várható értéket.

e) Megfordítjuk a kérdést: azt a bekapcsolási számot (k) keressük, amelynél tovább mindössze $0,05$ valószínűséggel használhatjuk az izzót. Formálisan:

$$0,05 = P(X > k), \quad k = ?$$

Az előzőekben láttuk, hogy

$$P(X > k) = (1 - p)^k$$

Így

$$0,05 = (1 - p)^k = (1 - 0,04)^k = 0,996^k$$

Mindkét oldal 10-es alapú logaritmusát véve:

$$\lg 0,05 = k \cdot \lg 0,996$$

$$k = \frac{\lg 0,05}{\lg 0,996} = 747,4342$$

Ami azt jelenti, hogy 747-nél többször $0,05$, legfeljebb 747-szer pedig $0,95$ valószínűséggel használhatjuk az izzót.

Általánosítva a c) feladatrészben követett gondolatmenetet, ha X:Geo(p), akkor

$$P(X > r) = (1-p)^r$$

7. feladat Egy villamoson $p = 0,05$ valószínűséggel jelennek meg ellenőrök és 8000 forintra büntetik a bliccelőket. Mi a valószínűsége annak, hogy a bírság összege fedezi a bliccelő által a lebukásig okozott kárt, ha a jegy ára 350 forint? (Az ellenőrök a büntetéskor a jegyet nem vetetik meg a bliccelővel.)

Megoldás: A bliccelő, minden egyes alkalommal, amikor nem büntetik meg, 350 forintos kárt okoz. A büntetés 8000 forintos összegéből 22 jegyre futja, 23-ra már nem. Vagyis ha a legkésőbb a 22. próbálkozásig megbüntetik a bliccelőt, akkor maximum $22 \cdot 350 = 7700$ forintos kárt okoz, ezt fedezi a bírság. Ha viszont legkorábban a 23. utazáskor büntetik meg, akkor legalább $23 \cdot 350 = 8050$ forintos kárt okoz, ezt már nem fedezi a bírság.
A kérdés tehát az, mi a valószínűség-e annak, hogy legfeljebb 22-szer bliccelhet a lebukásig.
Legyen X a bliccelések száma a lebukásig. A $P(X \leq 22)$ valószínűséget kétféleképp számolhatjuk.
Egyrészt a geometriai sorozat első n elemének összegével:

$$P(X \leq 22) = \big[P(X=1) + P(X=2) + P(X=3) + \ldots + P(X=22)\big] =$$

$$= \big[0,95^0 \cdot 0,05 + 0,95^1 \cdot 0,05 + 0,95^2 \cdot 0,05 \ldots + 0,95^{21} \cdot 0,05\big] = \big[0,95^0 + 0,95^1 + 0,95^2 \ldots + 0,95^{21}\big] \cdot 0,05 =$$

$$= \frac{1-0,95^{22}}{1-0,95} \cdot 0,05 = 1 - 0,95^{22} = 0,6765$$

Ne feledjük, hogy a zárójelben 22 tag áll, így az első 22 tag összegét vesszük a geometriai sorozatból.
Másrészt

$$P(X \leq 22) = 1 - P(X > 22) = 1 - 0,95^{22} = 0,6765$$

Itt visszavezettük a megoldást az ellentett esemény valószínűségének kiszámítására: $P(X > 22)$ annak a valószínűsége, hogy 22-nél többször próbálkozik a bliccelő a lebukás előtt, ami pontosan azt jelenti, hogy az első 22 alkalommal nem bukik le.

3.1.3. Poisson-eloszlás

Definíció: Ha az X valószínűségi változó eloszlása valamely $\lambda > 0$ paraméter mellett:

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

ahol $k = 0, 1, 2, \ldots$, akkor X **Poisson-eloszlású** valószínűségi változó, λ **paraméterrel**.

Jelölés: X:Poisson(λ).
(Numerikus sorok összegének felhasználásával ellenőrizhető, hogy az eloszlásban a valószínűségek összege 1.)

A **várható érték**

$$E(X) = \lambda$$

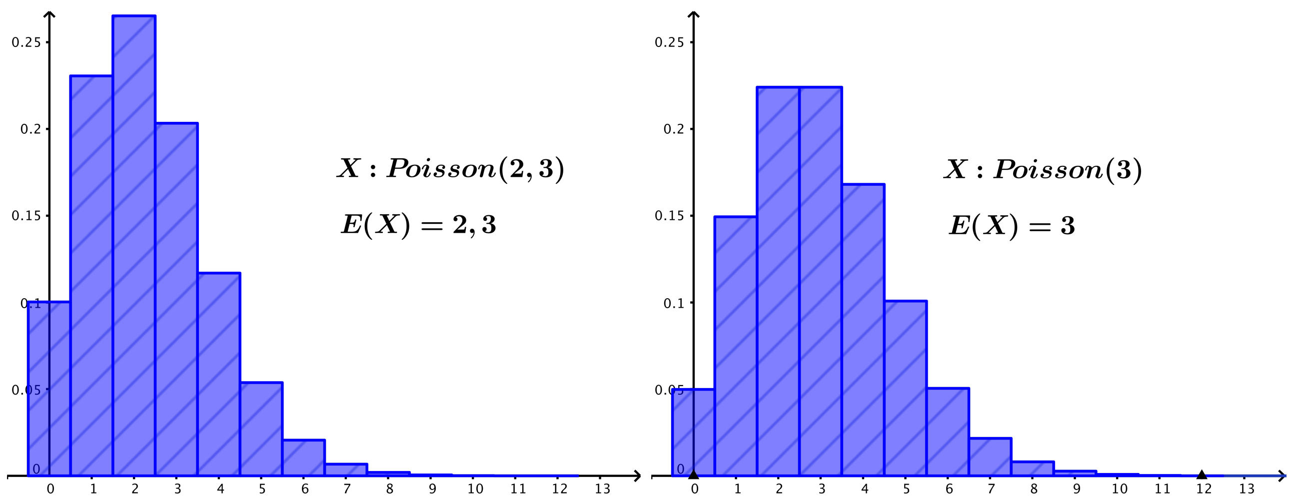
és a **szórás**

$$D(X) = \sqrt{\lambda}$$

alakban írható fel.

Eddig a nevezetes eloszlásokat bizonyos típusú kísérletekhez kötöttük. Felmerül a kérdés, milyen feladatok esetén használhatjuk a Poisson-eloszlást. Ha egy esemény bizonyos intervallumban (ami időbeli és térbeli intervallum is lehet) véletlenszerűen következik be és az intervallumban ismerjük az átlagos bekövetkezésének számát (λ), akkor az adott intervallumban az esemény bekövetkezéseinek száma λ paraméterű, Poisson-eloszlású valószínűségi változónak tekinthető.

A következő ábrákon különböző paraméterű Poisson-eloszlású valószínűségi változók eloszlását látjuk.



18. ábra. Poisson-eloszlás különböző paraméterekkel

A hisztogramokon is megfigyelhetjük, hogy a módusz a λ paraméterhez közeli egész szám. Pontosabban azt mondhatjuk, hogy ha X:Poisson(λ), akkor

- ha λ egész szám, akkor két módusz van, $\lambda - 1$ és λ,
- ha λ nem egész szám, akkor a módusz a λ egész része, azaz a hozzá legközelebbi, nála kisebb egész szám.

Példák a Poisson-eloszlásra:

- Vásárlók száma egy adott napon egy üzletben.
- Úthibák száma egy adott hosszúságú útszakaszon.
- Nyomdahibák száma egy bizonyos terjedelmű kiadványban.
- Ügyfélszolgálatra beérkező telefonhívások száma egy adott időtartam alatt.
- Közlekedési lámpa meghibásodásainak száma egy év alatt.
- Balesetek száma egy kereszteződésben adott idő alatt.

8. feladat Egy pékségben vásárolt túrós batyuban átlagosan 3 mazsola van. Mi a valószínűsége annak, hogy egy ott vásárolt túrós batyuban

a) pontosan két mazsola van?

b) van mazsola?

c) a mazsolák száma a várható értéknél kevesebb?

Megoldás: A mazsolák előfordulása egy batyuban véletlenszerű, az egy batyura eső mazsolák számának átlagos értéke ismert. Legyen az X valószínűségi változó a mazsolák száma egy batyuban, amely Poisson-eloszlásúnak tekinthető, $E(X) = \lambda = 3$ várható értékkel.

a)

$$P(X = 2) = \frac{3^2}{2!}e^{-3} = 0,2240$$

b)

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \frac{3^0}{0!}e^{-3} = 0,9502$$

c)

$$P\big(X < E(X)\big) = P(X < 3) = \big[P(X = 0) + P(X = 1) + P(X = 2)\big] = \frac{3^0}{0!}e^{-3} + \frac{3^1}{1!}e^{-3} + \frac{3^2}{2!}e^{-3} =$$

$$= \big[1 + 3 + \frac{3^2}{2}\big]e^{-3} = 0,4232$$

Érdemes megfigyelni az ábrákon, hogy a Poisson-eloszlás nem szimmetrikus eloszlás (egyetlen paraméterrel sem), emiatt a fenti valószínűségre nem is várhattunk 0,5-et.

9. feladat Egy városban hetente átlagosan 5 tűzeset fordul elő. Mi a valószínűsége annak, hogy

a) 4 hét alatt pontosan 18 tűzeset történik?

b) 1 nap legalább 2 tűzeset történik?

Megoldás: Legyen X az egy hét alatt előforduló tűzesetek száma, ekkor X Poisson-eloszlást követ, 5 várható értékkel (azaz $\lambda = 5$).

a) Jelölje X^* a négy hét alatt bekövetkező tűzesetek számát. Ha egy hét alatt átlagosan 5 tűzeset történik a városban, akkor 4 hét alatt átlagosan $4 \cdot 5 = 20$, így X^* is Poisson-eloszlású valószínűségi változónak tekinthető, $\lambda^* = 20$ paraméterrel.

$$P(X^* = 18) = \frac{20^{18}}{18!} e^{-20} = 0,0844$$

b) Jelölje X^{**} az egy nap alatt bekövetkező tűzesetek számát, ennek átlagos értéke $\frac{5}{7}$, így X^{**} is Poisson-eloszlású valószínűségi változónak tekinthető, $\lambda^{**} = \frac{5}{7}$ paraméterrel.

$$P(X^{**} \geq 2) = 1 - \left[P(X=0) + P(X=1)\right] = 1 - \left[\frac{\left(\frac{5}{7}\right)^0}{0!} e^{-\left(\frac{5}{7}\right)} + \frac{\left(\frac{5}{7}\right)^1}{1!} e^{-\left(\frac{5}{7}\right)}\right] = 0,1608$$

10. feladat Egy nem túl forgalmas benzinkúthoz óránként átlagosan $5,3$ autó érkezik.

a) Mekkora annak a valószínűsége, hogy egy óra alatt pontosan három autó érkezik a kúthoz?

b) Mekkora annak a valószínűsége, hogy fél óra alatt érkezik autó a kúthoz?

c) Mekkora az az időtartam, amely alatt 0,95 valószínűséggel érkezik autó a benzinkúthoz?

Megoldás:

a) A benzinkúthoz óránként érkező autók száma legyen X. X Poisson-eloszlást követ, $E(X) = 5,3$ várható értékkel, így $\lambda = 5,3$.

$$P(X = 3) = \frac{5,3^3}{3!} e^{-5,3} = 0,1239$$

b) X^* legyen a félóránként érkező autók száma, $\lambda^* = \frac{5,3}{2} = 2,65$.

$$P(X^* \geq 1) = 1 - P(X^* < 1) = 1 - P(X^* = 0) = 1 - \frac{2,65^0}{0!} e^{-2,65} = 0,9293$$

c) Jelölje x a keresett időtartamot. $\hat{X}$ legyen az x idő alatt érkező autók száma, ennek $\hat{\lambda}$ paraméterét nem ismerjük. A feladatban megadott valószínűséget felírva:

$$0,95 = P(\hat{X} \geq 1) = 1 - P\big(\hat{X} < 1\big) = 1 - P\big(\hat{X} = 0)\big) = 1 - \frac{\hat{\lambda}^0}{0!} e^{-\hat{\lambda}} = 1 - e^{-\hat{\lambda}}$$

Rendezés után:

$$e^{-\hat{\lambda}} = 1 - 0,95 = 0,05$$

Mindkét oldal e-alapú logaritmusát véve:

$$-\hat{\lambda} = \ln 0,05 = -2,9957$$

$$\hat{\lambda} = 2,9957$$

A kérdés viszont x-re vonatkozott, amiről annyit tudunk, hogy $\hat{\lambda}$-ra igaz, hogy úgy aránylik az $5,3$-hoz, ahogy az x idő az egy órához:

$$\frac{x}{1} = \frac{\hat{\lambda}}{5,3}$$

$$x = \frac{\hat{\lambda}}{5,3} = \frac{2,9957}{5.3} = 0,5652$$

Mivel eddig órában számoltunk, az eredményt is órában kaptuk meg. Percre átszámolva ≈ 34 perc az az időtartam, ami alatt kb. $0,95$ valószínűséggel érkezik autó a benzinkúthoz.

11. feladat Egy focicsapat meccenként lőtt góljainak számát az alábbi táblázat tartalmazza::

lőtt gólok száma	0	1	2	3	4	5	6	7
meccsek száma	14	18	29	18	10	7	3	1

Határozzuk meg az egy mérkőzésre jutó gólok számának átlagértékét, majd a Poisson-eloszlás segítségével számoljuk ki, hogy milyen gyakoriságokat várunk az eloszlásból!

Megoldás:
Elsőként vegyük észre a következőt: a véges mintánkat akarjuk modellezni egy olyan valószínűségi változóval, **amelynek végtelen sok lehetséges értéke van** (tudjuk, hogy egy Poisson-eloszlású valószínűségi változó tetszőleges nemnegatív egész értéket felvehet).
Felmerül a kérdés, hogy ezt megtehetjük-e? A focicsapatunk hétnél több gólt a megfigyelt mérkőzések egyikén sem lőtt, holott ez elméletben lehetséges. Ha Poisson-eloszlásúnak tekintjük a lőtt gólok számát, akkor viszont az eloszlásból a hétnél több lőtt gól valószínűsége nem nulla lesz. Mit tehetünk ilyen esetben? Megnézzük, hogy a Poisson-eloszlásból mekkora valószínűséget kapunk a 7-nél több gólra. Ha **ez a valószínűség kellően kicsi (majdnem 0)**, akkor elhanyagolhatónak tekintjük. Ekkor modellezhetünk a végtelen eloszlással.

A lőtt gólok számát tekintjük az X valószínűségi változónak. Azt feltételezzük, hogy X Poisson-eloszlású valószínűségi változó, melynek λ paraméterét a mintából számolt átlaggal becsüljük. A feladat szerint az X várható értékének az egy mérkőzésre eső gólok számának átlagát tekintjük, ami az összesen lőtt gólok számának és az összes megfigyelt mérkőzés számának hányadosa:

$$E(X) = \frac{0 \cdot 14 + 1 \cdot 18 + 2 \cdot 29 + \ldots + 6 \cdot 3 + \cdot 1}{14 + 18 + 29 + \ldots + 3 + 1} = \frac{230}{100} = 2,3$$

Nézzük meg, a $\lambda = 2,3$ paraméterű Poisson-eloszlásból mekkora valószínűségeket kapunk a lőtt gólok számának lehetséges értékeihez.

$$p_0 = P(X=0) = \frac{2.3^0}{0!} \cdot e^{-2.3} = 0,1003$$
$$p_1 = P(X=1) = \frac{2.3^1}{1!} \cdot e^{-2.3} = 0,2306$$
$$p_2 = P(X=2) = \frac{2.3^2}{2!} \cdot e^{-2.3} = 0,2652$$
$$p_3 = P(X=3) = \frac{2.3^3}{3!} \cdot e^{-2.3} = 0,2033$$
$$p_4 = P(X=4) = \frac{2.3^4}{4!} \cdot e^{-2.3} = 0,1169$$
$$p_5 = P(X=5) = \frac{2.3^5}{5!} \cdot e^{-2.3} = 0,0538$$
$$p_6 = P(X=6) = \frac{2.3^6}{6!} \cdot e^{-2.3} = 0,0206$$
$$p_7 = P(X=7) = \frac{2.3^7}{7!} \cdot e^{-2.3} = 0,0068$$
$$p_8 = P(X > 7) = 1 - P(X \leq 7) =$$
$$= 1 - (P(X=0) + P(X=1) + \ldots + P(X=7)) = 0,0025$$

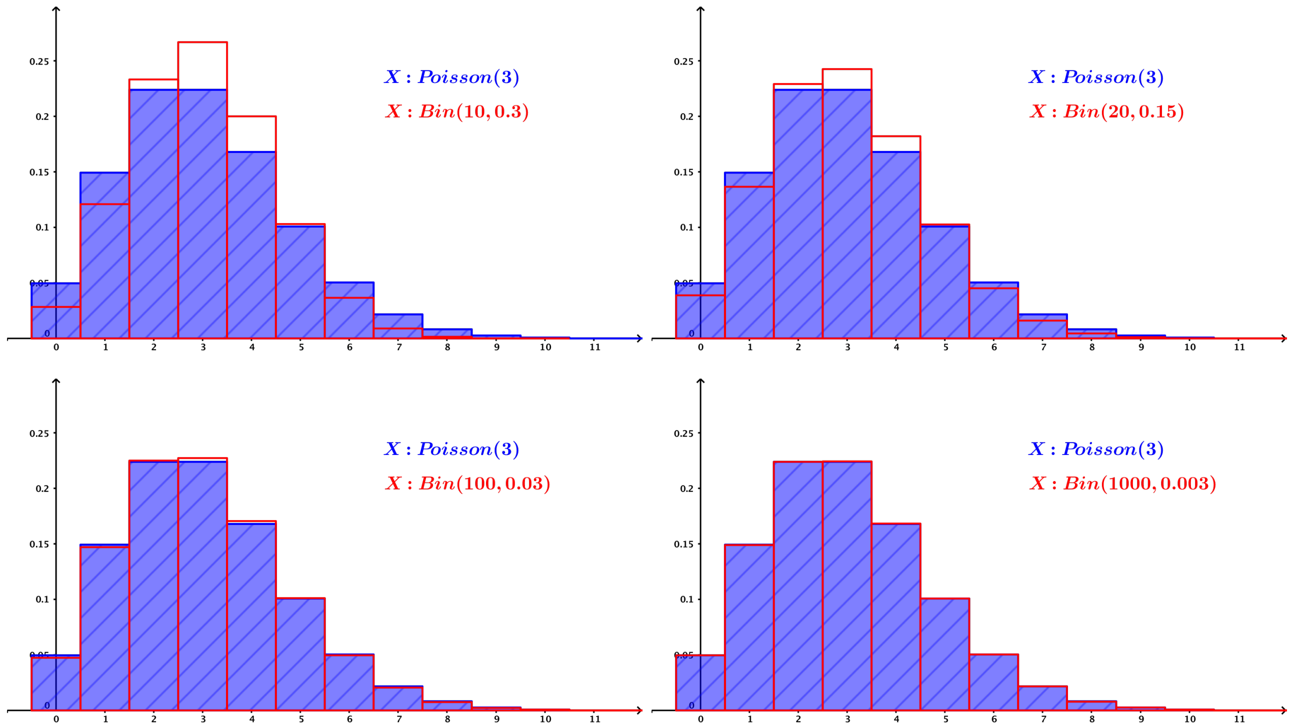
Az elvárt gyakoriságok (a kapott valószínűségeket a mérkőzések számával, százzal szorozzuk):

lőtt gólok száma	0	1	2	3	4	5	6	7
megfigyelt gyakoriság	14	18	29	18	10	7	3	1
elvárt gyakoriság	10,03	23,06	26,52	20,33	11,69	5,38	2,06	0,68

Látjuk, hogy a megfigyelt és az elvárt gyakoriságok között kisebb-nagyobb eltérések vannak. Hogy ezeket az eltéréseket hogyan értelmezhetjük, azzal a χ^2-próba illeszkedésvizsgálatánál fogalkozunk.

3.1.4. Kapcsolat a Poisson- és a binomiális eloszlás között

Ha megnézzük a binomiális és a Poisson-eloszláshoz tartozó hisztogramokat, akkor felfedezhetünk némi hasonlóságot közöttük. Az alábbi ábrákon azonos várható értékű Poisson- és binomiális eloszlású valószínűségi változók eloszlása látható.



19. ábra. Azonos várható értékű binomiális és Poisson-eloszlások

Az ábrákon látszik, hogy egy binomiális eloszlású valószínűségi változót a vele azonos várható értékű Poisson-eloszlású valószínűségi változó annál jobban közelít, minél nagyobb a binomiális eloszlás n paramétere. Ennél kicsit többet is mondhatunk:

Állítás: Legyen X binomiális eloszlású valószínűségi változó, n és p paraméterekkel. Ha n elég nagy és p (vagy $1-p$) elég kicsi, akkor X-et közelíthetjük $\lambda = np$ paraméterű Poisson-eloszlással, azaz:

$$P(X = k) = p^k \cdot (1-p)^{n-k} \cdot \binom{n}{k} \approx \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{ahol} \quad \lambda = np$$

12. feladat Egy irodaházban 1200 számítógép üzemel. Közülük mindegyik, egymástól függetlenül 0,003 valószínűséggel hibásodik meg naponta. (A hibás gépet csak a nap végén javítják.) Mi a valószínűsége annak, hogy hétfőn

a) pontosan két számítógép hibásodik meg?

b) nem történik meghibásodás?

c) legalább 4 számítógép meghibásodik?

Megoldás: Legyen X a hétfőn (vagy tetszőleges napon) a meghibásodások (A esemény bekövetkezéseinek) száma, $n = 1200$ kísérlet esetén. A meghibásodások egymástól függetlenül következnek be, így X binomiális eloszlást követ, $p = P(A) = 0,003$ és $n = 1200$ paraméterekkel.

a)

$$P(X=2) = 0,003^2 \cdot 0,997^{1198} \cdot \binom{1200}{2} = 0,1770$$

Ugyanakkor n elég nagy, p elég kicsi, így X közelíthető $\lambda = np = 1200 \cdot 0,003 = 3,6$ paraméterű Poisson-eloszlással:

$$P(X=2) \approx \frac{3,6^2}{2!} e^{-3,6} = 0,1771$$

Látjuk, hogy a Poisson-eloszlással közelített értéket lényegesen egyszerűbb kiszámolni, a közelítés pontossága pedig elég jónak tekinthető.

b)

$$P(X=0) = 0,003^0 \cdot 0,997^{1200} \cdot \binom{1200}{0} = 0,0272$$

Poisson-eloszlással közelítve:

$$P(X=0) \approx \frac{3,6^0}{0!} e^{-3,6} = 0,0273$$

c) A pontos megoldás:

$$P(X \geq 4) = 1 - P(X < 4) = 1 - \big[P(X=0) + P(X=1) + P(X=2) + P(X=3)\big] =$$

$$= 1 - \Big[0,003^0 \cdot 0,997^{1200} \cdot \binom{1200}{0} + 0,003^1 \cdot 0,997^{1199} \cdot \binom{1200}{1} + 0,003^2 \cdot 0,997^{1198} \cdot \binom{1200}{2} +$$

$$+ 0,003^3 \cdot 0,997^{1197} \cdot \binom{1200}{3}\Big] = 0,4850$$

Poisson-eloszlással közelítve:

$$P(X \geq 4) = 1 - P(X < 4) = 1 - \big[P(X=0) + P(X=1) + P(X=2) + P(X=3)\big] \approx$$

$$\approx 1 - \left[\frac{3,6^0}{0!} + \frac{3,6^1}{1!} + \frac{3,6^2}{2!} + \frac{3,6^3}{3!}\right] \cdot e^{-3,6} = 0,4848$$

4. Nevezetes eloszlások II.

4.1. Nevezetes folytonos eloszlások

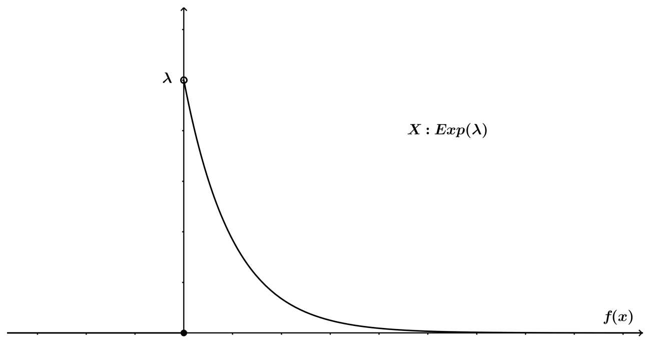
Folytonos eloszlású valószínűségi változót akkor tekintünk adottnak, ha vagy az eloszlás-, vagy a sűrűségfüggvényét ismerjük. Korábban láttuk, hogy az eloszlásfüggvénnyel egyszerűbb valószínűségeket kiszámolni, a sűrűségfüggvény viszont szükséges a várható érték és a szórás kiszámításához, valamint a grafikonja szemléletesebb. A következő nevezetes eloszlásoknál kihasználjuk az adott sűrűségfüggvény jellemzőit, hogy a várható értéket és a szórást egyszerűen, az eloszlásra jellemző paraméterekkel írjuk fel, így ezeknél az eloszlásoknál a számításokhoz elegendő lesz az eloszlásfüggvényt felhasználni.

4.1.1. Exponenciális eloszlás

Definíció: Legyen λ teszőleges pozitív szám ($\lambda > 0$). Ha az X valószínűségi változó **sűrűségfüggvény**e

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{ha } 0 < x \\ 0 & \text{különben} \end{cases}$$

akkor X **exponenciális eloszlású** valószínűségi változó, λ **paraméterrel**.
Jelölés: X:Exp(λ).



20. ábra. Exponenciális eloszlású valószínűségi változó sűrűségfüggvénye

(Ellenőrizzük, hogy az sűrűségfüggvény görbe alatti területe valóban 1!)

Az X:Exp(λ) **eloszlásfüggvény**e:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{ha } 0 < x \\ 0 & \text{különben.} \end{cases}$$

A várható érték és a szórás a folytonos eloszlású valószínűségi változónál tanultak szerint számolható:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x = \int_{0}^{\infty} x \cdot \lambda e^{-\lambda x}\mathrm{d}x = \ldots = \frac{1}{\lambda}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\mathrm{d}x = \int_{0}^{\infty} x^2 \cdot \lambda e^{-\lambda x}\mathrm{d}x = \ldots = \frac{2}{\lambda^2}$$

$$D(X) = \sqrt{E(X^2) - E^2(X)} = \sqrt{\frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2} = \frac{1}{\lambda}$$

Így a **várható érték**:

$$E(X) = \frac{1}{\lambda}$$

A **szórás**:

$$D(X) = \frac{1}{\lambda}$$

Megjegyzés: Tehát az exponenciális eloszlású valószínűségi változó várható értéke és szórása megegyezik.

Exponenciális eloszlású valószínűségi változó lehet jellemzően egy műszer élettartama, várakozási idő stb..

1. feladat Egy LED-es izzó élettartama exponenciális eloszlás követ, 25 000 óra várható élettartammal. Mi a valószínűsége annak, hogy egy ilyen izzó élettartama

a) kevesebb, mint 20 000 óra?

b) 20 000 és 25 000 óra közé esik?

c) legalább a várható érték?

d) több, mint 30 000 óra, ha már 20 000 órája működik?

Megoldás: Az X valószínűségi változó legyen az izzó élettartama (órában), $E(X) = 25\,000$. A paraméter a várható értékből számolható. Tudjuk, hogy

$$E(X) = \frac{1}{\lambda}$$

így

$$\lambda = \frac{1}{E(X)}$$

tehát

$$\lambda = \frac{1}{25\,000} = 0,00004$$

Az eloszlásfüggvény ennek segítségével:

$$F(x) = \begin{cases} 1 - e^{-0,00004x} & \text{ha } 0 < x \\ 0 & \text{különben} \end{cases}$$

a) A keresett valószínűség az eloszlásfüggvény értéke 20 000-nél:

$$P(X < 20\,000) = F(20\,000) = 1 - e^{-0,00004 \cdot 20\,000} = 1 - e^{-0,8} = 0,5507$$

b)

$$P(20\,000 < X < 25\,000) = F(25\,000) - F(20\,000) = \left(1 - e^{-0,00004 \cdot 25\,000}\right) - \left(1 - e^{-0,00004 \cdot 20\,000}\right) =$$

$$= \left(1 - e^{-1}\right) - \left(1 - e^{-0,8}\right) = 0,6321 - 0,5507 = 0,0814$$

c)

$$P(X \geq E(X)) = P(X \geq 25\,000) = 1 - P(X < 25\,000) = 1 - F(25\,000) =$$

$$= 1 - (1 - e^{-0,00004 \cdot 20\,000}) = 1 - (1 - e^{-1}) = 0,3679$$

d) A feltételes valószínűségnél tanult definíciót kihasználva:

$$P(X > 30\,000 \mid X > 20\,000) = \frac{P(\{X > 30\,000\} \cap \{X > 20\,000\})}{P(X > 20\,000)} = \frac{P(X > 30\,000)}{P(X > 20\,000)} =$$

$$= \frac{1 - F(30\,000)}{1 - F(20\,000)} = \frac{1 - (1 - e^{-0{,}00004 \cdot 30\,000})}{1 - (1 - e^{-000004 \cdot 20\,000})} = \frac{e^{-0{,}00004 \cdot 30\,000}}{e^{-0{,}00004 \cdot 20\,000}} = e^{-0{,}00004 \cdot (30\,000 - 20\,000)} =$$

$$= e^{-0{,}00004 \cdot 10\,000} = 0,6703$$

Vegyük észre, hogy

$$e^{-0{,}00004 \cdot 10\,000} = 1 - (1 - e^{-0{,}00004 \cdot 10\,000}) = 1 - F(10\,000) = P(X > 10\,000)$$

Az exponenciális eloszlás legfontosabb tulajdonsága az "**örökifjú**", vagy "**emlékezet nélküli**" tulajdonság:

$$P(X > a + b \mid X > a) = P(X > b)$$

minden pozitív a és b esetén.

Az előző példában ez azt jelenti, ha egy izzó már 20 000 órája működik, akkor annak valószínűsége, hogy további 10 000 órán túl (azaz összesen több, mint 30 000 óráig) működik, nem függ az előtte eltelt idő hosszától. A számítás is ezt mutatta.

2. feladat Egy orvosi műszer élettartama exponenciális eloszlást követ. A gyártó a tervezett elavulás jegyében úgy tervezte meg a műszert, hogy csak kb. 15%-uk élettartama érje el a 10 évet.

a) Mekkora a műszer várható élettartama?

b) Mi a valószínűsége annak, egy ilyen műszer két éven belül tönkremegy?

c) Feltéve, hogy egy ilyen műszer az első 5 éven belül nem hibásodik meg, mi a valószínűsége, hogy a következő 3 éven belül elromlik?

Megoldás: Az X valószínűségi változó a műszer élettartama évben számolva. Sem az $E(X)$ várható érték, sem a λ paraméter nem ismert, így a konkrét eloszlásfüggvény sem ismert. A hiányzó paraméter helyett viszont megadott a feladat egy valószínűséget:

$$0,15 = P(X \geq 10)$$

ebből

$$0,85 = P(X < 10)$$

Használjuk ki, hogy X exponenciális eloszlást követ:

$$0,85 = P(X < 10) = F(10) = 1 - e^{-\lambda \cdot 10}$$

Rendezés után:

$$e^{-\lambda \cdot 10} = 1 - 0,85 = 0,15$$

Mindkét oldal e-alapú logaritmusát vesszük:

$$-\lambda \cdot 10 = ln(0,15)$$

$$\lambda = -\frac{ln(0,15)}{10} = 0,1897$$

Miután kiszámoltuk a paraméter értékét, az eloszlásfüggvény:

$$F(x) = \begin{cases} 1 - e^{-0,1897 \cdot x} & \text{ha } 0 < x \\ 0 & \text{különben} \end{cases}$$

a)
$$E(X) = \frac{1}{\lambda} = \frac{1}{0,1897} = 5,2715$$

A várható élettartam tehát 5,2715 év (kb. 5 és egy negyed év).

b)
$$P(X < 2) = F(2) = 1 - e^{-0,1897 \cdot 2} = 1 - e^{-0,3794} = 0,3157$$

c) Használjuk ki az örökifjú tulajdonságot, és vegyük észre, hogy most az ellentettel kell számolnunk:

$$P(X < 5 + 3 \mid X > 5) = P(X < 3) = F(3) = 1 - e^{-0,1897 \cdot 3} = 1 - e^{-0,5691} = 0,4340$$

Kapcsolat az exponenciális és a Poisson-eloszlás között: Az exponenciális eloszlás paraméterére ugyanazt a jelölést használtuk, mint a Poisson-eloszlás paraméterére. Vajon véletlen ez, vagy találhatunk kapcsolatot a két eloszlás között? Azt mondhatjuk, hogy Poisson-eloszlásban az események bekövetkezése között eltelt időt (a következő esemény bekövetkezéséig eltelt időt) exponenciális eloszlásúnak tekinthetjük. A következő példán nézzük, mit jelent ez:

3. feladat Egy augusztusi éjszakán a hullócsillagokat figyeljük a domboldalon. Tapasztalatok szerint óránkét átlagosan 2 hullócsillagot lehet látni. Mi a valószínűsége annak, hogy fél óra alatt egyet sem látunk?

I. Megoldás: Legyen az X valószínűségi változó a fél óra alatt észlelt hullócsillagok száma, melynek átlagos értéke 1 (egy óra alatt kettő az átlag, akkor fél óra alatt ennek a fele, azaz 1). X-et Poisson-eloszlásúnak tekinthetjük,

$$\lambda = \frac{1}{E(X)} = \frac{1}{1} = 1$$

paraméterrel. Így

$$P(X = 0) = \frac{1^0}{0!} e^{-1} = e^{-1} = 0,3679$$

II. Megoldás: Felhasználva, hogy két esemény bekövetkezése között eltelt idő exponenciális eloszlású, legyen $\hat{X}$ a következő hullócsillag észleléséig eltel idő. Átlagos értéke fél óra, azaz $E(\hat{X}) = 0,5$ (ennyi idő telik el két csillaghullás között átlagosan). A $\hat{\lambda}$ paraméter így

$$\hat{\lambda} = \frac{1}{E(\hat{X})} = \frac{1}{0,5} = 2$$

az eloszlásfüggvény:

$$F(x) = \begin{cases} 1 - e^{-2x} & \text{ha } 0 < x \\ 0 & \text{különben} \end{cases}$$

A kérdés ezzel a valószínűségi változóval úgy fogalmazható meg, hogy mi a valószínűsége annak, hogy fél óránál többet kell várni a hullócsillag észlelésére. A keresett valószínűség:

$$P(\hat{X} > 0,5) = 1 - P(\hat{X} < 0,5) = 1 - F(0,5) = 1 - (1 - e^{-2 \cdot 0,5}) = 1 - (1 - e^{-1}) = e^{-1} = 0,3679$$

Látható, hogy a két eredmény pontosan megegyezik.

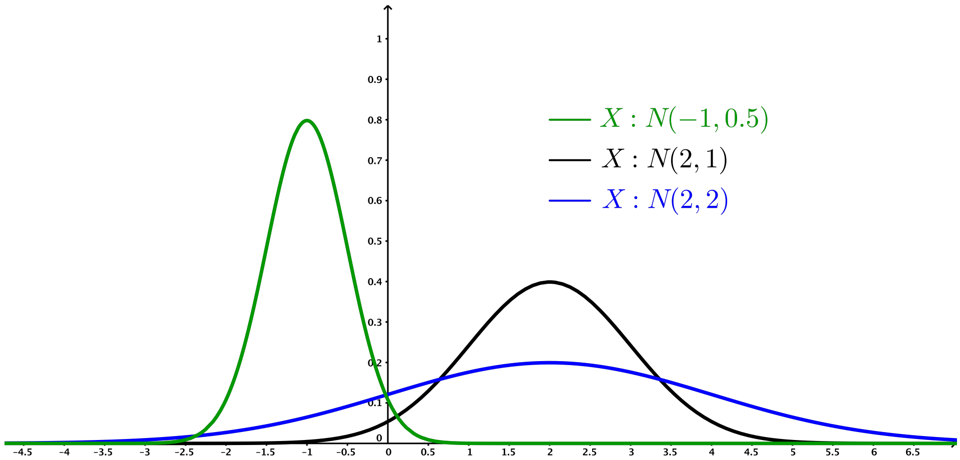
4.1.2. Normális eloszlás

Az egyik legfontosabb és leggyakrabban használt eloszlás a statisztikában a normális eloszlás.

Definíció: Legyenek m tetszőleges, σ pozitív valós számok. Ha az X valószínűségi változó **sűrűségfüggvénye**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

akkor X **normális eloszlású** valószínűségi változó m **és** σ **paraméterekkel**.
Jelölés: X:N(m,σ).



21. ábra. Normális eloszlású valószínűségi változó sűrűségfüggvénye különböző paraméterekkel

Az exponenciális eloszláshoz hasonlóan az eloszlásfüggvényt szeretnénk használni, de itt sajnos $f(x)$ elemi módszerekkel nem integrálható, így az **eloszlásfüggvény**t csak az alábbi (számolásra nem igazán alkalmas, integrálfüggvényes) alakban tudjuk felírni:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} \mathrm{d}t$$

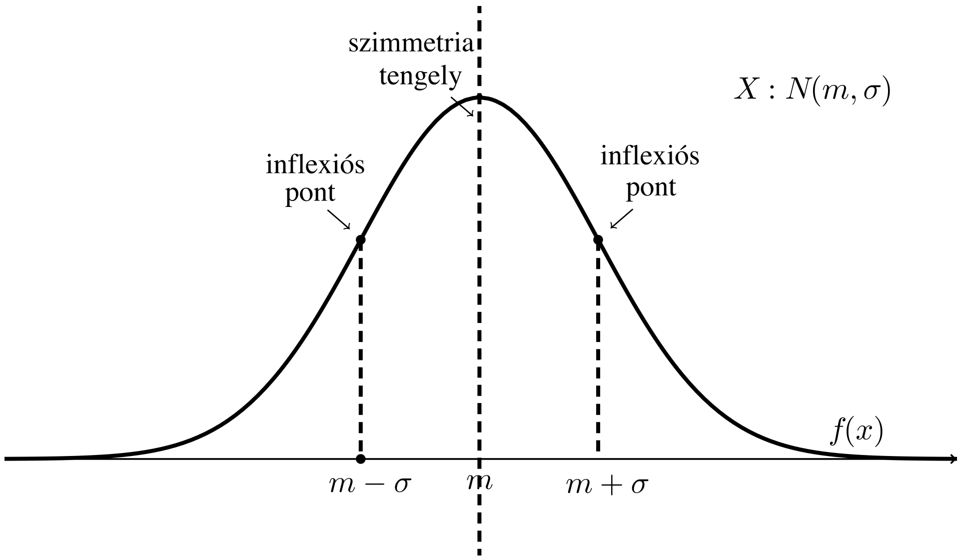
A **várható érték** és a **szórás**:

$$E(X) = m$$

$$D(X) = \sigma$$

Az X:N(m,σ) valószínűségi változó sűrűségfüggvényének speciális tulajdonságai:

- szimmetrikus az $x = m$ egyenesre (szimmetriatengelye a várható értéknél van),
- maximumát az $x = m$ helyen veszi fel, a maximum értéke $f(m) = \dfrac{1}{\sigma\sqrt{2\pi}}$,
- inflexiós pontjai az $x = m - \sigma$ illetve $x = m + \sigma$ helyeken vannak.



22. ábra. Normális eloszlású valószínűségi változó sűrűségfüggvényének jellemzői

Ahhoz, hogy valószínűségeket tudjunk számolni, szükség van egy speciális, normális eloszlású valószínűségi változó megkülönböztetésére, ez pedig a **standard normális eloszlású** valószínűségi változó.

Definíció: Ha X:N(m, σ), akkor az

$$X^* = \frac{X-m}{\sigma}$$

valószínűségi változó **standard normális eloszlású**, 0 várható értékkel és 1 szórással, azaz X^*:N$(0,1)$.

Az X^*:N$(0,1)$ valószínűségi változó **sűrűségfüggvény**ét $\varphi(x)$-el jelöljük:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}}$$

eloszlásfüggvényét pedig $\Phi(x)$-el:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t)^2}{2}} \mathrm{d}t$$

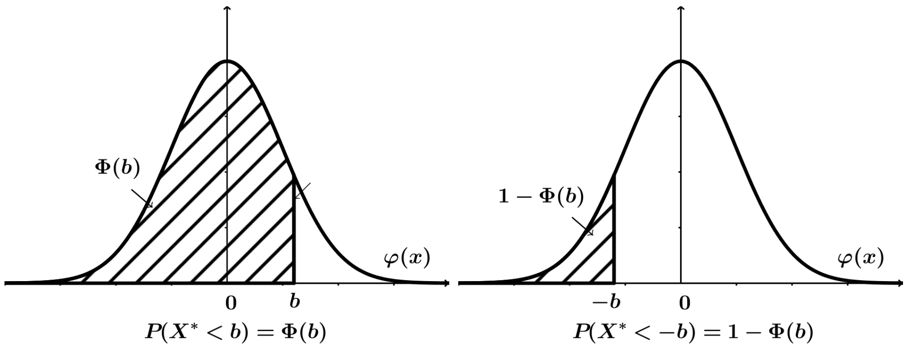
A standard normális eloszlású valószínűségi változó sűrűségfüggvényére teljesülnek, hogy

- az y-tengelyre szimmetrikus,
- inflexiós pontjai az $x = -1$ és $x = 1$ helyeken vannak,
- maximum helye az $x = 0$ hely, maximum értéke $\varphi(0) = \frac{1}{\sqrt{2\pi}}$.

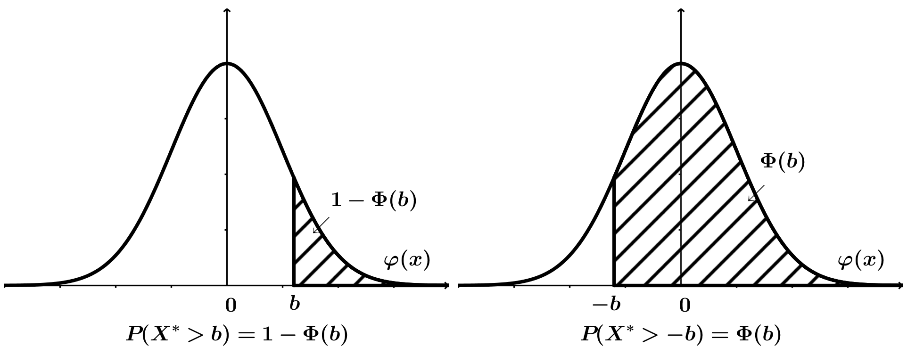
Bár a standard normális eloszlású valószínűségi változó sűrűségfüggvénye egyszerűbbnek tűnik, az integrálása még mindig nagyon nehéz, az eloszlásfüggvényt így most is csak integrálfüggvényes alakban tudtuk felírni. Az integrálási nehézségeket egy táblázat segít áthidalni. Tudjuk, hogy

$$P(X^* < b) = \Phi(b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-\frac{(x)^2}{2}} \mathrm{d}x$$

(A $\Phi(b)$ értéke a $\varphi(x)$ sűrűségfüggvény grafikonján az b-től balra eső görbe alatti területet jelenti.) Ennek az integrálnak az értékét pedig táblázatból kereshetjük ki, bármely nemnegatív b esetén (természetesen bizonyos lépésközönként). Megjegyezzük, hogy a táblázat a $\Phi(x)$ eloszlásfüggvény értékeit tartalmazza, ugyanakkor a valószínűségeket ábrázolni a $\varphi(x)$ sűrűségfüggvény grafikonján fogjuk. Az alábbi ábrákon láthatjuk, hogy a különböző módon felírt valószínűségeket hogyan vezethetjük vissza a $\Phi(x)$ eloszlásfüggvény értékeire:



23. ábra.



24. ábra.

A táblázat megadásakor kihasználjuk a standard normális eloszlás sűrűségfüggvényének szimmetriáját, illetve azt, hogy a sűrűségfüggvény görbe alatti területe 1.

Érdemes megjegyezni, hogy a fentiek alapján **előjel korrekció**t használhatunk:

$$\Phi(-b) = 1 - \Phi(b)$$

A 8.1 *A standard normális eloszlás eloszlásfüggvényének értékei* című táblázatot a jegyzet végén találjuk, használatát nézzük most meg.

4. feladat Legyen X^* standard normális eloszlású valószínűségi változó. ($X^* : N(0,1)$.)

a) $P(X^* < 1,96)$=?

b) $P(X^* < -1,96)$=?

c) $P(X^* > 2)$=?

d) $P(X^* > -1)$=?

e) $P(-1,96 < X^* < 1,96)$=?

f) $P(-0,5 < X^* < 1)$=?

g) $P(0,5 < X^* < 1,5)$=?

h) $P(-1,5 < X^* < -0,5)$=?

Megoldás:

a)
$$P(X^* < 1,96) = \Phi(1,96) = 0,9750$$

A standard normális eloszláshoz tartozó 8.1 táblázatban az első oszlopban megkeressük a változó értékét az első tizedesjegyig $(1,9)$, majd az első sorban megkeressük a második tizedesjegyet $(0,06)$. A sor és oszlop metszetében találjuk az eloszlásfüggvény értékét: $0,9750$. Hasonlóan járunk el a továbbiakban is.

b)
$$P(X^* < -1,96) = 1 - \Phi(1,96) = 1 - 0,975 = 0,025$$

c)
$$P(X^* > 2) = 1 - \Phi(2) = 1 - 0,9772 = 0,0228$$

d)
$$P(X^* > -1) = \Phi(1) = 0,8413$$

e)
$$P(-1,96 < X^* < 1,96) = \Phi(1,96) - \Phi(-1,96) = \Phi(1,96) - \big[1 - \Phi(1,96)\big] = 2 \cdot \Phi(1,96) - 1 =$$
$$= 2 \cdot 0,9750 - 1 = 0,9500$$

f)
$$P(-0,5 < X^* < 1) = \Phi(1) - \Phi(-0,5) = \Phi(1) - \big[1 - \Phi(0,5)\big] = \Phi(1) + \Phi(0,5) - 1 =$$
$$= 0,8413 + 0,6915 - 1 = 0,5328$$

g)
$$P(0,5 < X^* < 1,5) = \Phi(1,5) - \Phi(0,5) = 0,9332 - 0,6915 = 0,2417$$

h)
$$P(-1,5 < X^* < -0,5) = \Phi(-0,5) - \Phi(-1,5) = \big[1 - \Phi(0,5)\big] - \big[1 - \Phi(1,5)\big] =$$
$$= (1 - 0,6915) - (1 - 0,9332) = 0,2417$$

Gyakran nem arra vagyunk kíváncsiak, hogy az eloszlásfüggvény milyen értéket vesz fel egy adott helyen, hanem arra, hogy hol vesz fel egy értéket. Ehhez gyakorlatilag az eloszlásfüggvény inverzét kell használnunk. (Az exponenciális eloszlásnál is láttunk hasonló példát, a 2. feladatban.) A táblázat most is megkönnyíti a dolgunkat:

5. feladat Legyen X^* standard normális eloszlású valószínűségi változó. $(X^* : N(0,1))$.

a) $P(X^* < b) = 0,9$, b=?

b) $P(X^* > a) = 0,21$, a=?

c) $P(X^* > a) = 0,73$, a=?

d) $P(X^* < b) = 0,2$, b=?

e) $P(a < X^* < b)$=0,9?

f) $P(-a < X^* < a)$=0,9?

Megoldás:

a) $P(X^* < b) = \Phi(b) = 0,9$, a 8.1 táblázatban most az eloszlásfüggény értékei között keressük a $0,9$-et (tehát a táblázat 'belsejében'). A pontos érték nincs a táblázatban, ilyenkor a hozzá legközelebbit keressük. (Szokás még ilyenkor interpolációt alkalmazni, de ennél célravezetőbb szükség esetén olyan táblázatot keresni, amely több tizedesjegyet tartalmaz.) A $0,9$-hez legközelebbi érték a táblázatban a 0,8997, ezt választjuk. A hozzá tartozó sorból $1,2$, az oszlopból a $0,08$ érték olvasható le, így

$$b = 1,28$$

b) A továbbiakban felhasználjuk a korábbi összefüggéseket.

$$P(X^* > a) = 1 - P(X^* < a) = 1 - \Phi(a) = 0,21$$

rendezés után

$$\Phi(a) = 1 - 0,21 = 0,79$$

a táblázatból

$$a = 0,81$$

c) Mivel a valószínűség

$$P(X^* > a) = 1 - P(X^* < a) = 1 - \Phi(a) = 0,73$$

rendezés után

$$\Phi(a) = 1 - 0,73 = 0,27$$

Ez az érték azonban nincs a táblázatban. Miért? Azért, mert $0,5$-nél kisebb, tehát a értéke negatív szám. Mit tehetünk ekkor? Használjuk az előjel korrekciót:

$$\Phi(a) = 1 - \Phi(-a) = 0,27$$

ebből

$$\Phi(-a) = 0,73$$

$$-a = 0,61$$

$$a = -0,61$$

A számolás jelentősen lerövidíthető, ha észrevesszük, hogy $0,5$-nél nagyobb eloszlásfüggvény érték $P(X^* > a)$ esetben csak negatív a esetén lehetséges.

d) Vegyük észre, hogy tartozik $P(X^* < b)$ értéke most kisebb, mint $0,5$, így b negatív szám. Ezzel

$$P(X^* < b) = \Phi(b) = 1 - \Phi(-b) = 0,2$$

$$\Phi(-b) = 1 - 0,2 = 0,8$$

$$\Phi(-b) = 0,8$$

$$-b = 0,84$$

$$b = -0,84$$

e) Ennek a kérdésnek nincs egyértelmű megoldása. (Végtelen sok megoldás van, de ezeket csak 'találgatással' tudjuk meghatározni, lévén a megoldás egy kétismeretlenes lineáris egyenletrendszerre vezet.)

f) Az előző kérdés speciális esete. A különbség az, hogy csak egy ismeretlen van, de ennél sokkal fontosabb, hogy ennek eredményeként a várható értékre (0-ra) szimmetrikus az intervallum. Ekkor

$$P(-a < X^* < a) = \Phi(a) - \Phi(-a) = \Phi(a) - \big[1 - \Phi(a)\big] = 2 \cdot \Phi(a) - 1 = 0,9$$

Rendezés után:

$$\Phi(a) = \frac{0,9 + 1}{2} = 0,95$$

Innen

$$a = 1,65$$

Figyelembe véve, hogy az

$$X^* = \frac{X - m}{\sigma}$$

valószínűségi változó tetszőleges $X : N(m, \sigma)$ esetén standard normális eloszlású és

$$P(X < b) = P\Big(\frac{X - m}{\sigma} < \frac{b - m}{\sigma}\Big) = P\Big(X^* < \frac{b - m}{\sigma}\Big)$$

teljesül, egy transzformáció után bármely normális eloszlás esetén visszavezethetjük a számításokat a standard normális eloszlásra.

6. feladat Legyen X normális eloszlású valószínűségi változó, 5 várható értékkel 4 szórással.

a) $P(X < 7)$=?

b) $P(X < 2)$=?

c) $P(X > 6)$=?

d) $P(X > 1)$=?

e) $P(5 < X < 8)$=?

Megoldás:

$$m = 5 \quad \sigma = 4 \quad X : N(5, 4)$$

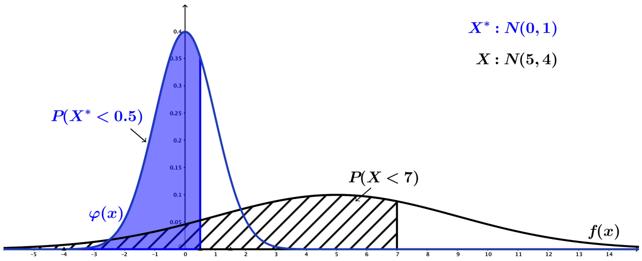
a) Ahhoz, hogy számolni tudjunk, első lépésben standardizáljuk (transzformáljuk) az X-et oly módon, hogy kivonjuk belőle m-et és osztjuk σ-val. A transzformációt az egyenlőtlenség mindkét oldalán elvégezzük:

$$P(X < 7) = P\Big(\frac{X-m}{\sigma} < \frac{7-5}{4}\Big) = P(X^* < 0,5)$$

Innen a már látott módon számolhatunk tovább:

$$P(X^* < 0,5) = \Phi(0,5) = 0,6915$$

Az alábbi ábrán láthatjuk az eredeti, illetve a transzformált valószínűségi változóhoz tartozó sűrűségfüggvényen a keresett valószínűségeket. (Azért, hogy a két sűrűségfüggvény ugyanabban a koordinátarendszerben jól látható legyen, az x és y tengely aránya lényegesen eltér, de ez nem befolyásolja, hogy a megfelelő területek nagysága azonos.)



25. ábra.

b)

$$P(X < 2) = P\Big(\frac{X-m}{\sigma} < \frac{2-5}{4}\Big) = P(X^* < -0,75) = \Phi(-0,75) =$$
$$= 1 - \Phi(0,75) = 1 - 0,7734 = 0,2266$$

c) A továbbiakban rövidítsünk annyit a jelölésünkön, hogy $\dfrac{X-m}{\sigma} = X^*$-ot csak X^* alakban írjuk le a transzformáció után.

$$P(X > 6) = P\Big(X^* > \frac{6-5}{4}\Big) = P(X^* > 0,25) = \Big(1 - P(X^* < 0,25)\Big) = 1 - \Phi(0,25) =$$
$$= 1 - 0,5987 = 0,4013$$

d)

$$P(X > 1) = P\Big(X^* > \frac{1-5}{4}\Big) = P(X^* > -1) = 1 - P(X^* < -1) = \Phi(1) = 0,8413$$

e)

$$P(5 < X < 8) = P\Big(\frac{5-5}{4} < X^* < \frac{8-5}{4}\Big) = P(0 < X^* < 0,75) = 0,7734 - 0,5 = 0,2734$$

7. feladat Legyen X normális eloszlású valószínűségi változó, 20 várható értékkel 3 szórással.

a) $P(X < b) = 0,54$, b=?

b) $P(X > a) = 0,22$, a=?

c) $P(m - c < X < m + c) = 0,98$, c=?

Megoldás: $m = 20$, $\sigma = 3$. $X : N(20, 3)$.

a)
$$P(X < b) = P\Big(\frac{X - m}{\sigma} < \frac{b - 20}{3}\Big) = P\Big(X^* < \frac{b - 20}{3}\Big)$$

Tudjuk, hogy
$$P\Big(X^* < \frac{b - 20}{3}\Big) = \Phi\Big(\frac{b - 20}{3}\Big) = 0,54$$

A táblázatból (figyeljünk arra, hogy most az eloszlásfüggvény értéke van megadva!):
$$\frac{b - 20}{3} = 0,1$$
$$b = 0,1 \cdot 3 + 20 = 20,3$$

b) Hasonlóan járunk el, mint az előző feladatrészben:
$$P(X > a) = P\Big(\frac{X - m}{\sigma} > \frac{a - 20}{3}\Big) = P\Big(X^* > \frac{a - 20}{3}\Big)$$

Ez a valószínűség most:
$$P\Big(X^* > \frac{a - 20}{3}\Big) = 0,22$$
$$1 - \Phi\Big(\frac{a - 20}{3}\Big) = 0,22$$
$$\Phi\Big(\frac{a - 20}{3}\Big) = 1 - 0,22 = 0,78$$
$$\frac{a - 20}{3} = 0,77$$
$$a = 0,77 \cdot 3 + 20 = 22,31$$

c) A keresett intervallum a várható értékre szimmetrikus, így egyértelmű megoldást várunk. Használjuk ki, hogy $m = 20$:
$$P(m - c < X < m + c) = P(20 - c < X < 20 + c) =$$
$$= P\Big(\frac{(20 - c) - 20}{3} < X^* < \frac{(20 + c) - 20}{3}\Big) =$$
$$= P\Big(\frac{-c}{3} < X^* < \frac{c}{3}\Big)$$

Így
$$P\Big(\frac{-c}{3} < X^* < \frac{c}{3}\Big) = 0,98$$
$$\Phi\Big(\frac{c}{3}\Big) - \Phi\Big(\frac{-c}{3}\Big) = 0,98$$

c pozitív szám, $-c$ ekkor negatív, alkalmazható rá az előjel korrekció:
$$\Phi\Big(\frac{c}{3}\Big) - \Big[1 - \Phi\Big(\frac{c}{3}\Big)\Big] = 0,98$$
$$2\Phi\Big(\frac{c}{3}\Big) - 1 = 0,98$$

Rendezzük az egyenletet:
$$\Phi\Big(\frac{c}{3}\Big) = \frac{0,98 + 1}{2}$$
$$\Phi\Big(\frac{c}{3}\Big) = 0,99$$

A táblázatból:
$$\frac{c}{3} = 2,33$$
$$c = 3 \cdot 2,33 = 6,99$$

A kapott eredmény azt jelenti, hogy egy $X : N(20,3)$ valószínűségi változó értéke a várható értékre szimmetrikus

$$(20-6,99; 20+6,99)$$

tehát a

$$(13,01; 26,99)$$

intervallumba nagy, $0,98$ valószínűséggel esik.

Alkalmazzuk szöveges feladatokra a tanultakat!

8. feladat Egy populációban az IQ normális eloszlásúnak tekinthető, 100 pont várható értékkel és 15 pont szórással. A populáció hány százalékának IQ-ja

a) magasabb 120 pontnál?

b) alacsonyabb 80 pontnál?

c) esik 90 és 130 pont közé?

Megoldás: A megoldás kulcsa, hogy ismerjük a normális eloszlás paramétereit és tudjuk felírni a megadott valószínűségi változóval a kérdéseket. A feladat megadja, hogy az X valószínűségi változó az IQ értéke, normális eloszlásúnak tekinthető, $m = 100$, $\sigma = 15$ paraméterekkel.

a) A kérdést felírjuk X-szel, majd az előző feladatokban látott módon számolunk:

$$P(X > 120) = 1 - P(X < 120) = 1 - P\Big(X^* < \frac{120-100}{15}\Big) = 1 - P\Big(X^* < \frac{4}{3}\Big) = 1 - P(X^* < 1,33) =$$

$$= 1 - \Phi(1,33) = 1 - 0,9082 = 0,0918$$

Tehát a populáció $9,18\%$-a rendelkezik 120 pontnál magasabb IQ-val.

b) A $P(X < 80)$ valószínűséget keressük. Ha szemfülesek vagyunk, akkor kihasználhatjuk a sűrűségfüggvény szimmetriáját. X várható értéke ugyanis 100, erre szimmetrikus a $(80,120)$ intervallum, ezért $P(X < 80) = P(X > 120)$. Az előző feladat eredményét felhasználva: $P(X < 80) = 0,0918$. Ha nem vesszük észre a szimmetriát, akkor a hagyományos módon is számolhatuk, kihasználva az előjel korrekciót:

$$P(X < 80) = P\Big(X^* < \frac{80-100}{15}\Big) = P(X^* < -1,3333) = \Phi(-1,33) = 1 - \Phi(1,33) =$$

$$= 1 - 0,9082 = 0,0918$$

c)

$$P(90 < X < 130) = P\Big(\frac{90-100}{15} < X^* < \frac{130-100}{15}\Big) = P(-0,6667 < X^* < 2) =$$

$$= \Phi(2) - \Phi(-0,67) = \Phi(2) - \Big(1 - \Phi(0,67)\Big) = 0,9772 - 1 + 0,7486 = 0,7258$$

9. feladat Egy alkatrész hossza normális eloszlást követ, 100 mm várható értékkel és $0,8$ mm szórással. Selejtnek tekintünk egy alkatrészt, ha hossza a várható értéktől legalább 1,5 mm-rel eltér.

a) Az alkatrészek hány százaléka hosszabb 98 mm-nél?

b) Hány százalék selejt?

c) Mekkora az a hossz, amelynél az alkatrészek mindössze 1%-a rövidebb?

Megoldás: Legyen X az alkatrész hossza, tudjuk, hogy $m = 100$, $\sigma = 0,8$, így $X : N(100; 0,8)$.
A kérdéseket írjuk fel X segítségével!

a)

$$P(X > 98) = P\Big(X^* > \frac{98-100}{0,8}\Big) = P(X^* > -2,5) = 1 - P(X^* < -2,5) = 1 - \Phi(-2,5) =$$

$$= 1 - \big(1 - \Phi(2,5)\big) = \Phi(2,5) = 0,9938$$

Tehát az alkatrészek $99,38\%$-a hosszabb, mint $98\,\mathrm{mm}$.

b) Selejt helyett először a nem selejtek arányát számoljuk ki. Nem selejt egy alkatrész, ha hossza a $(100-1,5; 100+1,5)$, azaz a $(98,5; 101,5)$ intervallumba esik. Ennek valószínűsége:

$$P(98,5 < X < 101,5) = P\Big(\frac{98,5-100}{0,8} < X^* < \frac{101,5-100}{0,8}\Big) = P(-1,8750 < X^* < 1,8750) =$$

$$= \Phi(1,875) - \Phi(-1,875) = \Phi(1,875) - \big(1 - \Phi(1,875)\big) = 2\cdot\Phi(1,875) - 1 = 2\cdot 0,9699 - 1 = 0,9398$$

A selejtarány tehát $1 - 0,9398 = 0,0602$, azaz $6,02\%$.

c) A keresett hosszt jelölje b. Tudjuk, hogy

$$0,01 = P(X < b)$$

Keressük b-t!

$$P(X < b) = P\Big(X^* < \frac{b-100}{0,8}\Big) = \Phi\Big(\frac{b-100}{0,8}\Big) = 0,01$$

Ez azt jelenti, hogy $\dfrac{b-100}{0,8}$ negatív szám ($0,01$ nincs benne a táblázatban), tehát előjel korrekció szükséges:

$$\Phi\Big(\frac{b-100}{0,8}\Big) = 1 - \Phi\Big(\frac{-(b-100)}{0,8}\Big) = 0,01$$

rendezés után

$$\Phi\Big(\frac{-b+100}{0,8}\Big) = 1 - 0,01 = 0,99$$

a táblázatból

$$\frac{-b+100}{0,8} = 2,33$$

így

$$b = 100 - 2,33\cdot 0,8 = 98,136$$

Tehát $98,136\,\mathrm{mm}$-nél az alkatrészek 1%-a rövidebb.

Mit tehetünk, ha az eloszlás valamelyik paramétere hiányzik?

10. feladat Egy termék tömege normális eloszlást követ, 70 g várható értékkel. Tudjuk, hogy a termékek 15%-ának tömege legalább 84 g. Határozzuk meg a szórást!

Megoldás: Legyen X a termék tömege. Amit tudunk: $X : N(70, \sigma)$, illetve $P(X \geq 84) = 0,15$. Érdemes az ellentettel számolni: $P(X < 84) = 0,85$. Innen

$$P(X < 84) = P\Big(X^* < \frac{84-70}{\sigma}\Big) = P\Big(X^* < \frac{14}{\sigma}\Big) = \Phi\Big(\frac{14}{\sigma}\Big) = 0,85$$

A táblázatból

$$\frac{14}{\sigma} = 1,04$$

$$\sigma = \frac{14}{1,04} = 15,56$$

A szórás tehát: $\sigma = 15,56$.

Megjegyzés: A feladatok megoldása során többször találkoztunk olyan esettel, mikor a standardizálás után 0-ra szimmetrikus intervallumot kaptunk. (Figyeljük meg, hogy ilyenkor az eredeti valószínűségi változó is a várható értékére szimmetrikus intervallumba esett.) Az alábbi formulával standard normális eloszlásnál rövidíthetünk a számoláson:

$$P(-a < X^* < a) = 2\Phi(a) - 1$$

Használjuk fel ezt a következő valószínűségek kiszámításához!

12. feladat Legyen $X : N(m, \sigma)$. Mekkora annak a valószínűsége, hogy X értéke a várható értékétől kevesebb, mint

a) egy szórással,

b) két szórással,

c) három szórással

tér el?

Megoldás:

a)

$$P(m-\sigma < X < m+\sigma) = P\Big(\frac{(m-\sigma)-m}{\sigma} < X^* < \frac{(m+\sigma)-m}{\sigma}\Big) = P(-1 < X^* < 1) =$$
$$= 2\Phi(1) - 1 = 2 \cdot 0,8413 - 1 = 0,6826$$

b)

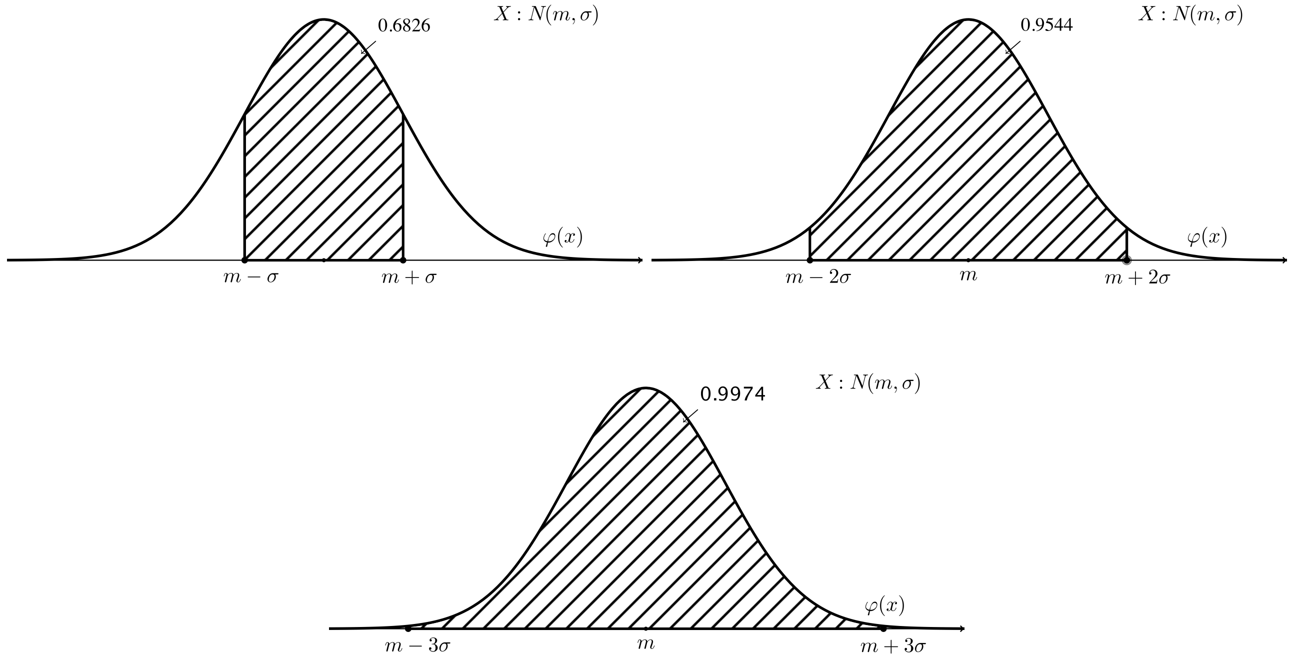
$$P(m-2\sigma < X < m+2\sigma) = P\Big(\frac{(m-2\sigma)-m}{\sigma} < X^* < \frac{(m+2\sigma)-m}{\sigma}\Big) = P(-2 < X^* < 2) =$$
$$= 2\Phi(2) - 1 = 2 \cdot 0,9772 - 1 = 0,9544$$

c)

$$P(m-3\sigma < X < m+3\sigma) = P\Big(\frac{(m-3\sigma)-m}{\sigma} < X^* < \frac{(m+3\sigma)-m}{\sigma}\Big) = P(-3 < X^* < 3) = 2\Phi(3)-1 =$$

$$= 2 \cdot 0,9987 - 1 = 0,9974$$

A sűrűségfüggvény grafikonján a kapott eredmények:



26. ábra.

Az eredményünk minden normális eloszlású valószínűségi változóra igaz, paramétertől függetlenül. Azt is érdemes megjegyezni, hogy a várható értéktől való eltérés és a szórás aránya a transzformáció után is ugyanannyi marad.

5. Mintaelemek átlagának statisztikai viselkedése

Az előző fejezetekben megismertünk néhány nevezetes eloszlást, megnéztük, hogyan számolhatjuk valószínűségi változók várható értékét, szórását (akár általános, akár nevezetes esetben). Egy sokaság megfigyelése során azonban gyakran valószínűségi változók összegét, átlagát vizsgáljuk. Adódnak az alábbi kérdések:

- Mit mondhatunk valószínűségi változók összegének és átlagának eloszlásáról?
- 'Örökli-e' az összeg és az átlag az eredetileg megfigyelt valószínűségi változók eloszlását?
- 'Örökli-e' az összeg és az átlag az eredeti valószínűségi változók paramétereit, várható értékét, szórását?

Általánosságban nem a válasz ezekre a kérdésekre, de bizonyos esetekben az eredeti eloszlás várható értékéből és szórásából tudjuk származtatni az összeg és az átlag várható értékét és szórását. Sőt, néhány feltétel teljesülése esetén az összeg és az átlag eloszlására is tudunk következtetni.

5.1. Valószínűségi változók összegének várható értéke és szórása

Elsőként nézzük, hogy valószínűségi változók összegének várható értékét és szórását hogyan számolhatjuk ki az eredeti várható értékek és szórások segítségével.

Tétel: Ha az $X_1, X_2, \ldots, X_n$ **azonos eloszlású valószínűségi változók**nak létezik $E(X_1)$, $E(X_2)$, $\ldots$, $E(X_n)$ várható értéke, akkor összegüknek is létezik, és az **összegük várható értéke megegyezik a várható értékek összegével**. Azaz

$$E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$$

A szórásnégyzetekre hasonló összefüggés igaz, de itt már megköveteljük, hogy a valószínűségi változók függetlenek is legyenek.

Tétel: Ha az $X_1, X_2, \ldots, X_n$ **azonos eloszlású, független valószínűségi változók**nak létezik $D(X_1)$, $D(X_2)$, $\ldots$, $D(X_n)$ szórása, akkor az összegüknek is létezik, és az **összegük szórásnégyzete egyenlő a szórásnégyzetek összegével**. Azaz

$$D^2(X_1 + X_2 + \ldots, X_n) = D^2(X_1) + D^2(X_2) + \ldots + D^2(X_n)$$

amit

$$D(X_1 + X_2 + \ldots + X_n) = \sqrt{D^2(X_1) + D^2(X_2) + \ldots + D^2(X_n)}$$

alakba is írhatunk.

Megjegyzés: Ha a valószínűségi változók várható értéke és szórása azonos, tehát

$$E(X_1) = E(X_2) = \ldots = E(X_n) = m$$

és

$$D(X_1) = D(X_2) = \ldots = D(X_n) = \sigma$$

akkor a fenti formulák az alábbi alakra egyszerűsödnek:

$$E(X_1 + X_2 + \ldots + X_n) = n \cdot m$$

$$D(X_1 + X_2 + \ldots + X_n) = \sqrt{n \cdot \sigma^2} = \sigma \cdot \sqrt{n}$$

Speciálisan normális eloszlású valószínűségi változók összegére ennél többet is mondhatunk:

Tétel: Ha $X_1, X_2, \ldots, X_n$ **független, normális eloszlású valószínűségi változók**, rendre (m_1, σ_1), (m_2, σ_2),...,(m_n, σ_n) paraméterekkel, akkor **összegük is normális eloszlású** valószínűségi változó,

$$\left(m_1 + m_2 + \ldots + m_n, \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}\right)$$

paraméterekkel.

Egyszerűbben leírva: ha $X_i : N(m_i, \sigma_i)$, minden $i = 1, 2, \ldots, n$ esetén, akkor

$$X_1 + X_2 + \ldots + X_n : N\big(m_1 + m_2 + \ldots + m_n, \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots, \sigma_n^2}\,\big)$$

Megjegyzés: Azonos paraméterek esetén az összeg várható értéke és szórása normális eloszlás esetén is egyszerűbb alakba írható: Ha $X_1, X_2, \ldots, X_n$ független, normális eloszlású valószínűségi változók, melyekre $X_i : N(m, \sigma)$, minden $i = 1, 2, \ldots, n$ esetén, akkor

$$X_1 + X_2 + \ldots + X_n : N\big(n \cdot m, \sigma \cdot \sqrt{n}\big)$$

Alkalmazzuk ezeket az összefüggéseket először különböző paraméterű valószínűségi változókra!

1. feladat Egy csokoládé kétféle kiszerelésben kapható. A kisebb tömege normális eloszlást követ 45 g várható értékkel és 2 g szórással. A nagyobb egy promóció keretében ideiglenesen kapható, tömege normális eloszlást követ, 95 g várható értékkel és 3 g szórással. Két kisebb és egy nagyobb kiszerelésű csokoládét vásárolunk. Mi a valószínűsége annak, hogy a két kisebb tömege összesen nagyobb, mint a nagy kiszerelésűé?

Megoldás: Három valószínűségi változónk van, mindhárom normális eloszlást követ, különböző paraméterekkel. Legyenek

X_1: az egyik kisebb csokoládészelet tömege, $m_1 = 45$, $\sigma_1 = 2$, paraméterekkel,

X_2: a másik kisebb szelet tömege, ekkor $m_2 = 45$, $\sigma_2 = 2$, paraméterekkel,

X_3: a nagyobb szelet tömege, ekkor $m_3 = 95$, $\sigma_3 = 3$, paraméterekkel.

Így $X_1 : N(45, 2)$, $X_2 : N(45, 2)$, és $X_3 : N(95, 3)$.

Kérdés: $P(X_1 + X_2 > X_3) = P(X_1 + X_2 - X_3 > 0) = ?$ Legyen

$$Y = X_1 + X_2 - X_3$$

Tudjuk, hogy Y normális eloszlású valószínűségi változó, melyre

$$E(Y) = E(X_1 + X_2 - X_3) = E(X_1) + E(X_2) - E(X_3) = 45 + 45 - 95 = -5$$

$$D(Y) = D(X_1 + X_2 - X_3) = \sqrt{D^2(X_1) + D^2(X_2) + D^2(X_3)} =$$
$$= \sqrt{2^2 + 2^2 + 3^2} = \sqrt{17} = 4,1231$$

Így $Y : N(-5; 4,1231)$, ezért:

$$P(X_1 + X_2 - X_3 > 0) = P(Y > 0) = P\Big(Y^* > \frac{0-(-5)}{4,1231}\Big) =$$

$$= P(Y^* > 1,2127) = 1 - \Phi(1,2127) = 1 - 0,8869 = 0,1131$$

0,1131 tehát a valószínűsége annak, hogy a két kisebb csokoládé össztömege több, mint a nagyobb szelet tömege.
A számolásnál kihasználtuk, hogy

$$E(-X_3) = -E(X_3)$$

és

$$D(-X_3) = |-1| \cdot D(X_3) = D(X_3)$$

Nézzük, azonos paraméterű valószínűségi változók összegére hogyan alkalmazhatjuk a tanult összefüggéseket!

2. feladat Egy bizonyos csokoládét hármasával csomagolva árulnak. Egy ilyen csokoládé tömege normális eloszlást követ, 100 g várható értékkel és 4 g szórással. Mi a valószínűsége annak, hogy egy hármas kiszerelést vásárolva több, mint 305 g csokoládét kapunk?

Megoldás: Három valószínűségi változónk van, mindhárom normális eloszlást követ, azonos paraméterekkel. Legyenek

X_1: az első csokoládészelet tömege, $m_1 = 100$, $\sigma_1 = 4$, paraméterekkel,

X_2: második szelet tömege, ekkor $m_2 = 100$, $\sigma_2 = 4$,

X_3: a harmadik szelet tömege, ekkor $m_3 = 100$, $\sigma_3 = 4$.

$X_1 : N(100, 4)$, $X_2 : N(100, 4)$, $X_3 : N(100, 4)$. Azonos paraméterű mindhárom valószínűségi változó, ezt kihasználva egyszerűsítjük a jelölést:

$$m_1 = m_2 = m_3 = 100 = m$$

$$\sigma_1 = \sigma_2 = \sigma_3 = 4 = \sigma$$

A hármas csomagolásban a csokoládé tömegét jelölje Y, melyre

$$Y = X_1 + X_2 + X_3$$

normális eloszlást követ, várható értéke és szórása:

$$E(Y) = E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = 3 \cdot m = 3 \cdot 100 = 300$$

$$D(Y) = D(X_1 + X_2 + X_3) = \sqrt{D^2(X_1) + D^2(X_2) + D^2(X_3)} = \sigma \cdot \sqrt{3} = 4 \cdot \sqrt{3} =$$
$$= 6,9282$$

A valószínűséget a normális eloszlás segítségével számoljuk, kihasználva, hogy $Y : N(300; 6,9282)$:

$$P(X_1 + X_2 + X_3 > 305) = P(Y > 305) = P\Big(Y^* > \frac{305-300}{6,9282}\Big) =$$

$$= P(Y^* > 0,7217) = 1 - \Phi(0,7217) = 1 - 0,7642 = 0,2358$$

A keresett valószínűség tehát 0,2358.

A megoldást látva felvetődik a kérdés, hogy az $Y = X_1 + X_2 + X_3$ felírás helyett miért nem alkalmaztuk az $Y = 3 \cdot X_1$ alakot, ha a három valószínűségi változó ugyanolyan eloszlású, ugyanolyan paraméterekkel.

A válaszhoz a következő összefüggést kell megismernünk:
Ha $X : N(m, \sigma)$ és a tetszőleges valós szám, akkor

$$a \cdot X : N(a \cdot m, \mid a \mid \cdot \sigma)$$

azaz egy (m, σ) paraméterű, normális eloszlású valószínűségi változót bármely a valós számmal szorzunk, $a \cdot X$ is normális eloszlást követ, $a \cdot m$ várható értékkel és $\mid a \mid \cdot \sigma$ szórással.

(Azt korábban is láttuk, hogy ha X valószínűségi változónak létezik várható értéke és szórása, akkor $E(a \cdot X) = a \cdot E(X)$ és $D(a \cdot X) = \mid a \mid \cdot D(X)$, ehhez a fenti állítás annyit tesz hozzá, hogy $a \cdot X$ is normális eloszlású, ha X normális eloszlású.)

A különbséget az alábbi feladaton is láthatjuk.

3. feladat Legyen $X : N(50; 4,3)$, X_1, X_2, X_3 pedig három, egymástól független, valószínűségi változó, melyek várható értéke rendre 50, szórása rendre $4,3$ (azaz $X_i : N(50; 4,3)$, $i = 1, 2, 3$ esetén).

a) $P(3 \cdot X < 146)$=?

b) $P(X_1 + X_2 + X_3 < 146)$=?

Megoldás: Amit tudunk:

$$E(X) = 50 = E(X_1) = E(X_2) = E(X_3)$$
$$D(X) = 4,3 = D(X_1) = D(X_2) = D(X_3)$$

a) Legyen $Y = 3 \cdot X$.

$$E(Y) = E(3 \cdot X) = 3 \cdot E(X) = 3 \cdot 50 = 150$$
$$D(Y) = D(3 \cdot X) = \mid 3 \mid \cdot D(X) = 3 \cdot 4,3 = 12,9$$

Kihasználjuk, hogy Y normális eloszlást követ, 150 várható értékkel és 12,9 szórással:

$$P(3 \cdot X < 146) = P(Y < 146) = P\Big(Y^* < \frac{146 - 150}{12,9}\Big) = P(Y^* < -0,3101) =$$
$$= 1 - \Phi(0,3101) = 1 - 0,6217 = 0,3783$$

b) Legyen $Z = X_1 + X_2 + X_3$.

$$E(Z) = E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = 3 \cdot 50 = 150$$
$$D(Z) = D(X_1 + X_2 + X_3) = \sqrt{D^2(X_1) + D^2(X_2) + D^2(X_3)} = 4,3 \cdot \sqrt{3} = 7,4478$$

$$P(X_1 + X_2 + X_3 < 146) = P(Z < 146) = P\Big(Z^* < \frac{146 - 150}{7,4478}\Big) =$$
$$= P(Z^* < -0,5371) = 1 - \Phi(0,5371) = 1 - 0,7054 = 0,2946$$

Látható, hogy a két valószínűségi változó csak a szórásában tér el egymástól, a várható értékében nem. Ugyanakkor ez az eltérés jelentősen befolyásolja a valószínűséget.

5.2. Valószínűségi változók átlagának várható értéke és szórása

4. feladat Egy dobozban 10 golyó van, 2 piros, 3 zöld, 5 fekete. Egy golyót húzunk a dobozból, az X valószínűségi változó értéke legyen 0, ha a kihúzott golyó piros, 1, ha zöld és 2, ha fekete. Húzunk egy golyót, majd feljegyezzük a hozzá tartozó valószínűségi változó értékét (X_1). Majd az eredeti kísérletet még egyszer elvégezzük (tehát visszatevéssel húzunk), és újra feljegyezzük a valószínűségi változó értékét (X_2). Írjuk fel az $\frac{X_1+X_2}{2}=\overline{X}$ valószínűségi változó várható értékét és szórását!

Megoldás: A feladatból X eloszlását könnyen felírhatjuk:

$$X : \begin{cases} 0 & 1 & 2 \\ 0,2 & 0,3 & 0,5 \end{cases}$$

Az eloszlásból X várható értékét és szórását is kiszámolhatjuk:

$$E(X) = 0 \cdot 0,2 + 1 \cdot 0,3 + 2 \cdot 0,5 = 1,3$$

$$E(X^2) = 0^2 \cdot 0,2 + 1^2 \cdot 0,3 + 2^2 \cdot 0,5 = 2,3$$

$$D(X) = \sqrt{2,3 - 1,3^2} = \sqrt{0,61} = 0,7810$$

Tekintsük át, hogy a mintavételnek milyen eredménye lehet. Például ha először piros golyót húzunk, akkor X_1 értéke 0, ha másodszor is pirosat, akkor X_2 értéke is nulla. Ennek valószínűsége a függetlenséget kihasználva $P(X_1 = 0) \cdot P(X_2 = 0) = 0,2 \cdot 0,2 = 0,04$.
Az összes lehetséges kimenetel X_1, X_2 értékei szerint: (0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2).
Írjuk fel ezekhez a megfelelő valószínűségeket!

(X_1, X_2)	$\overline{X} = \frac{X_1+X_2}{2}$	valószínűség
(0,0)	0	0,04
(0,1)	0,5	0,06
(0,2)	1	0,1
(1,0)	0,5	0,06
(1,1)	1	0,09
(1,2)	1,5	0,15
(2,0)	1	0,1
(2,1)	1,5	0,15
(2,2)	2	0,25

A táblázatból leolvashatjuk az átlag eloszlását (vegyük figyelembe, hogy az átlag értéke pl. $0,5$ kétféleképpen lehet, ezért a hozzá tartozó valószínűség: $0,06 + 0,06 = 0,12$). Így

$$\overline{X} : \begin{cases} 0 & 0,5 & 1 & 1,5 & 2 \\ 0,04 & 0,12 & 0,29 & 0,3 & 0,25 \end{cases}$$

A szokásos módon számolhatjuk a várható értéket és a szórást:

$$E(\overline{X}) = 0 \cdot 0,04 + 0,5 \cdot 0,12 + 1 \cdot 0,29 + 1,5 \cdot 0,3 + 2 \cdot 0,25 = 1,3$$

$$E(\overline{X}^2) = 0^2 \cdot 0,04 + 0,5^2 \cdot 0,12 + 1^2 \cdot 0,29 + 1,5^2 \cdot 0,3 + 2^2 \cdot 0,25 = 1,995$$
$$D(\overline{X}) = \sqrt{1,995 - 1,3^2} = \sqrt{0,3050} = 0,5523$$

Azt láthatjuk, hogy az átlag várható értéke megegyezik az eredeti valószínűségi változó várható értékével, a szórást megvizsgálva pedig azt vehetjük észre, hogy $\dfrac{D(X)}{D(\overline{X})} = \dfrac{0,7810}{0,5523} = 1,4141 = \sqrt{2}$. Ez azt jelenti, hogy $D(\overline{X}) = \dfrac{D(X)}{\sqrt{2}}$.

Általánosítsuk a feladat megoldása során kapott eredményt:
A valószínűségi változók összegére vonatkozó összefüggéseket felhasználva belátható, hogy ha valamely X valószínűségi változó megfigyelésére n elemű véletlen (független) mintát veszünk, melynek $X_1, X_2, \ldots, X_n$ (X-el azonos eloszlású) elemeinek várható értéke rendre m, szórása σ, akkor a minta átlagának($\overline{X}$) várható értéke és szórása: $E(\overline{X}) = m$, $D(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$.

Tétel: Ha az $X_1, X_2, \ldots, X_n$ azonos eloszlású, **független valószínűségi változók**nak létezik **várható értéke és szórása**, melyek **rendre megegyeznek**, azaz $E(X_1) = E(X_2) = \ldots = E(X_n) = m$ és $D(X_1) = D(X_2) = \ldots = D(X_n) = \sigma$, **akkor az átlaguk**nak is létezik **várható értéke és szórása**:

$$E\Big(\frac{X_1 + X_2 + \ldots + X_n}{n}\Big) = m$$

és

$$D\Big(\frac{X_1 + X_2 + \ldots, X_n}{n}\Big) = \frac{\sigma}{\sqrt{n}}$$

Tétel: Ha az $X_1, X_2, \ldots, X_n$ **normális eloszlású**, független valószínűségi változók várható értéke rendre m, szórása rendre σ, akkor az **átlaguk**, az

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

valószínűségi változó is **normális eloszlású**, m várható értékkel és $\dfrac{\sigma}{\sqrt{n}}$ szórással. Azaz $\overline{X} : N\big(m, \dfrac{\sigma}{\sqrt{n}}\big)$.

Megjegyzés: Az átlag várható értéke tehát megegyezik az eredeti várható értékkel, nem függ n-től, azonban az **átlag szórása** n **növelésével csökken**, tehát minél több megfigyelésünk van, a valószínűségi változók átlagának szórása annál kisebb. A gyakorlatban ez azt jelenti, hogy minél több megfigyelésünk van, az átlag várható értéktől való eltérése várhatóan annál kisebb lesz.

5. feladat Egy műszer élettartama normális eloszlást követ, 80 hónap várható értékkel és $2,5$ hónap szórással. 20 ilyen műszert tesztelünk, addig használjuk őket, míg tönkre nem mennek. Mindegyik esetében feljegyezzük az élettartamot. Mi a valószínűsége annak, hogy az élettartamok átlaga

a) kevesebb, mint 79 hónap?

b) több, mint 82 hónap?

Megoldás: X jelölje az műszer élettartamát

$$E(X) = 80 = m$$

$$D(X) = 2,5 = \sigma$$

A megfigyelésből az i-edik műszer élettartamát jelölje X_i $(i = 1, 2, \ldots, 20)$. Van $n = 20$ valószínűségi változónk, melyekre teljesül, hogy

$$\left\{ \begin{array}{l} \text{normális eloszlású} \\ \text{független} \\ M(X_i) = 80 = m \\ D(X_i) = 2,5 = \sigma \end{array} \right\}$$

valószínűségi változók. Ekkor az

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_{20}}{20}$$

normális eloszlású valószínűségi változó, $m = 80$ várható értékkel és $\dfrac{\sigma}{\sqrt{n}} = \dfrac{2,5}{\sqrt{20}} = 0,5590$ szórással.

a)

$$P\Big(\frac{X_1 + X_2 + \ldots + X_{20}}{20} < 79\Big) = P(\overline{X} < 79) = P\Big(\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} < \frac{79 - 80}{0,5590}\Big) =$$

$$= P\Big(X^* < \frac{79 - 80}{0,5590}\Big) = P(X^* < -1,7889) = 1 - \Phi(1,7889) = 1 - 0,9633 = 0,0367.$$

A transzformált valószínűségi változót (ami most $\dfrac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$) a megszokott X^*-gal jelöltük, a továbbiakban már csak ezt az egyszerűsített jelölést használjuk.

b)

$$P\Big(\frac{X_1 + X_2 + \ldots + X_{20}}{20} > 82\Big) = P(\overline{X} > 82) = 1 - P\Big(X^* < \frac{82 - 80}{0,5590}\Big) =$$

$$= 1 - P(X^* < 3,5778) = 1 - \Phi(3,5778) = 1 - 0,9998 = 0,0002$$

Ötlet: Nézzük meg, hogy egyetlen ilyen műszer esetén mekkora annak a valószínűsége, hogy 79 hónapon belül tönkremegy, és hasonlítsuk össze az eredményt az (a) részben az átlagra kapott értékkel. Mi az oka az eltérésnek?

5.3. Központi határeloszlás tétele

Eddig azt láttuk, hogy normális eloszlású valószínűségi változók összege és átlaga is normális eloszlást követ. Felvetődik a kérdés, hogy tetszőleges eloszlású valószínűségi változók átlagának eloszlásáról mit mondhatunk?

A következő összefüggést **központi határeloszlás tétele** néven ismerjük. Tulajdonképpen azt mondja ki, hogy **sok, azonos eloszlású, független, azonos várható értékű és azonos szórású valószínűségi változó átlaga (összege) közelítőleg normális eloszlású.** Az azonos eloszlás tetszőleges lehet.

Tétel: (Központi- vagy centrális határeloszlás tétele)
Ha az $X_1, X_2, \ldots, X_n$ valószínűségi változókra $n \geq 30$ esetén teljesül, hogy:

$$\left\{\begin{array}{l}\text{azonos eloszlásúak,}\\ \text{függetlenek,}\\ M(X_i) = m,\\ D(X_i) = \sigma,\end{array}\right\} \text{ minden } (i = 1, 2, \ldots, n)\text{-re,}$$

akkor az

- $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ valószínűségi változó közelítőleg normális eloszlású, m várható értékkel és $\dfrac{\sigma}{\sqrt{n}}$ szórással, azaz $\overline{X} \approx N\big(m, \dfrac{\sigma}{\sqrt{n}}\big)$.
- $X_1 + X_2 + \ldots + X_n$ valószínűségi változó közelítőleg normális eloszlású, $n \cdot m$ várható értékkel és $\sigma \cdot \sqrt{n}$ szórással, azaz $X_1 + X_2 + \ldots + X_n \approx N\big(n \cdot m, \sigma \cdot \sqrt{n}\big)$.

6. feladat Adott $n = 40$ darab azonos eloszlású, független, valószínűségi változó, melyek várható értéke rendre 100, szórása rendre 12. Mekkora lehet annak a valószínűsége, hogy az átlaguk

a) 100,5-nél kevesebb?

b) 99 és 100 közé esik?

c) 99 és 101 közé esik?

Megoldás: A feladatból tudjuk, hogy

$$X_1, X_2, \ldots, X_{40}\colon \left\{\begin{array}{l}\text{azonos eloszlású}\\ \text{függetlenek}\\ M(X_i) = 100 = m\\ D(X_i) = 12 = \sigma\end{array}\right\} \text{ valószínűségi változók.}$$

A központi határeloszlás tétel szerint az átlaguk,

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_{40}}{40}$$

közelítőleg normális eloszlást követ,

$$m = 100$$

várható értékkel és

$$\frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{40}} = 1,8974$$

szórással. A normális eloszlás szerint számolunk:

a)

$$P\Big(\frac{X_1+X_2+\ldots+X_{40}}{40} < 100,5\Big) = P(\overline{X} < 100,5) = P\Bigg(X^* < \frac{100,5-100}{\frac{12}{\sqrt{40}}}\Bigg) =$$

$$= P\Big(X^* < \frac{100,5-100}{1,8974}\Big) = P(X^* < 0,2635) \approx \Phi(0,2635) = 0,6026$$

b)

$$P\Big(99 < \frac{X_1+X_2+\ldots+X_{40}}{40} < 100\Big) = P\Big(99 < \overline{X} < 100\Big) =$$

$$= P\Big(\frac{99-100}{1,8974} < X^* < \frac{100-100}{1,8974}\Big) = P(-0,5270 < X^* < 0) \approx \Phi(0) - \Phi(-0,5270) =$$

$$= \Phi(0) - \big[1-\Phi(0,5270)\big] = 0,5 - 1 + 0,7019 = 0,2019$$

c)

$$P\Big(99 < \frac{X_1+X_2+\ldots+X_{40}}{40} < 101\Big) = P\Big(99 < \overline{X} < 101\Big) = P\Big(\frac{99-100}{1,8974} < X^* < \frac{101-100}{1,8974}\Big) =$$

$$= (-0,5270 < X^* < 0,5270) \ \approx 2\cdot\Phi(0,5270) - 1 = 2\cdot 0,7019 - 1 = 0,4038$$

7. feladat Egy új típusú gumiabroncs élettartamának eloszlása nem ismert, a várható értéke 43 500 km, a szórása 345 km. 100 ilyen gumiabroncsot tesztelünk, megmérjük az élettartamukat.

a) Mekkora lehet annak a valószínűsége annak, hogy az élettartamuk átlaga legalább az átlag várható értéke?

b) Melyik az az élettartam, amelynél a gumiabroncsok élettartamának átlaga kb. 0,95 valószínűséggel kisebb?

c) Kb. 0,95 valószínűséggel milyen, várható értékre szimmetrikus intervallumba esik az abroncsok élettartamának átlaga?

d) Hogyan változik a (c) kérdésben meghatározott intervallum, ha a tesztelt abroncsok számát $n = 400$-ra növeljük?

Megoldás: Legyen X_i: az i-edik abroncs élettartama, $(i = 1, 2, \ldots, n)$, $n = 100$. Tudjuk, hogy

$$X_1, X_2, \ldots, X_{100}\colon \left\{\begin{array}{l} \text{azonos eloszlású} \\ \text{független} \\ M(X_i) = 43\,500 = m \\ D(X_i) = 345 = \sigma \end{array}\right\} \text{valószínűségi változók.}$$

Így

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_{100}}{100}$$

közelítőleg normális eloszlást követ,

$$m = 43\,500$$

várható értékkel és

$$\frac{\sigma}{\sqrt{n}} = \frac{345}{\sqrt{100}} = 34,5$$

szórással.

a)

$$P\Big(\frac{X_1 + X_2 + \ldots + X_{100}}{100} \geq m\Big) = 1 - P\Big(\overline{X} < 43500\Big) = 1 - P\Big(X^* < \frac{43500 - 43500}{34,5}\Big) =$$

$$= 1 - P(X^* < 0) \approx 1 - \Phi(0) = 1 - 0,5 = 0,5$$

A továbbiakban a valószínűségi változók átlagát már nem írjuk ki külön lépésben, csak $\overline{X}$-ként.

b) A keresett élettartamot jelölje b. Keressük azt a b-t, amelyre

$$P(\overline{X} < b) \approx 0,95$$

teljesül.

Kihasználjuk, hogy az átlag közel normális eloszlású, és az egyenlet bal oldalán transzformálunk:

$$P(\overline{X} < b) = P\Big(X^* < \frac{b - 43500}{34,5}\Big) \approx \Phi\Big(\frac{b - 43500}{34,5}\Big)$$

Így

$$\Phi\Big(\frac{b - 43500}{34,5}\Big) \approx 0,95$$

A táblázatból:

$$\frac{b - 43500}{34,5} \approx 1,65$$

innen

$$b \approx 43500 + 1,65 \cdot 34,5 = 43556,925$$

A $(-\infty; 43556,925)$ intervallumba kb. $0,95$ valószínűséggel várjuk az élettartamok átlagának értékét.

(Az ilyen intervallumokat szokás egyoldali konfidencia-intervallumnak nevezni.)

A száz gumiabroncs élettartamának átlaga kb. $0,95$ valószínűséggel kisebb, mint 43 556,925 km.

c) Keressük azt a c-t, amelyre

$$0,95 \approx P(m - c < \overline{X} < m + c) = P(43500 - c < \overline{X} < 43500 + c)$$

A jobb oldal:

$$P(43500 - c < \overline{X} < 43500 + c) =$$

$$= P\Big(\frac{(43500 - c) - 43500}{34,5} < X^* < \frac{(43500 + c) - 43500}{34,5}\Big) \approx$$

$$\approx P\Big(\frac{-c}{34,5} < X^* < \frac{c}{34,5}\Big) \approx 2 \cdot \Phi\Big(\frac{c}{34,5}\Big) - 1$$

A következő egyenletet kaptuk:

$$2 \cdot \Phi\Big(\frac{c}{34,5}\Big) - 1 \approx 0,95$$

Rendezés után:

$$\Phi\Big(\frac{c}{34,5}\Big) \approx \frac{0,95+1}{2} = 0,9750$$

A táblázatból:

$$\frac{c}{34,5} \approx 1,96$$

$$c = 1,96 \cdot 34,5 = 67,62$$

A keresett intervallum:

$$(43\,500 - 67,62; 43\,500 + 67,62)$$

azaz

$$(43\,432,38; 43\,567,62)$$

(Az ilyen intervallumokat szokás kétoldali konfidencia-intervallumnak nevezni.)

Gondoljuk át, mit jelent a kapott eredmény:
A gumiabroncsok tesztelése előtt, az ismert várható értékből és szórásból meg tudjuk mondani, hogy a $(43432,38; 43\,567,62)$ intervallumba nagy, kb. $0,95$ valószínűséggel várjuk a mérési eredményeink átlagát. Ez azt jelenti, hogy az átlag ebbe az intervallumba fog esni? Természetesen nem, kicsi, kb. $0,05$ valószínűséggel ezen kívüli átlagélettartamot is kaphatunk. Ugyanakkor ennek kicsi az esélye, általában azt várjuk, hogy egy nagy valószínűségű esemény következik be. Később a hipotézisvizsgálatnál megnézzük, hogy hogyan értelmezzük azt az esetet, amikor 'nem várt', kis valószínűségű eredményt kapunk a kísérleteink, méréseink eredményeként.

d) Az $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ átlag várható értékét nem befolyásolja az n értéke, a szórást annál inkább. Az

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_{400}}{400}$$

valószínűségi változó várható értéke továbbra is

$$m = 43\,500$$

szórása viszont,

$$\frac{\sigma}{\sqrt{n}} = \frac{345}{\sqrt{400}} = 17,25$$

a felére csökkent.
(n-et a négyszeresére növeltük, a szórás a felére csökkent ($\sqrt{4} = 2$).)
Ezzel a szórással a (c) megoldása a következőképpen módosul:

$$P(43500 - c < \overline{X} < 43500 + c) =$$

$$= P\Big(\frac{(43500 - c) - 43500}{17,25} < X^* < \frac{(43500 + c) - 43500}{17,25}\Big) \approx$$

$$\approx P\Big(\frac{-c}{17,25} < X^* < \frac{c}{17,25}\Big) \approx 2\cdot\Phi\Big(\frac{c}{17,25}\Big) - 1$$

Így

$$2\cdot\Phi\Big(\frac{c}{17,25}\Big) - 1 \approx 0,95$$

$$\Phi\Big(\frac{c}{17,25}\Big) \approx \frac{0,95+1}{2} = 0,9750$$

$$\frac{c}{17.25} \approx 1,96$$

$$c = 1,96 \cdot 17.25 = 33,81$$

Az intervallum sugara felére csökkent:

$$(43\,500 - 33,81; 43\,500 + 33,81)$$

azaz

$$(43\,466,19; 43\,533,81)$$

Ez azt jelenti, hogy minél több gumiabroncsot tesztelünk, az élettartamuk átlagát annál kisebb sugarú intervallumba várjuk, ugyanakkora valószínűség mellett.

Ötlet: Oldjuk meg az előző feladat (c) részét, ha a valószínűséget $0,98$-ra növeljük. Figyeljük meg, hogyan változik az intervallum sugara!

Megjegyzés: Ha a valószínűséget változtatjuk, az intervallum sugara szintén változni fog, mégpedig a valószínűség növelésével nő az intervallum sugara, csökkentésével csökken az intervallum sugara.

8. feladat Egy főútvonalon a burkolathibák száma 100 m-ként átlagosan $1,8$. Egy 12 km-es szakaszon összeszámoljuk a burkolathibákat.

a) Mekkora lehet annak a valószínűsége, hogy a 12 km-es szakaszon legalább 230 burkolathibát találunk?

b) Kb. $0,95$ valószínűséggel legalább hány burkolathibát találunk?

Megoldás: Első ránézésre a feladat megoldásához nincs szükség a központi határeloszlás tételére. Legyen

$$X: \text{burkolathibák száma 12 km-en.}$$

X Poisson-eloszlást követ, $120 \cdot 1,8 = 216$ várható értékkel, azaz $\lambda = 216$. Ez pedig azt jelenti, hogy a Poisson-eloszlást felhasználva pontos megoldást adhatunk a kérdésekre. Ugyanakkor ha felírjuk az (a) kérdést és alkalmazzuk a Poisson-eloszlást, akkor azt látjuk, hogy 231 tagú összegre vezet, túl sok a számítási igény:

$$P(X \geq 230) = P(X=0) + P(X=1) + P(X=2) + \ldots + P(X=230) =$$

$$= \sum_{k=0}^{230} \frac{216^k}{k} \cdot e^{-230} =?$$

Pontos megoldást (legalábbis számológéppel) tehát ebben az esetben akkor sem tudunk adni, ha ismerjük az eloszlást és annak paraméterét.

Próbáljuk meg másképp megközelíteni a feladatot. A 12 km-t rakjuk össze 100 m-es darabokból. Legyenek

$$X_i : \text{burkolathibák száma az } i\text{-edik } 100\,\text{m-en}$$

$i = 1, 2, \ldots 120$ esetén. (Tehát pl. X_5 az ötödik 100 m-en, azaz 400 és 500 m között a hibák száma.)
Van $n = 120$ darab valószínűségi változónk (a korábbi egy helyett), ugyanakkor a korábbi X-et felírhatjuk ezek segítségével:

$$X = X_1 + X_2 + \ldots X_{120}$$

Mit tudunk az X_i valószínűségi változókról?

$$X_1, X_2, \ldots, X_{120}: \left\{ \begin{array}{l} \text{azonos, Poisson-eloszlású} \\ \text{független} \\ M(X_i) = 1,8 = m \\ D(X_i) = \sqrt{1,8} = \sigma \end{array} \right\} \text{valószínűségi változók.}$$

(Felhasználtuk, hogy a Poisson-eloszlásból $D(X_i) = \sqrt{1,8}$, a tizedestört alakot most a kerekítési hibák elkerülése miatt csak később írjuk fel.)
Ez azt jelenti, hogy az $X = X_1 + X_2 + \ldots X_{120}$-re alkalmazhatjuk a központi határeloszlás tételének összegre vonatkozó alakját. (Figyeljünk arra, hogy most az összeg közel normális eloszlású, így az összeghez tartozó paraméterekkel kell számolni a transzformáció során!)
Tehát

$$X = X_1 + X_2 + \ldots X_{120}$$

közelítőleg normális eloszlású valószínűségi változó,

$$n \cdot m = 120 \cdot 1,8 = 216$$

várható értékkel és

$$\sigma \cdot \sqrt{n} = \sqrt{1,8} \cdot \sqrt{120} = 14,6969$$

szórással.
(A megoldás során az összeg standardizáltját a szokásos X^*-gal jelöljük.)

a) Írjuk fel a kérdést , majd a normális eloszlásnál megszokott lépéseket kövessük:

$$P(X \geq 230) = P(X_1 + X_2 + \ldots X_{120} \geq 230) = 1 - P(X^* < \frac{230 - 216}{14,6969}) = 1 - P(X^* < 0,9526) \approx$$
$$\approx 1 - \Phi(0,9526) = 1 - 0,8289 = 0,1711$$

A keresett valószínűség tehát kb. $0,1711$.
(A pontos megoldás 4-tizedesjegyre kerekítve, Geogebra programmal számolva: 0,1618)

Látjuk, hogy elég 'jó' becslést adtunk a központi határeloszlás tételének segítségével. Azonban e becslés javítható egy korrekcióval, mivel a Poisson-eloszlás diszkrét, a normális eloszlás folytonos. Ezzel mi most nem foglalkozunk. (A Moivre-Laplace formulánál megnézzük, hogy hogyan működik a korrekció.)

b) Keressük azt a a értéket, amelyre
$$P(X \geq a) \approx 0,95$$
A bal oldal:
$$P(X \geq a) = P(X_1 + X_2 + \dots X_{120} \geq a) = 1 - P\Big(X^* < \frac{a-216}{14,6969}\Big)$$
A kapott egyenlet:
$$1 - P\Big(X^* < \frac{a-216}{14,6969}\Big) \approx 0,95$$
Vegyük észre, hogy ez azt jelenti, hogy a $\frac{a-216}{14,6969}$ tört értéke negatív, így alkalmazzuk az előjel korrekciót:
$$1 - P\Big(X^* < \frac{a-216}{14,6969}\Big) = P\Big(X^* < \frac{-a+216}{14,6969}\Big) \approx \Phi\Big(\frac{-a+216}{14,6969}\Big)$$
Az egyenlet tehát
$$\Phi\Big(\frac{-a+216}{14,6969}\Big) \approx 0,95$$
A táblázatból
$$\frac{-a+216}{14,6969} \approx 1,65$$
$$a = 216 - 1,65 \cdot 14,6969 = 191,7501$$
Kb. 0,95 valószínűséggel legalább 192 úthibát találunk a 12 km-es szakaszon.

Megjegyzés: Ha a központi határeloszlás tételét alkalmazzuk, akkor a valószínűségi változók számának (n) növelésével 'jobb' becslést adhatunk.

5.4. Moivre-Laplace formula

A Moivre- Laplace formula már a központi határeloszlás tétele előtt is ismert volt. Ugyanakkor a központi határeloszlás tételének speciális eseteként is kimondható. Mi ez utóbbit nézzük meg, egy nagyon egyszerű feladat kapcsán:

9. feladat Egy ládában tranzisztorok vannak, melyek 16%-a hibás. Visszatevéssel húzunk 300-szor a ládából. Mekkora lehet annak a valószínűsége, hogy a kihúzott hibás tranzisztorok száma legalább 40 és legfeljebb 55?

Megoldás: Első ránézésre egyszerű a feladat, korábban láttunk hasonlót. Most azonban a húzások száma meglehetősen nagy. Nézzük meg, hogyan befolyásolja ez a megoldhatóságot!

- **Pontos megoldás:** Legyen X a kihúzott hibás tranzisztorok száma. Ekkor az A: 'hibás a kihúzott tranzisztor' esemény bekövetkezéseinek száma a valószínűségi változó, független kísérletek esetén. Ez binomiális eloszlást követ, $n = 300$ és $p = P(A) = 0,16$ paraméterekkel.
$$P(40 \leq X \leq 55) = \Big[P(X=40) + P(X=41) + \ldots + P(X=55)\Big] =$$
$$\Big[0,16^{40} \cdot (0,84)^{260} \cdot \binom{300}{40} + 0,16^{41} \cdot (0,84)^{259} \cdot \binom{300}{41} + \ldots + 0,16^{55} \cdot (0,84)^{245} \cdot \binom{300}{55}\Big] =?$$
A megoldással most nem is az a legnagyobb probléma, hogy túl sok tagból áll, hanem az, hogy egyiket sem tudjuk kiszámítani a számológépünkkel. (Próbáljuk meg a $\binom{300}{40}$-et beírni a számológépünkbe, nem tudja az eredményt kiszámolni.) Fel tudjuk tehát írni a pontos megoldást, de nem tudjuk kiszámolni, így a keresett valószínűség értékét még közelítőleg sem tudjuk megadni. Mit tehetünk?

- **Közelítő megoldás központi határeloszlás tételével:** Alkalmazzuk a burkolathibák számánál látott trükköt: bontsuk szét a feladatot részekre, ebben az esetben az egyes dobásokat tekintsük külön-külön: Legyenek

$$X_i : \begin{cases} 1 & \text{ha az i. húzás hibás tranzisztor} \\ 0 & \text{ha az i. húzás nem hibás tranzisztor} \end{cases}$$

Ezen valószínűségi változók eloszlása azonos, $i = 1, 2, \ldots, 300$ esetén:

$$X_i : \begin{cases} 0 & 1 \\ 0,84 & 0,16 \end{cases}$$

így a várható értékük is megegyezik:

$$E(X_i) = 0 \cdot 0,84 + 1 \cdot 0,16 = 0,16$$

Felhasználva, hogy

$$E(X_i^2) = 0^2 \cdot 0,84 + 1^2 \cdot 0,16 = 0,16$$

a szórásuk is azonos:

$$D(X_i) = \sqrt{0,16 - 0,16^2} = \sqrt{0,16 \cdot (1 - 0,16)}$$

(A szórást egyenlőre ne írjuk tizedestört alakba.)

Így az $n = 300$ valószínűségi változóra teljesül:

$$X_1, X_2, \ldots, X_{300}: \left\{ \begin{array}{l} \text{azonos eloszlású} \\ \text{független} \\ M(X_i) = 0,16 = m \\ D(X_i) = \sqrt{0,16 \cdot (1 - 0,16)} = \sigma \end{array} \right\} \text{valószínűségi változók.}$$

Az összegükre így a központi határeloszlás tétele alapján teljesül, hogy

$$X_1 + X_2 + \ldots + X_{300}$$

közelítőleg normális eloszlást követ,

$$n \cdot m = 300 \cdot 0,16 = 48$$

várható értékkel és

$$\sigma \cdot \sqrt{n} = \sqrt{300} \cdot \sqrt{0,16 \cdot (1 - 0,16)} = 6,3498$$

szórással.

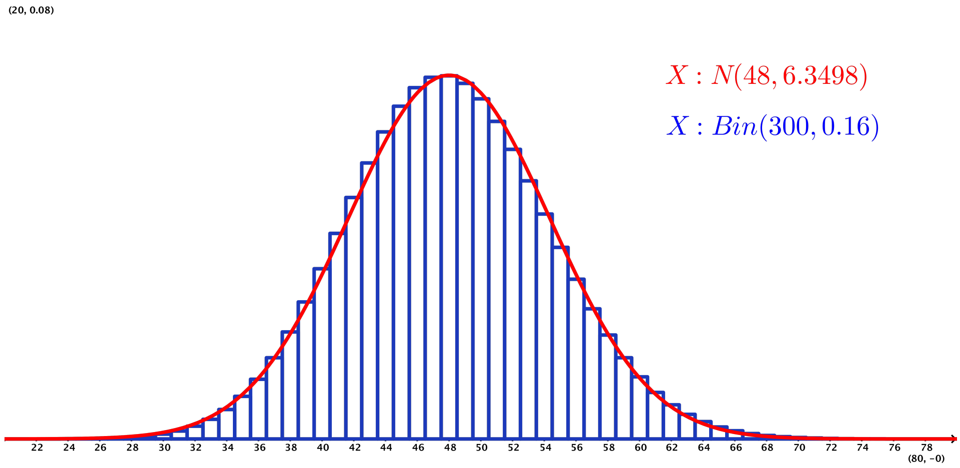
Mielőtt azonban alkalmaznánk a központi határeloszlás tételét, nézzük meg kicsit jobban az $X_1 + X_2 + \ldots + X_{300}$ összeget. Ebben pontosan annyi valószínűségi változó értéke 1, ahány húzás hibás tranzisztort eredményezett, a többi értéke pedig nulla. Vagyis ez az összeg pontosan azt mondja meg, hány hibás tranzisztort húzunk. Igen ám, de ez a kihúzott hibás tranzisztorok száma összesen, ami pedig a pontos megoldásnál X-el jelölt valószínűségi változó. Ez azt jelenti, hogy X, amelynek a pontos eloszlása binomiális, közelíthető normális eloszlással. Ha megnézzük az összeg várható értékét $(n \cdot p)$, az éppen a binomiális eloszlás várható értéke, az összeg szórása $\left(\sqrt{n} \cdot \sqrt{p \cdot (1 - p)}\right)$ pedig éppen a binomiális eloszlás szórása. Tulajdonképpen azt kaptuk, hogy az eredeti binomiális eloszlású valószínűségi változó olyan normális eloszlással közelíthető, mely 'örökli' a binomiális eloszlás várható értékét és szórását. A közelítő megoldás:

$$P(40 \leq X_1 + X_2 + \ldots X_{300} \leq 55) = P(40 \leq X \leq 55) = P\Big(\frac{40 - 48}{6,3498} \leq X^* \leq \frac{55 - 48}{6,3498}\Big) =$$

$$= P(-1,2599 \leq X^* \leq 1,1024) \approx \Phi(1,1024) - \Phi(-1,2599) = \Phi(1,1024) - \Big[1 - \Phi(1,2599)\Big] =$$

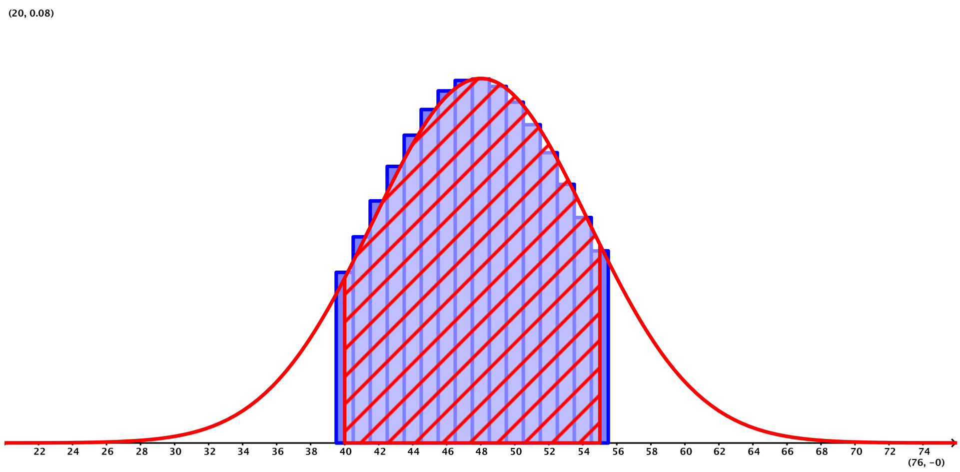
$$0,8643 - 1 + 0,8962 = 0,7605$$

- **Közelítő megoldás korrekcióval:** Ha alaposan megnézzük a számításunkat, akkor látjuk, hogy egy diszkrét valószínűségi változót közelítettünk folytonossal. Az alábbi ábrán a feladatban szereplő, $n = 300$ és $p = 0,16$ paraméterű binomiális eloszlás sűrűséghisztogramját (kék), illetve az $m = 48$ és $\sigma = 6,3498$ paraméterű normális eloszlás sűrűségfüggvényét (piros) látjuk. A sűrűség-



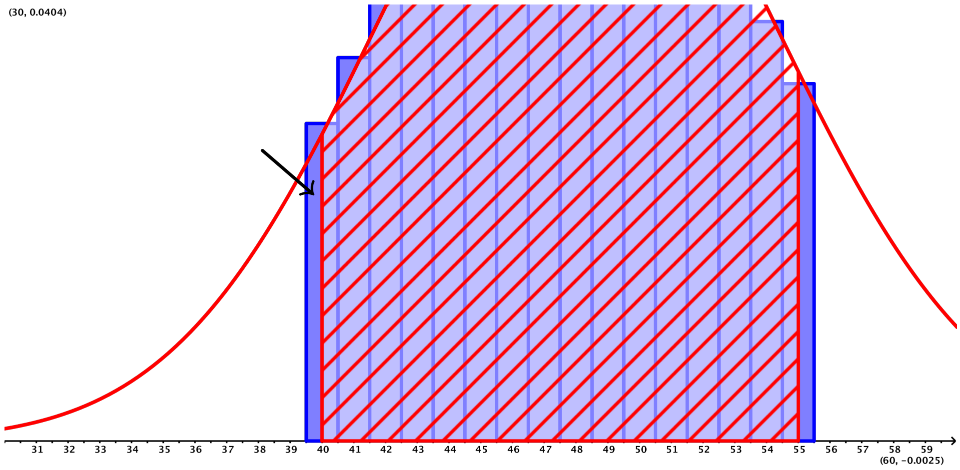
27. ábra.

hisztogramon a téglalapok területe, a sűrűségfüggvény grafikonján a görbe alatti terület adja az adott intervallumhoz tartozó valószínűséget. A keresett $P(40 \leq X_1 + X_2 + \ldots X_{300} \leq 55)$ valószínűséget ábrázoltuk mindkét eloszlás esetében a következő ábrán (kékkel színezve a binomiális, pirossal csíkozva a normális eloszlás esetén):



28. ábra.

A két terület közelítőleg azonos, egyedül az intervallumok szélénél látható nagyobb eltérés. Nézzük meg a következő ábrán az intervallum baloldali szélét kinagyítva:



29. ábra.

Itt az látszik, hogy az integrál a görbe alatti területet (piros vonalazott) 40-től számolja, a sűrűséghisztogramon viszont a 40 lehetséges értékhez tartozó valószínűséget megadó téglalap területe 'túllóg' ezen, 39,5-ig. Ezzel a normális eloszlással ezt a (nyíllal) jelzett területet nem becsültük, ami feltehetőleg arra vezet, hogy a valószínűség becsült értéke kisebb a pontos értéknél. Hogyan lehet ezt a nyilvánvaló pontatlanságot csökkenteni? Egyszerűen annyit teszünk, hogy 40 helyett 39,5-re változtatjuk az intervallum határát. (Binomiális eloszlásnál ez nem jelent érdemi változást, mert nem lehet fél hibás tranzisztort húzni, a normális eloszlás becslésének hibáját pedig csökkenti ez a felírás.) Ugyanígy kitoljuk a jobboldali intervallum végpontot is:

$$P(40 \leq X_1 + X_2 + \ldots X_{300} \leq 55) = P(40 \leq X \leq 55) = P(39,5 < X < 55,5) =$$

$$= P\Big(\frac{39,5-48}{6,3498} < X^* < \frac{55,5-48}{6,3498}\Big)) = P(-1,3386 < X^* < 1,1811) \approx$$

$$\approx \Phi(1,1811) - \Phi(-1,3386) = \Phi(1,1811) - \Big[1 - \Phi(1,3386)\Big] = 0,8810 - 1 + 0,9099 = 0,7909$$

A pontos megoldás (Geogebrával számolva) 4-tizedesjegyre kerekítve 0,7923. Látható, hogy a korrekció lényegesen javította a becslésünket.

Általánosítsuk a fenti feladatban szerzett tapasztalatainkat.

Moivre-Laplace formula Ha az X binomiális eloszlású valószínűségi változó paramétereire teljesül, hogy $n \geq 30$, p nem túl kicsi és nem túl nagy, akkor X közel normális eloszlású valószínűségi változó, $n \cdot p$ várható értékkel és $\sqrt{n \cdot p \cdot \ (1-p)}$ szórással. Azaz

$$P(a \leq X \leq b) \approx \Phi\Big(\frac{b - n \cdot p}{\sqrt{n \cdot p \cdot \ (1-p)}}\Big) - \Phi\Big(\frac{a - n \cdot p}{\sqrt{n \cdot p \cdot \ (1-p)}}\Big)$$

Alkalmazva a korrekciót:

$$P(a \leq X \leq b) = P(a - 0,5 < X < b + 0,5) \approx \Phi\Big(\frac{b + 0,5 - n \cdot p}{\sqrt{n \cdot p \cdot \ (1-p)}}\Big) - \Phi\Big(\frac{a - 0,5 - n \cdot p}{\sqrt{n \cdot p \cdot \ (1-p)}}\Big)$$

A formulák helyett a feladatok megoldásánál talán hasznosabb azt megtanulni, hogy ha binomiális eloszlású valószínűségi változót közelítünk normális eloszlással, akkor a normális eloszlásnál tanultak szerint járunk el, a paramétereket a binomiális eloszlásból származtatva. A következő feladatban ezt látjuk.

10. feladat Egy légitársaság tapasztalatai szerint a jeggyel rendelkező utasok 0,96 valószínűséggel jelennek meg (egymástól függetlenül) ténylegesen a beszállásnál. Egy alkalommal a 320 férőhelyes gépre 325 jegyet adnak el. Mi a valószínűsége annak, hogy minden megjelent utas felfér a gépre?

Megoldás: A X valószínűségi változó a megjelent utasok száma. (Vagyis az A: 'megjelenik egy utas a beszállásnál' eseményt figyeljük meg 325 esetben.) Ez binomiális eloszlást követ, $n = 325$ és $p = P(A) = 0,96$ paraméterekkel. A kérdés az, hogy mi annak a valószínűsége, hogy a megjelent utasok száma legfeljebb a férőhelyek száma, azaz 320. A binomiális eloszlást normálissal közelítjük és alkalmazzuk a korrekciót is.
Határozzuk meg első lépésként a binomiális eloszlás várható értékét és szórását!

$$E(X) = n \cdot p = 325 \cdot 0,96 = 312$$

$$D(X) = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{325 \cdot 0,96 \cdot 0,04} = \sqrt{12,48} = 3,5327$$

Ezeket felhasználjuk, mint a normális eloszlás paramétereit:

$$P(0 \leq X \leq 320) = P(-0,5 < X < 320,5) = P\Big(\frac{-0,5-312}{3,5327} < X^* < \frac{320,5-312}{3,5327}\Big)) =$$

$$= P(-88,4593 < X^* < 2,4061) \approx$$

$$\approx \Phi(2,4061) - \Phi(-88,4593) = \Phi(2,4061) - \Big[1 - \Phi(88,4593)\Big] = 0,9920 - 1 + 1 = 0,9920$$

Megjegyzés: $\Phi(-88,4593)$ értéke (4-tizedesre kerekítve) 0, így a becslés értékét nem befolyásolta, ki sem kellett volna írni a feladat felírásakor. Ilyen esetekben általában azt mondhatjuk, ha az intervallum végpontja a várható értéktől a szórás 4-szeresénél távolabb van, akkor csak abban az esetben érdemes kiírni, ha több tizedesjegyet tartalmazó standard normális eloszlás táblázatot használunk.

6. Hipotézisvizsgálat I.

- Megfelel-e egy termék valós mérete a gyártó állításának?
- Normális eloszlást követ-e egy termék mérete?
- Van-e összefüggés az iskolai végzettség és az egészségi állapot között?
- Két ugyanolyan terméket gyártó gyáregységben azonosnak tekinthető-e a selejtarány?
- Minőségellenőrzés során eleget tesz-e a gyártás bizonyos követelményeknek?

Sokáig tudnánk sorolni a hasonló kérdéseket, akár a mindennapjainkból, akár tudományos kutatásokból. Az a közös bennük, hogy mindegyik esetben egy véges (sokszor egészen kis számú) minta alapján akarunk következtetni egy sokkal nagyobb, akár végtelen sokaságra. Gyakran már a mintavétel előtt van sejtésünk egy (vagy akár több) eloszlásra, annak paramétereire vonatkozóan, amit a minta segítségével szeretnénk igazolni. Ebben a fejezetben megnézzük, hogy mindezt hogyan tehetjük meg.

6.1. Statisztikai minta paramétereinek becslése

Egy sokaság ismeretlen paramétereit a sokaságból vett véletlen minta (minták) segítségével becsülhetjük.

Definíció: Statisztikai mintán a megfigyelt X valószínűségi változóval azonos eloszlású, független $X_1, X_2, \ldots X_n$ valószínűségi változók összességét értjük.

$X_1, X_2, \ldots X_n$-et mintaelemeknek is szokás nevezni. X egyes paramétereinek becsléséhez tulajdonképpen értelmezünk egy-egy többváltozós függvényt, amelynek az értékét az adott mintaelemeken kiszámoljuk. A függvény adott mintához tartozó értékével becsüljük a keresett paraméter értékét. Ebben az esetben **pontbecslés**ről beszélünk, mert az ismeretlen paraméter értékét egyetlen valós számmal becsüljük. (Szokás még **intervallum becslés**t alkalmazni, amikor a becsült paraméter értékére egy intervallumot adunk meg. Erre példa a korábban látott konfidencia-intervallum, ezzel mi nem foglalkozunk.)

A mintaelemekből sokféle pontbecslést adhatunk. Ezek közül melyiket tekintjük jó becslésnek? A becsléseknek különféle szempontok szerint többféle osztályozása van. Mi ebből kettőt említünk meg:

a) **Torzítatlan**nak nevezünk egy **becslés**t, ha a becslő paraméter várható értéke a becsült paraméterrel egyezik meg.

b) **Hatásos**nak nevezünk egy **becslés**t, ha a becslő paraméter szórása a becsült paraméter körül minimális.

A továbbiakban a megfigyelt valószínűségi változó várható értékét és szórását akarjuk a mintából becsülni. Ehhez két statisztikai függvényt hívunk segítségül:

Definíció: Az $X_1, X_2, \ldots X_n$ mintaelemek **mintaátlaga** az $\hat{m}_n$-nel jelölt valós szám, melyre

$$\hat{m}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

teljesül.

A mintaátlag hatásos és torzítatlan becslése a várható értéknek.

Definíció: Az $X_1, X_2, \ldots X_n$ mintaelemek **tapasztalati szórása** a $\hat{\sigma}_n$-nel jelölt nemnegatív szám, melyre

$$\hat{\sigma}_n = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \hat{m}_n)^2} = \sqrt{\frac{(X_1 - \hat{m}_n)^2 + (X_2 - \hat{m}_n)^2 + \ldots + (X_n - \hat{m}_n)^2}{n}}$$

teljesül.
Számításoknál használhatjuk a

$$\hat{\sigma}_n = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} X_i \right)^2}$$

formulát is.

Gondoljuk át azonban, hogy a tapasztalati szórás négyzete a mintaelemeknek a **belőlük kiszámolt átlagtól** való eltérésének négyzetösszegének átlagát jelenti. A mintaelemek általában közelebb vannak a saját átlagukhoz, mint a várható értékhez. Ebből következik, hogy a tapasztalati szórás értéke szinte biztosan kisebb, mint a szórás elméleti értéke. Tehát növelni kellene a tapasztalati szórás értékét. Ehhez pedig a definícióban látott összeg nevezőjét csökkentjük.

Definíció: Az $X_1, X_2, \ldots X_n$ mintaelemek **korrigált tapasztalati szórása** az $\hat{s}_n$-nel jelölt nemnegatív szám, melyre

$$\hat{s}_n = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \hat{m}_n)^2} = \sqrt{\frac{(X_1 - \hat{m}_n)^2 + (X_2 - \hat{m}_n)^2 + \ldots + (X_n - \hat{m}_n)^2}{n-1}}$$

teljesül.

$\hat{s}_n$ már torzítatlan becslése a szórásnak.
A tapasztalati szórás és a korrigált tapasztalati szórás közötti kapcsolat:

$$\hat{s}_n = \sqrt{\frac{n}{n-1}} \cdot \hat{\sigma}_n$$

Tudjuk, hogy $\lim\limits_{n \to \infty} \sqrt{\frac{n}{n-1}} = 1$, tehát nagy n esetén nincs lényeges különbség $\hat{\sigma}_n$ és $\hat{s}_n$ között, kicsi n esetén azonban lényeges a különbség. (Ha mindig $\hat{s}_n$-ot használunk a szórás becslésére, akkor biztosan nem hibázunk.)

6.2. A statisztikai próba

A **statisztikai hipotézis** egy vagy több valószínűségeloszlásra vonatkozó feltevés.

A feltevés vonatkozhat magára az eloszlásra (normális eloszlású-e a megfigyelt valószínűségi változó, függetlenek-e valószínűségi változók stb.), vagy annak valamely paraméterére (a várható érték adott szám-e, megegyezik-e két valószínűségi változó várható értéke, szórása stb.).

A **statisztikai próba** segítségével döntünk a hipotézisünk elfogadásáról vagy elvetéséről.

Mivel véges minta alapján dolgozunk, a hipotézisvizsgálat végén soha nem modhatjuk azt, hogy a hipotézisünk igaz vagy nem igaz. Annyit mondhatunk, hogy a vizsgált minta alapján a hioptézisünk elfogadható vagy nem fogadható el.

Tulajdonképpen egy hipotézisvizsgálat esetén a sejtésünk alapján megfogalmazzuk, hogy nagy valószínűséggel milyen eredményt várunk a mintavétel során, majd megnézzük, hogy az adott minta (mintából kiszámított érték) megfelel-e az elvárásainknak. (Emlékezzük, hogy pl. a központi határeloszlás tétele alapján meghatároztuk, hogy valószínűségi változók átlagának értékét milyen intervallumba várjuk nagy (pl. $0,95$) valószínűséggel. Ezután már csak az a dolgunk, hogy megnézzük, a mintából kapott átlag belefér-e az adott intervallumba.) Ha nem, az két dolgot jelenthet: a mintából számított érték olyan, amit nagyon kis valószínűséggel várunk a mintavétel során, vagy a feltevésünk hibás volt. Hipotézisvizsgálat során azt mondjuk, hogy inkább az utóbbit hisszük (ami természetesen nem biztos, hogy igaz).

Több száz statisztikai próbát ismerünk, azonban a hipotézisvizsgálat menete, lépései alapvetően ugyanazok mindegyik esetében, eltérés csak a technikai részletekben, a megvalósításban van.

6.2.1. A statisztikai próba elméleti lépései

Tekintsük át, hogy hogyan építünk fel egy statisztikai próbát, melyek az elméleti lépései:

a) **H_0 null- és H_1 ellenhipotézis megfogalmazása**.
Első lépésben megfogalmazzuk a sejtésünket: A feltételezett eloszlásra, illetve annak valamely paraméterére felállítunk egy H_0-lal jelölt nullhipotézist.
Mivel nem tudjuk, hogy a sejtésünk helyes-e, szükséges a H_1 ellenhipotézist is meghatároznunk, mégpedig úgy, hogy a H_0 null- és a H_1 ellenhipotézis egymást kizárják. Közülük pontosan egyet fogadunk el, vagyis ha a nullhipotézist elutasítjuk, azzal az ellenhipotézist elfogadjuk és fordítva.
A H_0 és a H_1 megfogalmazása attól függ, mi a vizsgálat célja, mi a sejtésünk. Ha lehet, a null- és ellenhipotézist úgy fogalmazzuk meg, hogy a nullhipotézis elvetése legyen fontos, szemben a H_0 elfogadásával.

Az alábbi példán nézzük meg, hogyan fogalmazhatjuk meg H_0-t és H_1-et a feltevésünk függvényében.

Példa Egy üzemben 250 mm névleges hosszúságú terméket gyártanak. Azt akarjuk eldönteni, hogy megfelel-e ennek a gyártás (tehát valóban 250 mm a hossz várható értéke). Állítsuk fel a null- és az ellenhipotézist, ha a sejtéseink a következők:

(a) A gyártósor megfelelően van beállítva.

(b) A gyártósor nem megfelelően van beállítva, a termék hosszának várható értéke kevebb, mint 250 mm.

(c) A gyártósor nem megfelelően van beállítva, a termék hosszának várható értéke több, mint 250 mm.

Megoldás: A megfigyelt valószínűségi változó most a termék hossza, ezt jelöljük X-szel. A feltevésünk mindhárom kérdés esetén ennek $E(X) = m_0$ várható értékére vonatkozik.

(a) A nullhipotézis ekkor az, hogy a gyártósor jól van beállítva, a termék hosszának várható értéke 250 mm. Az ellenhipotézis ekkor ennek az ellentettje, vagyis az, hogy a gyártósor rosszul van beállítva, tehát a hossz várható értéke nem 250 mm.
Ekkor hipotéziseinket az alábbi alakba írhatjuk :

$$H_0\colon E(X) = 250 \quad \text{(A hossz várható értéke } m_0 = 250.)$$
$$H_1\colon E(X) \neq 250 \quad \text{(A hossz várható értéke nem 250.)}$$

(b) Ha a feltételezésünk csak egyenlőtlenséget tartalmaz, akkor azt az ellenhipotézisben fogalmazhatjuk meg. Most az a feltevésünk, hogy a hossz várható értéke kevesebb, mint az előírt 250 mm, azaz $E(X) < 250$. Ez egyenlőséget nem tartalmaz, így csak H_1-ben fogalmazhatjuk meg. H_0 ennek az ellentettje: $E(X) \geq 250$. Formálisan:

$$H_0\colon E(X) \geq 250 \quad \text{(A hossz várható értéke legalább } m_0 = 250.)$$
$$H_1\colon E(X) < 250 \quad \text{(A hossz várható értéke kevesebb, mint 250.)}$$

(c) Hasonlóan járunk el, mint az előző esetben. A sejtésünk, hogy a hossz várható értéke több, mint 250 mm, ez csak az ellenhipotézisben fogalmazható meg, a nullhipotézisben azt kell megfogalmaznunk, hogy a hossz várható értéke legfeljebb 250 mm.

$$H_0\colon E(X) \leq 250 \quad \text{(A hossz várható értéke legfeljebb } m_0 = 250.)$$
$$H_1\colon E(X) > 250 \quad \text{(A hossz várható értéke nagyobb, mint 250.)}$$

A gyakorlatban azt mondhatjuk, hogy a **nullhipotézis**t úgy kell felállítani, hogy **mindig tartalmazza az egyenlőséget**, az **ellenhipotézis** pedig soha **ne tartalmazzon egyenlőséget**.

b) **A próbafüggvény kiválasztása**.
A próbafüggvény (jel.:$\hat{\alpha}_n$) egy alkalmasan megválasztott statisztikai függvény, amellyel H_0 fennállását vizsgáljuk.
Gyakorlatilag a próbafüggvény egy valószínűségi változó, mely az egyes minták esetében különböző értéket vesz fel. (Gondoljuk pl. arra, hogy az átlag értéke különböző mintavételeknél más és más lehet.)
A hipotézisvizsgálatnál a feladat feltételeinek megfelelő próbafüggvényt kell kiválasztanunk. (Azt, hogy melyik próbafüggvényt választjuk, sok mindentől függ, például attól, hogy mire vonatkozik a feltevésünk, hány mintaelemünk van, milyen paramétereket ismerünk vagy nem ismerünk, több minta esetén függetlenek-e a minták stb.)
Az előző példa esetében, ismert elméleti szórás (σ) mellett, legalább 30 elemű elemű mintát tekintve a megfelelő próbafüggvény:

$$\hat{\alpha}_n = \frac{\hat{m}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$$

c) **A próbafüggvény eloszlásának meghatározása**.
Ezzel nincs dolgunk, a konkrét próba kiválasztása esetén a nullhipotézis fennállása mellett adott a próbafüggvény eloszlása.
A fenti példa esetében pl. közelítőleg standard normális eloszlású a próbafüggvény. (A központi határeloszlás tétele alapján.)

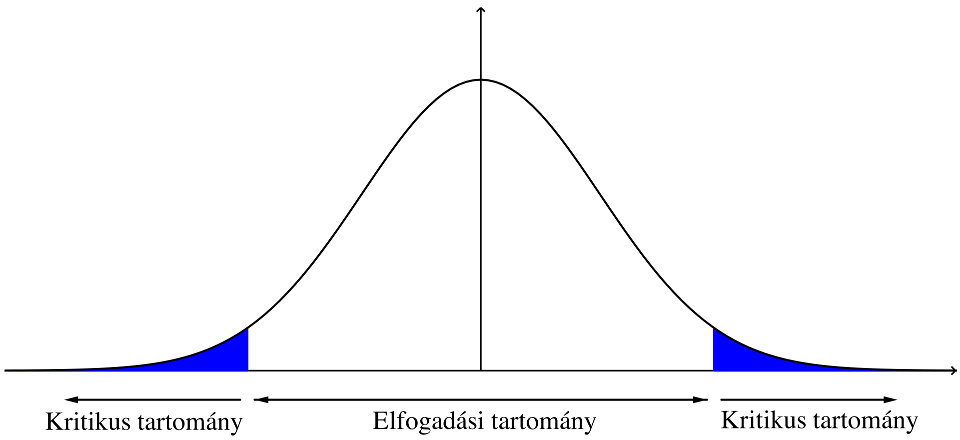
d) **A próba megbízhatósági- vagy szignifikanciaszintjének $((1-\varepsilon)\cdot 100\%)$ megadása**.
Ezt általában százalékos alakban adjuk meg. Azt határozza meg, hogy mekkora valószínűséghez tartozó intervallumot adunk meg, ahova a próbafüggvény értékét az adott eloszlás alapján várjuk.
A szignifikanciaszint általában nagy, így ε értéke kicsi. Ha nincs egyéb kikötés, a szignifikanciaszintet 95%-nak (0, 95-nek) tekinthetjük.

Megjegyzés: Az irodalomban a szignifikanciaszinten gyakran nem az $(1-\varepsilon)\cdot 100\%$ értéket, hanem az $\varepsilon \cdot 100\%$ értéket értik. (Pl. statisztikai szoftvereknél.) Általánosságban azt mondhatjuk, hogy ha a szignifikanciaszintnek kicsi (10% vagy az alatti) értéket adnak meg, akkor $\varepsilon \cdot 100\%$-ot értik rajta.

e) **Kritikus vagy elutasítási tartomány kijelölése**.
Itt határozzuk meg, hogy a próbafüggvény értékét adott valószínűséggel (szignifikanciaszintből) milyen intervallumba várjuk (elfogadási tartomány).
Ebben a lépésben gyakorlatilag a próbafüggvény teljes értékkészletét bontjuk két tartományra. Kritikus (vagy elutasítási) tartománynak azt a részintervallumot nevezzük, ahova $(1-\varepsilon)\cdot 100\%$ szignifikanciaszint mellett a próbafüggvény értéke ε valószínűséggel esik. Az intervallumok határait minden esetben a próbafüggvény eloszlásának ismeretében határozhatjuk meg. A kritikus tartomány komplementere az elfogadási tartomány.
A kritikus tartomány kijelölése azért nagyon fontos, mert ettől függ a hipotéziseinkről való döntés. Az az intervallum, amelybe a próbafüggvény értékét várjuk, lehet zárt intervallum, de lehet egyik határa ∞ vagy $-\infty$ is. Azt, hogy ezek közül melyik a kritikus tartomány, **mindig az ellenhipotézis határozza meg**. Az első példában láttuk, hogy az ellenhipotézis három alakban írható fel, ettől függően beszélhetünk egyoldali- illetve kétoldali próbáról.
Az ellenhipotézistől függően a kritikus tartomány kijelölésében a különböző eseteket a várható értékre vonatkozó, korábban látott nullhipotézis esetén mutatjuk be. A kritikus- és az elfogadási taromány minden esetben a próbafüggvénynek megfelelő eloszlás sűrűségfüggvényének grafikonján jelöljük. (Előfordul, hogy a tartományokat nem adják meg, sőt, nem is ábrázolják, egyszerűen a tartományok határát jelző kritikus értéket hasonlítják össze a próbafüggvény értékének abszolútértékével. A grafikon használata azonban sokkal szemléletesebb, megítélésünk szerint jobban szolgálja az érthetőséget. A továbbiakban minden esetben ábrázolni fogjuk a tartományokat.)

- **Kétoldali próba**. Az ellenhipotézis a feltételezett értéktől bármilyen irányú eltérést ($\neq$) tartalmaz.

$$H_0\colon E(X) = m_0$$
$$H_1\colon E(X) \neq m_0$$

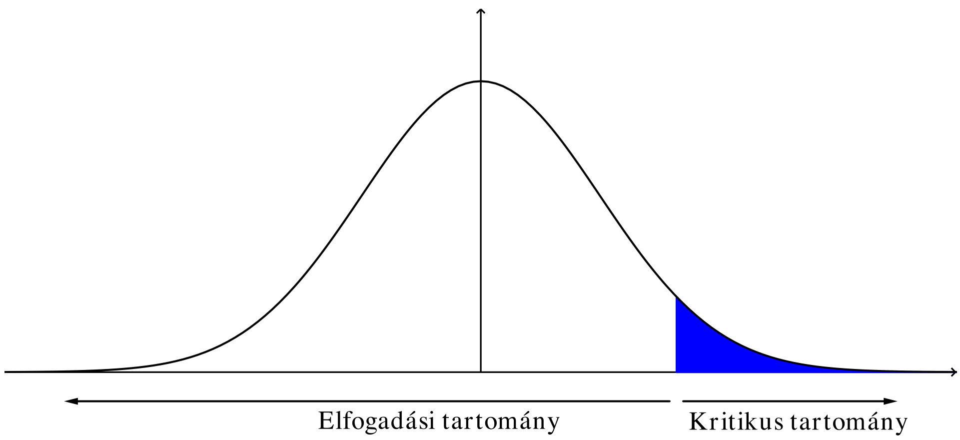


30. ábra. Kritikus tartomány kétoldali próba esetén

- **Egyoldali próba**.
  Az ellenhipotézis a feltételezett értéktől csak egyirányú eltérést tartalmaz. A kritikus tartomány kijelölése az ellenhipotézis egyenlőtlenségének irányától függ.

Jobboldali próbát alkalmazunk, ha az alábbi típusú ellenhipotézis esetén:

$$H_0\colon E(X) \leq m_0$$
$$H_1\colon E(X) > m_0$$

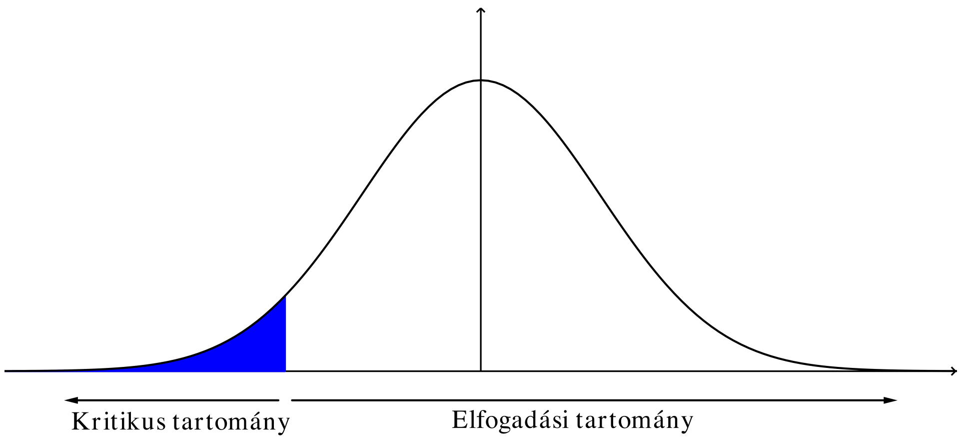


31. ábra. Kritikus tartomány jobboldali próba esetén

Figyeljünk tehát arra, hogy ha az ellenhipotézisben 'nagyobb' szerepel, akkor a kritikus tartomány a jobboldali tartomány!

Baloldali próbát az alábbi típusú ellenhipotézis esetén alkalmazunk:

$$H_0\colon E(X) \geq m_0$$
$$H_1\colon E(X) < m_0$$



32. ábra. Kritikus tartomány baloldali próba esetén

A legtöbb próba egyoldali- és kétoldali kritikus tartomány esetén is elvégezhető, de néhány speciális esetben a próba csak jobb oldali lehet. Ilyen lesz később például az F- és a χ^2-próba is.

f) **Az $\hat{\alpha}_n$ próbafüggvény értékének (számított érték) meghatározása az adott minta alapján.**
A mintából kapott adatok segítségével kiszámoljuk a próbafüggvény értékét. (Ez az érték tehát csak az adott mintához tartozik, másik mintavétel esetén változik (változhat) az értéke.)

g) **Döntés**.
Attól függően, hogy a próbafüggvény számított értéke a kritikus-, vagy az elfogadási tartományba esik, döntünk a H_0 hipotézis elfogadásáról illetve elvetéséről. H_0-t elfogadjuk, ha a számított érték az elfogadási tartományba esik, és elvetjük, ha a számított érték a kritikus tartományba esik.
A H_0 hipotézisről való döntéssel egyben a H_1 hipotézisről is döntünk. Ha H_0-t elfogadjuk, egyben H_1-et elvetjük, ha H_0-t elvetjük, egyben H_1-et elfogadjuk.

Megjegyzés: Ahogy korábban már megjegyeztük, a hipotézisvizsgálat során gyakorlatilag arról döntünk, hogy az ismert eloszlásból nagy, $1 - \varepsilon$ valószínűséggel kijöhet-e az adott minta.

6.2.2. A statisztikai próba gyakorlati lépései

Az elméleti lépések gyakorlati megvalósítása lényegesen kevesebb lépésben megoldható, mivel a fenti lépések a megoldás során összevonhatók.

I A H_0 null- és a H_1 ellenhipotézis felállítása.

II A próbafüggvény kiválasztása és számított értékének meghatározása a minta alapján.

III A kritikus- és elfogadási tartomány kijelölése a próbafüggvény eloszlása és a szignifikanciaszint ismeretében, számított érték elhelyezése a tartományokban.

IV Döntés a H_0 hipotézis elfogadásáról vagy elvetéséről. (Mindig a H_0 elfogadásáról döntünk!!!)
Ebben a lépésben ne feledkezzünk meg a kérdés szavakkal történő megválaszolásáról sem.

Nézzünk példát a kritikus tartomány kijelölésére egy- és kétoldali próba esetén, ha a próbafüggvény standard normális eloszlású:

Példa Határozzuk meg a kritikus tartományt

a) kétoldali,

b) jobboldali,

c) baloldali

próba esetén, 95%- os szignifikanciaszint mellett, ha a próbafüggvény standard normális eloszlású!

Megoldás:

a) Kétoldali próba esetén a hipotéziseink az alábbi alakba írhatók:

$$H_0\colon E(X) = m_0$$
$$H_1\colon E(X) \neq m_0$$

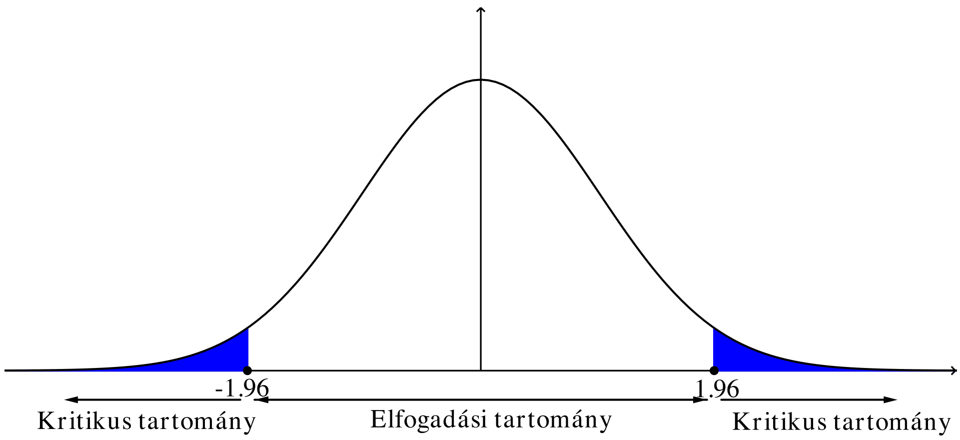
A kritikus tartomány határait jelölje $-u_t$, illetve u_t (kritikus érték). Keressük tehát azt a 0-ra szimmetrikus intervallumot, amelybe egy standard normális eloszlású valószínűségi változó $0,95$ valószínűséggel esik. Legyen $p = 1 - \varepsilon$ jelölést, azaz a szignifikanciaszintet jelölje p. Keressük azt az u_t értéket, melyre teljesül, hogy

$$\Phi(u_t) = \frac{1+p}{2} = \frac{1+0,95}{2} = 0,975$$

A standard normális eloszlás eloszlásfüggvényének értékeit tartalmazó táblázatból kiolvasható, hogy

$$u_t = 1,96$$

A kritikus tartomány:



33. ábra.

b) Jobboldali próba esetén a hipotéziseink:

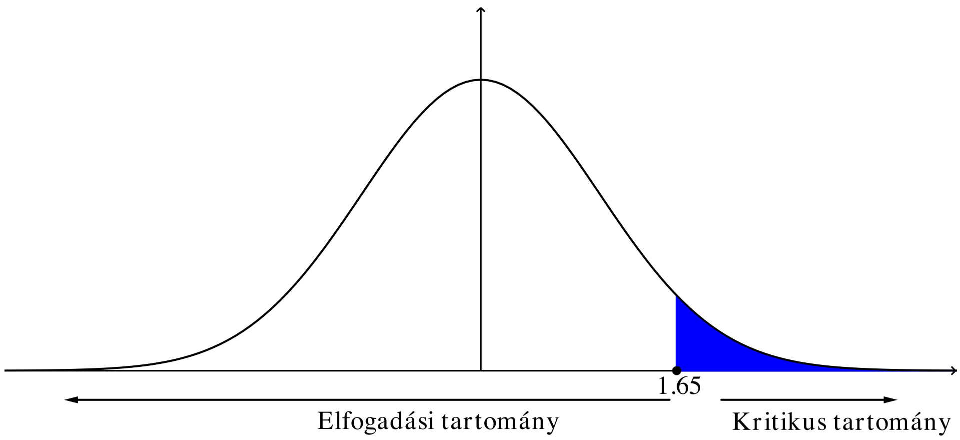
$$H_0\colon E(X) \leq m_0$$
$$H_1\colon E(X) > m_0$$

A kritikus tartomány ekkor $(u_t; \infty)$. Az előző feladatrész jelölését használjuk:

$$\Phi(u_t) = 1 - \varepsilon = p = 0,95$$

A táblázatból

$$u_t = 1,65$$



34. ábra.

c) Baloldali próba esetén érdemes kicsit jobban odafigyelni a kritikus érték előjelére. A hipotéziseink:

$$H_0\colon E(X) \geq m_0$$
$$H_1\colon E(X) < m_0$$

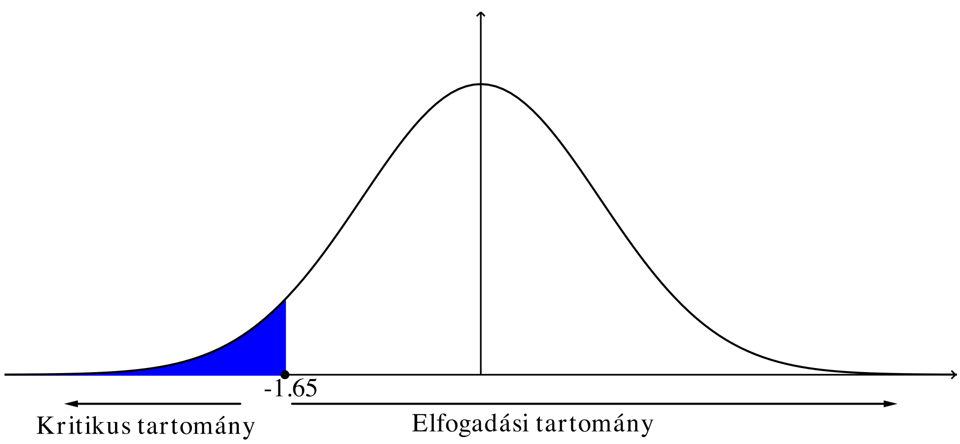
A kritikus tartomány ekkor $(-\infty, -u_t)$. A standard normális eloszlás szimmetriáját kihasználva

$$\Phi(u_t) = 1 - \varepsilon = p = 0,95$$

Ekkor a táblázatból kiolvasott érték **előjelét meg kell változtatnunk**, így

$$u_t = -1,65$$

A kritikus tartomány tehát a $(-\infty; -1,65)$.



35. ábra.

6.2.3. Hipotézisvizsgálat hibalehetőségei

Arra a kérdésre, hogy milyen hibát követhetünk el egy hipotézisvizsgálat során, gyakran hangzik el válaszként, hogy 'rossz próbát választottam', 'elszámoltam' stb. Természetesen ilyen jellegű hibákat is el lehet követni, de hipotézisvizsgálatnál ha a jól kiválasztott próbát jól hajtjuk végre, akkor is juthatunk hibás döntésre. Ennek az oka, hogy egy adott minta alapján döntünk, ami kis valószínűséggel, de rossz döntést eredményezhet (olyan próbafüggvény értékre vezet, ami kicsi eséllyel jön ki).

A hipotézisvizsgálat során (helyesen lefolytatva a próbát) kétféle hibát követhetünk el. **Elsőfajú** a **hiba**, ha a H_0 hipotézis igaz, ennek ellenére elutasítjuk. Ha a hibás H_0 hipotézist elfogadjuk, **másodfajú hibát** követünk el. Összefoglalva:

	H_0 igaz	H_0 hamis
H_0-t elfogadjuk	Helyes a döntés	**Másodfajú hiba**
H_0-t elutasítjuk	**Elsőfajú hiba**	Helyes a döntés

Ezeket a hibákat általában kis valószínűséggel követjük el. Az elsőfajú hiba valószínűsége megegyezik ε-nal, a másodfajú hiba valószínűségét már nem ilyen egyszerű megmondani.
Szokás ezeket a hibákat a bírósági ítéletekhez is hasonlítani. Ha az a feltevésünk (H_0), hogy a vádlott ártatlan, akkor az elítélése elsőfajú hibának, a bűnös vádlott felmentése másodfajú hibának tekinthető. Nyilvánvalóan arra kell törekedni, hogy mindkét típusú hiba valószínűségét minél alacsonyabban tartsuk. Az elsőfajú hiba valószínűségét a szignifikanciaszint növelésével csökkenthetjük (ahogy például gyógyszerkísérletek esetén teszik). A másodfajú hiba valószínűsége a minta elemszámának növelésével csökkenthető. Mindkét hiba valószínűségét azonban egyszerre, adott minta elemszám mellett nem tudjuk csökkenteni. Az elsőfajú hiba csökkentésével nő a másodfajú hiba valószínűsége. A leggyakrabban használt 95%-os szignifikanciaszint kompromisszumnak tekinthető a kétfajta hiba nagyságát illetően.

6.3. Paraméteres próbák I.

Paraméteres próbáról akkor beszélhetünk, ha a H_0 hipotézis és a H_1 ellenhipotézis a megfigyelt valószínűségi változó valamely paraméterére (valószínűség, várható érték, szórás stb.) vonatkozik. Azt mondhatjuk, hogy a legtöbb paraméteres próbának megvan a nemparaméteres párja. Mi szól a paraméteres próbák mellett?

- A paraméteres próbák nagyobb erejűek, mint nemparaméteres párjaik (kisebb a másodfajú hiba valószínűsége).
- A H_0 hipotézis gyakran nagyobb információ tartalommal bír, mint a nemparaméteres esetben.

A paraméteres próbák hátrányai:

- Több feltevést igényelnek, szűkebb körben alkalmazhatóak.
- Csak abban az esetben alkalmazhatóak, ha a mintából az átlagot, szórást tudjuk értelmezni. (Gondoljunk bele, ha azt vizsgáljuk, hogy függ-e a vizsgázó neme és egy vizsga kimenetele egymástól, a sikeres vizsga és sikertelen vizsga 'átlagáról' nincs értelme beszélni.)

Jelen jegyzetben a paraméteres próbák közül a várható értékre vonatkozó hipotézisek ellenőrzésére alkalmas legfontosabb próbákat tekintjük át.
Ismeretlen várható értékű, **normális eloszlású** valószínűségi változó várható értékére vonatkozó

$$H_0\colon\ E(X) = m_0$$

hipotézis helyességét akarjuk ellenőrizni. Keressük a próbafüggvényt.

6.3.1. Egymintás u-próba

Azt szeretnénk eldönteni, hogy egy n elemű, $\hat{m}_n$ mintaátlagú statisztikai minta származhat-e m_0 várható értékű normális eloszlásból.

a) Legyen az X normális eloszlású valószínűségi változó σ **szórása ismert**:
Ha H_0 hipotézis helyes, a mintaelemek m_0 várható értékű, σ szórású normális eloszlású valószínűségi változók. Korábban láttuk, hogy normális eloszlású valószínűségi változók összege is normális eloszlást követ, így az $\hat{m}_n$ valószínűségi változó normális eloszlású, m várható értékkel és $\frac{\sigma}{\sqrt{n}}$ szórással.
Alkalmazzunk a normális eloszlásnál tanult standardizálást. A transzformáció után kapott

$$\hat{u}_p = \frac{\hat{m}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$$

próbafüggvény standard normális eloszlású (ha H_0 fennáll).

Megjegyzés: A központi határeloszlás tétele miatt nagy elemszámú ($n \geq 30$) minta esetén $\hat{u}_p$ a megfigyelt valószínűségi változó eloszlásától függetlenül standard normális eloszlásúnak tekinthető (közelítőleg standard normális eloszlású). Ez azt jelenti, hogy az u-próba a fenti próbafüggvénnyel nagy minta elemszám esetén akkor is alkalmazható, ha a megfigyelt valószínűségi változó nem normális eloszlású.

b) Legyen a X a valószínűségi változó σ **szórása ismeretlen**, de a minta **elemszáma nagy** (($n \geq 30$)).
Továbbra is a központi határeloszlás tételére hivatkozunk, ez alapján a megfigyelt valószínűségi változó eloszlásától függetlenül a mintaátlag közelítőleg normális eloszlást követ. Szeretnénk standard normális eloszlásúvá transzformálni, de ehhez nem ismerjük a szórását. A szórást azonban az $\hat{s}_n$ korrigált tapasztalati szórással, így a mintaátlag szórását az $\frac{\hat{s}_n}{\sqrt{n}}$-nel becsülhetjük. Így az

$$\hat{u}_p = \frac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}}$$

próbafüggvény közel standard normális eloszlású (ha H_0 fennáll).

Megjegyzés: A korrigált és a tapasztalati szórás közötti összefüggést ($\hat{\sigma}_n = \sqrt{\frac{n}{n-1}} \cdot \hat{s}_n$) felhasználva a következőt állíthatjuk:
Ha n nagy, akkor $\sqrt{\frac{n}{n-1}} \approx 1$, tehát $\hat{s}_n \approx \hat{\sigma}_n$. Nagy elemszámú minta esetén tehát a korrigált tapasztalati szórás helyettesíthető a tapasztalati szórással. (Célszerű azonban minden esetben a korrigált tapasztalati szórást alkalmazni, ez minden esetben megfelelő, mivel a 'legjobb' becslés a szórásra.)

1. feladat Egy termék tömege normális eloszlást követ, 1,8 g szórással. Tíz esetben megvizsgáltuk a termék tömegét, a kapott eredmények (g):

9,8	10,1	10,2	9,8	9,7	10,4	10,1	10,4	9,6	10,2

95%-os szignifikanciaszinten elfogadható-e, hogy a termék tömegének várható értéke 10 g?

Megoldás: Ismert a minta elemszáma, a termék tömegének(X) szórása, nem ismert viszont a várható érték. Ezt a mintaátlaggal tudjuk becsülni:

$$\hat{m}_{10} = \frac{9,8 + 10,1 + \ldots + 10,2}{10} = \frac{100,3}{10} = 10,03$$

Így az adataink:

$$n = 10$$

$$\sigma = 1,8$$

$$\hat{m}_{10} = 10,03$$

Kövessük a hipotézisviszgálat gyakorlati lépéseit!

I *A null- és ellenhipotézis kimondása.* Azt feltételezzük, hogy a várható érték $m_0 = 10$. A feladatban nincs kikötés az ettől való eltérés irányára, az ellenhipotézis tehát az, hogy a várható érték nem 10. Formálisan a hipotéziseink így:

$$H_0\colon E(X) = 10$$
$$H_1\colon E(X) \neq 10$$

Az ellenhipotézis alapján a próba kétoldali.

II *A próba kiválasztása és a minta alapján a próbafüggvény értékének kiszámítása.*

A H_0 a várható értékre vonatkozik, a minta normális eloszlásból származik, és a szórás ismert, ezért u-próbát alkalmazunk. Az

$$\hat{u}_p = \frac{\hat{m}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$$

próbafüggvény standard normális eloszlású valószínűségi változó.

A mintából származó adatokat a próbafüggvénybe helyettesítjük:

$$\hat{u}_p = \frac{\hat{m}_n - m_0}{\frac{\sigma}{\sqrt{n}}} = \frac{10,03 - 10}{\frac{1,8}{\sqrt{10}}} = 0,0527$$

A mintából számított érték tehát $\hat{u}_p = 0,0527$.

III *A kritikus érték meghatározása.*

A próba kétoldali u-próba, $p = 0,95$ szignifikanciaszinttel. A korábbiakban láttuk a kritikus tartomány kijelölését kétoldali próba esetén, ha a próbafüggvény standard normális eloszlású.

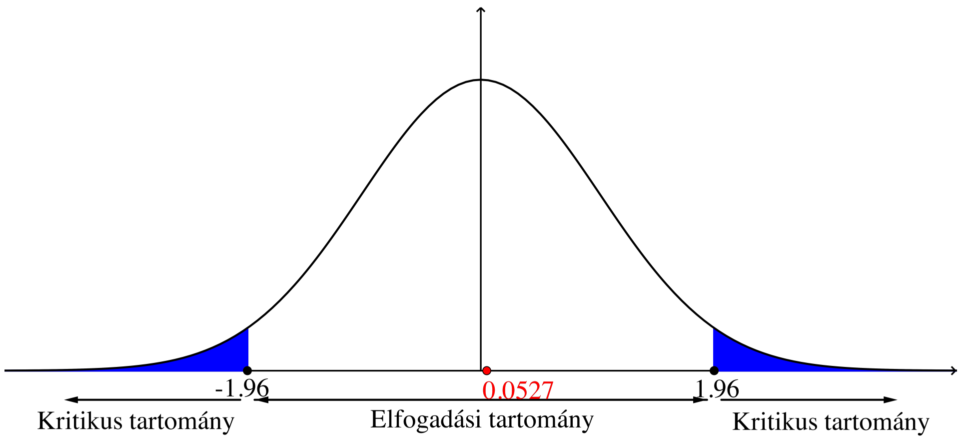
A kritikus érték meghatározásához a

$$\Phi(u_t) = \frac{1 + 0,95}{2} = 0,975$$

egyenlőségnek kell teljesülni. A standard normális eloszlás táblázatából u_t értéke kiolvasható:

$$u_t = 1,96.$$

Az alábbi ábrán látható a kritikus, illetve az elfogadási tartomány, benne a próbafüggvény értéke (pirossal jelölve).



36. ábra.

IV *A döntés.*

A próbafüggvény értéke az elfogadási tartományba esik, ezért a H_0 hipotézist elfogadjuk. 95%-os szignifikanciaszinten a minta nem cáfolja, hogy a valószínűségi változó várható értéke 100.

2. feladat Egy utcai neonlámpa típusról a gyártó cég azt állítja, hogy várható élettartama több, mint 2000 óra. 100 ilyen lámpát teszteltünk. A vizsgálat eredménye: a 100 lámpa élettartamának átlaga 2028 óra, a korrigált tapasztalati szórása 215 óra. (Feltételezzük, hogy a lámpa élettartama normális eloszlást követ.) 95%-os megbízhatósági szinten igaz-e, hogy a lámpa élettartama több, mint 2000 óra?

Megoldás: A mintából adott:

$$n = 100$$

$$\hat{m}_n = 2028$$

$$\hat{s}_n = 215$$

Lássuk a hipotézisvizsgálat lépéseit!

I *A null- és ellenhipotézis kimondása.* A gyártó állítását (az élettartam várható értéke *nagyobb*, mint 2000 óra) az ellenhipotézisben tudjuk csak megfogalmazni. (Emlékezzünk, a nullhipotézis mindig tartalmazza az egyenlőséget!) A nullhipotézis ennek az ellentettje, az élettartam várható értéke legfeljebb 2000 óra.

$$H_0\colon E(X) \leq 2000$$

$$H_1\colon E(X) > 2000$$

Az ellenhipotézis alapján a próba egyoldali, mégpedig jobboldali.

II *A próba kiválasztása és a minta alapján a próbafüggvény értékének kiszámítása.* A minta normális eloszlásból származik, a szórás nem ismert, de a minta elemszáma nagy ($100 \geq 30$), ezért u-próbát alkalmazunk. A próbafüggvény értéke:

$$\hat{u}_p = \frac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}} = \frac{2028 - 2000}{\frac{215}{\sqrt{100}}} = 1,3023$$

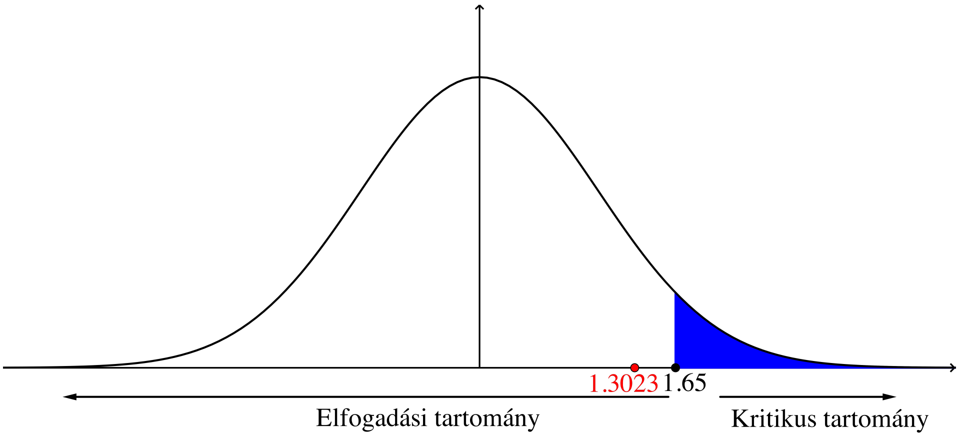
III *A kritikus érték meghatározása.*
A próba egyoldali (jobboldali) u-próba, a kritikus értékre teljesül, hogy

$$\Phi(u_t) = 0,95$$

A standard normális eloszlás táblázatából kiolvasott $1,65$ egyben az u_t értéke (nem kell az előjelen változtatni, mert a próba jobboldali):

$$u_t = +1,65.$$

Ennek alapján az alábbi ábrán látható a kritikus, illetve az elfogadási tartomány, melyben elhelyeztük (pirossal jelölve) a próbafüggvény értékét.



37. ábra.

IV *A döntés.*
A próbafüggvény értéke az elfogadási tartományba esett, a nullhipotézist elfogadjuk 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az ellenhipotézist elvetjük, a kérdésre tehát az a válasz, hogy az adott szignifikanciaszinten nem fogadható el a gyártó állítása, hogy a neonlámpa élettartamának várható értéke több, mint 2000 óra.

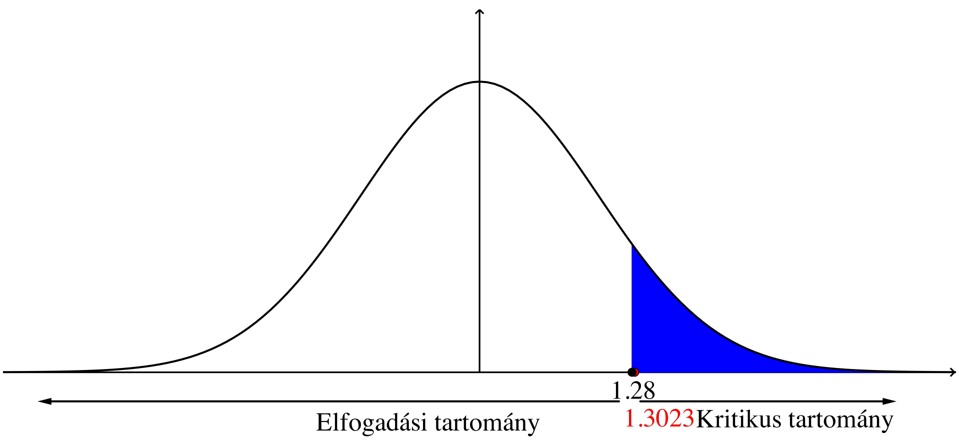
Vizsgáljuk meg, hogy a szignifikanciaszintet 90%- ra változtatva hogyan változik a döntésünk! Ha megnézzük a hipotézisvizsgálat lépéseit, akkor azt látjuk, hogy a szignifikanciaszintnek a kritikus tartomány kijelölésénél van szerepe, tehát az első két lépés nem változik.
Ha a szignifikanciaszintet 90%-ra csökkentjük (és egyben az elsőfajú hiba valószínűségét $0,1$-re növeljük), akkor a kritikus érték:

$$\Phi(u_t) = 0,90$$

így a táblázatból

$$u_t = +1,28$$

Az ábrán látjuk a kritikus tartományt, benne a számított értékkel:



38. ábra. Kritikus tartomány a 2. feladathoz 90%-os szignifikanciaszinttel

Ezzel a számított érték az elutasítási tartományba került, a döntésünk megváltozott. A H_0-t elutasítjuk, a H_1-et elfogadjuk.

Megjegyzés: A szignifikanciaszint megválasztása tehát jelentősen befolyásol(hat)ja a döntésünket. Ugyanakkor a 90%-os szignifikanciaszint esetén azt láttuk, hogy a számított érték és a kritikus érték nagyon közel van egymáshoz. Adott szignifikanciaszint mellett ilyenkor érdemes megismételni a mintavételt és az új mintával lefolytatni a hipotézisvizsgálatot.

3. feladat Egy horgászzsinór teherbírása névlegesen 20 kg. Ennek ellenőrzésére 120 mérést végeztek, melynek eredménye (kg): $\hat{m}_{120} = 19,7$, $\hat{s}_{120} = 1,32$. Igaz-e 95%-os szignifikanciaszinten, hogy a horgászzsinór teherbíró képességének várható értéke kevesebb, mint 20 kg?

Megoldás: Tudjuk, hogy

$$n = 120$$

$$\hat{m}_n = 19,7$$

$$\hat{s}_n = 1,32$$

I *A null- és ellenhipotézis kimondása.*

$$H_0\colon E(X) \geq 20$$

$$H_1\colon E(X) < 20$$

Az ellenhipotézis alapján a próba egyoldali (baloldali).

II *A próba kiválasztása és a minta alapján a próbafüggvény értékének kiszámítása.* A minta normális eloszlásból származik, a szórás nem ismert, de a minta elemszáma nagy ($120 \geq 30$), ezért u-próbát alkalmazunk. A próbafüggvény értéke:

$$\hat{u}_p = \frac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}} = \frac{19,7 - 20}{\frac{1,32}{\sqrt{120}}} = -2,4896$$

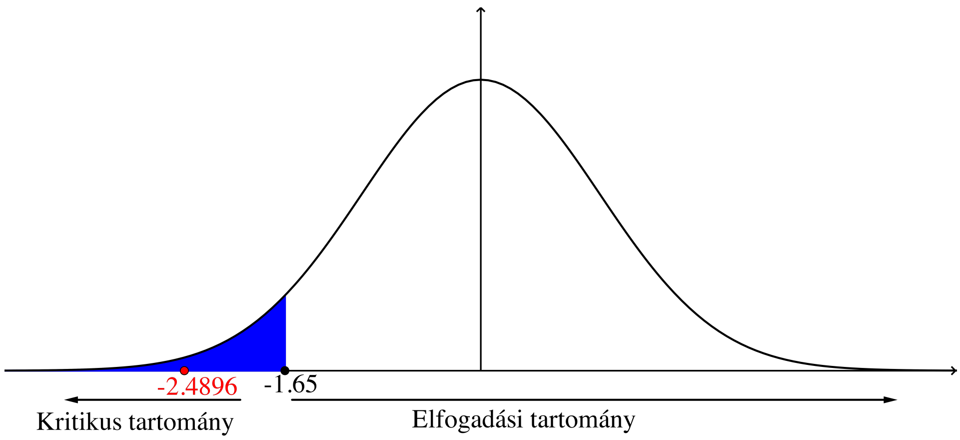
III *A kritikus érték meghatározása.*
A próba egyoldali (jobboldali) u-próba, a kritikus értékre teljesül, hogy

$$\Phi(u_t) = 0,95$$

A standard normális eloszlás táblázatából kiolvasott érték 1.65, de baloldali próbánk van, ezért a kritikus érték ennek a (-1)-szerese:

$$u_t = -1,65.$$

A kritikus, illetve az elfogadási tartomány a következő ábrán látható benne (a pirossal jelölt) a próbafüggvény értéke.



39. ábra.

IV *A döntés.*
A próbafüggvény értéke az elutasítási tartományba esett, a nullhipotézist elutasítjuk 95%-os szignifikanciaszinten. A ellenhipotézist így elfogadjuk, elfogadható, hogy a horgászzsinór teherbíró képessége kevesebb, mint 20 kg.

6.3.2. Kétmintás u-próba

Gyakran a probléma nem egy mintához kapcsolódik, hanem két minta paramétereinek összehasonlítása a feladat. Két különböző gyáregységben gyártott, ugyanolyan típusú motorok élettartamának várható értékeiről akarjuk eldönteni, hogy megegyeznek-e, esetleg az egyik gyáregységben a várható élettartam adott értékkel nagyobb-e. Ilyenkor ha a minták függetlenek, akkor kétmintás próbát alkalmazhatunk. Nézzük meg, milyen feltételek mellett alkalmazhatunk kétmintás u-próbát!

Legyenek X és Y **normális eloszlású**, $E(X) = m_1$ és $E(Y) = m_2$ ismeretlen várható értékű, $D(X) = \sigma_1$ és $D(Y) = \sigma_2$ szórású valószínűségi változók. A várható értékük különbségére vonatkozó

$$H_0\colon m_1 - m_2 = m_0$$

hipotézis helyességét akarjuk ellenőrizni.

- Ha az X és Y valószínűségi változók **szórása**, $D(X) = \sigma_1$ és $D(Y) = \sigma_2$ **ismert**, akkor az

$$\hat{u}_p = \frac{(\hat{m}_{n_1} - \hat{m}_{n_2}) - m_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

  próbafüggvény H_0 fennállása esetén standard normális eloszlást követ.
  n_1 az X megfigyeléséből, n_2 az Y megfigyeléséből származó független minták elemszáma. (Nem kell, hogy megegyezzenek!)

  Megjegyzés: Vegyük észre, hogy ha az **ismert szórások azonosak**, azaz $\sigma_1 = \sigma_2 = \sigma$, akkor a fenti próbafüggvény

$$\hat{u}_p = \frac{(\hat{m}_{n_1} - \hat{m}_{n_2}) - m_0}{\sigma \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

  alakba írható.

- Ha mindkét minta elemszáma nagy ($n_1 \geq 30$ $n_2 \geq 30$), a szórások nem ismertek, de értékük megegyezik ($\sigma_1 = \sigma_2 = \sigma$ nem ismert), akkor a próbafüggvény

$$\hat{u}_p = \frac{1}{\sqrt{\dfrac{\hat{s}^2_{n_1} \cdot (n_1 - 1) + \hat{s}^2_{n_2} \cdot (n_2 - 1)}{n_1 + n_2 - 2}}} \frac{(\hat{m}_{n_1} - \hat{m}_{n_2}) - m_0}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}},$$

H_0 fennállása esetén közelítőleg standard normális eloszlású.

4. feladat Két különböző üzemben ugyanolyan típusú csavart gyártanak. Az első üzemben 20 véletlenszerűen kiválasztott csavar hosszát mértük meg, csavarok hosszának átlagára $2,53$ cm adódott, a második üzemben 25 véletlenszerűen kiválasztott csavar hosszának átlagára $2,48$ cm-t kaptunk. Feltételezzük, hogy a csavarok hossza mindkét üzemben normális eloszlást követ, az elsőben $0,1$ cm, a másodikban $0,08$ cm szórásal. 95%-os szignifikanciaszinten elfogadható-e, hogy a két üzemben gyártott csavarok hosszának várható értéke megegyezik?

Megoldás: A minták adatai (minták elemszáma, a várható értékek és szórások) a feladatból kiolvashatóak.

Első minta:	Második minta:
$n_1 = 20$	$n_2 = 25$
$\hat{m}_1 = 2,54$	$\hat{m}_2 = 2,48$
$\sigma_1 = 0,1$	$\sigma_2 = 0,08$

a) *A null- és ellenhipotézis kimondása.*
Azt feltételezzük, hogy a várható értékek megegyeznek. Ez azt jelenti, hogy a várható értékek különbsége $m_0 = 0$. A hipotéziseink:

$$H_0\colon m_1 - m_2 = 0 \quad \text{(A várható értékek megegyeznek.)}$$
$$H_1\colon m_1 - m_2 \neq 0 \quad \text{(A várható értékek nem egyeznek meg.)}$$

A próba kétoldali.

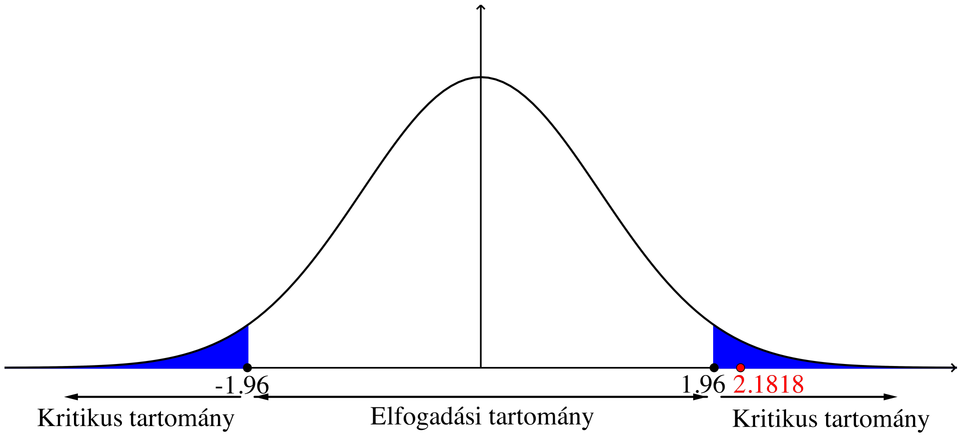
b) *A próba kiválasztása és a minta alapján a próbafüggvény értékének kiszámítása.*
A szórások ismertek, a minták függetlenek, normális eloszlásból származnak, így kétmintás u-próbát alkalmazunk.
Helyettesítsük az adatokat a próbafüggvénybe:

$$\hat{u}_p = \frac{(\hat{m}_{n_1} - \hat{m}_{n_2}) - m_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{(2,54 - 2,48) - 0}{\sqrt{\dfrac{0,1^2}{20} + \dfrac{0,08^2}{25}}} = 2,1818$$

c) *A kritikus érték meghatározása.*
A próba kétoldali, $\Phi(u_t) = \dfrac{1 + 0,95}{2} = 0,975.$
A standard normális eloszlás eloszlásfüggvényének értékeit tartalmazó táblázatból

$$u_t = 1,96.$$

Az alábbi ábrán látható a kritikus, illetve az elfogadási tartomány, a próbafüggvény számított értékével (piros).



40. ábra.

d) *A döntés.*
A H_0 hipotézist elfogadjuk, 95%-os szignifikanciaszinten nem fogadható el, hogy a két üzemben gyártott csavarok hosszának várható értéke azonos.

7. Hipotézisvizsgálat II.

7.1. Paraméteres próbák II.

Mielőtt rátérnénk a hipotézisvizsgálatra, ismerjünk meg egy olyan nevezetes folytonos eloszlást, amelyet a következő részben használni fogunk.

A **Student-eloszlás** vagy más néven **t-eloszlás** elsősorban a statisztikában használatos, standard normális eloszlásból származtatott eloszlás. Eloszlás-, és sűrűségfüggvénye egy n szabadsági foktól függően írható fel, és a normális eloszláshoz hasonlóan nehezen kezelhető függvények.

Az n szabadsági fokú t-eloszlás sűrűségfüggvénye:

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$$

ahol

$$\Gamma(x) = \int\limits_0^\infty e^{-t} \cdot t^{x-1} \mathrm{d}t\,, \qquad x > 0$$

az Euler-féle gammafüggvény.

Student-eloszlás esetén elsősorban az érdekel bennünket, hogy hol vesz fel az eloszlásfüggvény egy adott értéket, így az eloszlásfüggvény inverzének értékeire van szükségünk. Ennek leggyakrabban előforduló értékeit a jegyzet végén található 8.2 *A Student-féle eloszlásfüggvény inverzének értékei t-próbához* táblázatból olvashatjuk ki.

A standard normális eloszláshoz hasonlóan most is az eloszlásfüggvénnyel számolunk, de sűrűségfüggvény grafikonján ábrázolunk. (Emlékezzünk: az eloszlásfüggvény adott helyen felvett értéke megegyezik a sűrűségfüggvény grafikonján az adott helytől balra eső területtel.)
A táblázat használatát a következő példa mutatja.

Példa Adott az f szabadsági fokú t-eloszlású valószínűségi változó. Határozzuk meg, hogy az eloszlásfüggvénye hol veszi fel a p értéket, ha

a) $f = 9$, p=0,95,

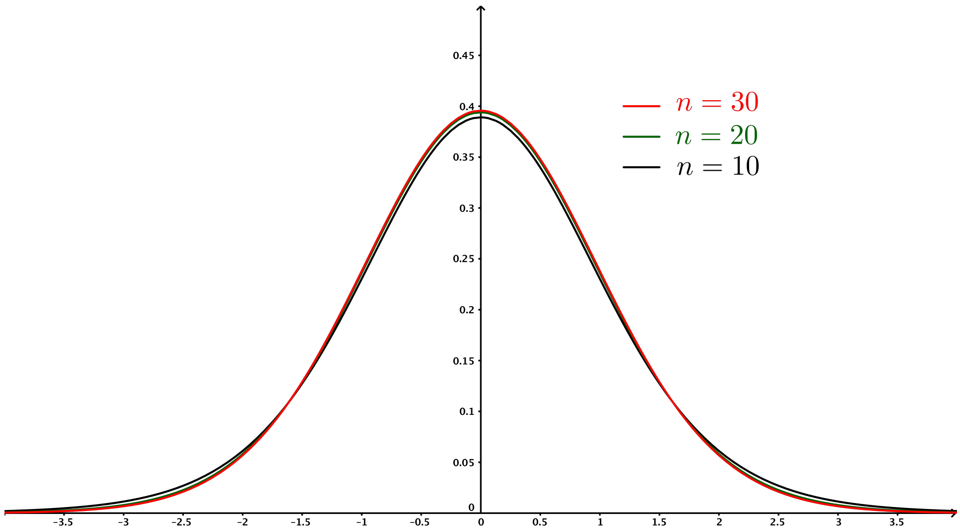
b) $f = 24$, p=0,9750!

Megoldás: A 8.2 *A Student-féle eloszlásfüggvény inverzének értékei t-próbához* táblázat első sorában az eloszlásfüggvény értékei, első oszlopában pedig a szabadsági fok található. Az első oszlopból keressük ki a megadott szabadsági fokot, az első sorból pedig a megadott valószínűséget. (A keresett értéket jelöljünk t_t-vel).

a) $t_t = 1,83$,

b) $t_t = 2,06$.

A *Student*-eloszlás sűrűségfüggvényét különböző szabadsági fokokkal az alábbi ábra mutatja. (Vegyük észre a standard normális eloszláshoz való hasonlóságot, ami a szabadsági fok növelésével egyre nagyobb. $n = 30$-tól a két eloszlás közötti eltérés olyan kicsi, hogy gyakran standard normális eloszlást használunk a *Student*-eloszlás közelítésére.)



41. ábra. Student-eloszlás sűrűségfüggvénye különböző szabadsági fokokkal

Térjünk vissza a hipotézisvizsgálathoz. Továbbra is a várható értékre vonatkozó hipotéziseket akarunk tesztelni. Egyetlen minta esetén korábban láttuk, hogy normális eloszlásból származó valószínűségi változó esetében ismert szórás, illetve nagy minta elemszám esetén u-próbát alkalmazhatunk, standard normális vagy közelítőleg standard normális eloszlású próbafüggvénnyel. Azonban a gyakorlatban sokszor előfordul, hogy a megfigyelt, normális eloszlású valószínűségi változónak egyetlen paraméterét sem ismerjük, ugyanakkor csak kis elemszámú minta áll a rendelkezésünkre. A várható értékre vonatkozó H_0: $E(X) = m_0$ hipotézis vizsgálatára ebben az esetben az u-próba nem alkalmas, helyette t-próbát kell alkalmaznunk.

7.1.1. Egymintás t-próba

A X normális eloszlású valószínűségi változó megfigyelésére n elemű mintát veszünk. Az X σ **szórása nem ismert** és a minta **elemszáma kicsi** ($n < 30$). Ha a H_0: $E(X) = m_0$ hipotézis helyes, akkor a

$$\hat{t}_p = \frac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}}$$

próbafüggvény **(n-1) szabadsági fokú, Student-** vagy más néven t-**eloszlású**.

Megjegyzés: Mivel n kicsi, $\hat{s}_n$ korrigált tapasztalati szórás ebben az esetben nem helyettesíthető a $\hat{\sigma}_n$ tapasztalati szórással.

Az egymintás u- és t-próba alkalmazását (a próbafüggvénnyel és annak eloszlásával) a következő táblázat mutatja:

	σ **ismert**	σ **nem ismert**
$\mathbf{n < 30}$	$\hat{u}_p = \dfrac{\hat{m}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$ standard normális eloszlású	$\hat{t}_p = \dfrac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}}$ Student-eloszlású
$\mathbf{n \geq 30}$	$\hat{u}_p = \dfrac{\hat{m}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$ $\approx$ standard normális eloszlású	$\hat{u}_p = \dfrac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}}$ $\approx$ standard normális eloszlású

Ha csak a próbát akarjuk kiválasztani, a fenti táblázat egyszerű alakja:

	σ **ismert**	σ **nem ismert**
$\mathbf{n < 30}$	u-próba	t-próba
$\mathbf{n \geq 30}$	u-próba	u-próba

Az következő feladatok az egymintás t-próba alkalmazását mutatják:

1. feladat Egy normális eloszlású valószínűségi változó megfigyelésére $n = 20$ mérést végeztek. A minta-átlagra $\hat{m}_{20} = 4972,2$ adódott, a korrigált tapasztalati szórásra $\hat{s}_{20} = 82,4$. 95%-os szignifikanciaszinten elfogadható-e, hogy a valószínűségi változó várható értéke 5000?

Megoldás: A feladatból az adataink:

$$n = 20$$

$$\hat{m}_{20} = 4972,2$$

$$\hat{s}_{20} = 82,4$$

I *A null- és ellenhipotézis kimondása.*
A kérdésből adódóan a feltételezésünk, hogy a valószínűségi változó várható értéke pontosan 5000. Így

$$\begin{aligned} H_0 &: E(X) = 5000 \quad \text{(A várható érték 5000.)} \\ H_1 &: E(X) \neq 5000 \quad \text{(A várható érték nem 5000.)} \end{aligned}$$

Az ellenhipotézis alapján a próba kétoldali.

II *A próba kiválasztása és a minta alapján a próbafüggvény értékének kiszámítása.*
A feltevés a várható értékre vonatkozik, a szórás nem ismert, a minta elemszáma pedig kicsi ($n < 30$), ezért t-próbát alkalmazunk.
A próbafüggvény értéke:

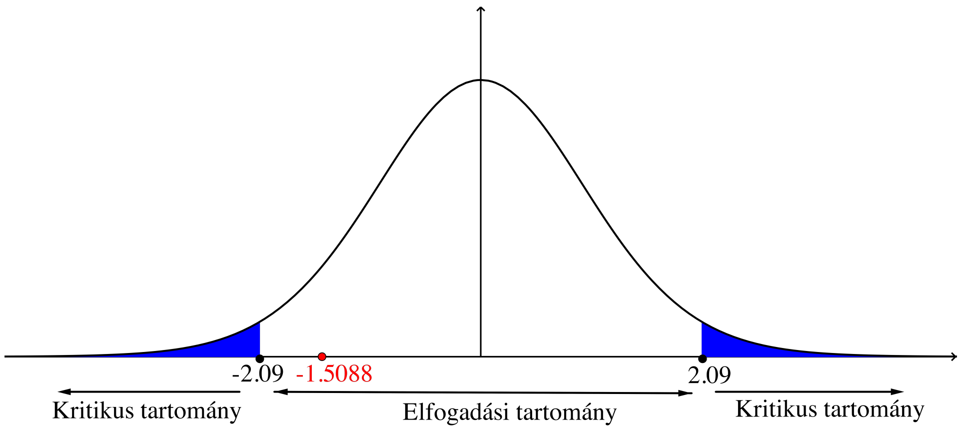
$$\hat{t}_p = \frac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}} = \frac{4972,2 - 5000}{\frac{82,4}{\sqrt{20}}} = -1,5088$$

III *A kritikus érték meghatározása.*

A próbafüggvény Student-eloszlású, $f = n - 1 = 20 - 1 = 19$ szabadsági fokkal. Mivel a próba kétoldali, $p = \dfrac{1 + 0,95}{2} = 0,975$. A *Student*-eloszlás táblázatából

$$t_t = 2,09$$

Az alábbi ábrán látható a kritikus, illetve az elfogadási tartomány, melyben elhelyeztük a próbafüggvény értékét is (a szokásos piros jelöléssel).



42. ábra.

IV *A döntés.*

A próbafüggvény számított értéke az elfogadási tartományba esik, H_0 hipotézist elfogadjuk. 95%-os szignifikanciaszinten elfogadható, hogy a valószínűségi változó várható értéke 5000.

2. feladat Egy gyártósoron egy szerelési feladatra megadott szintidő 20 perc. Az üzem vezetése szerint ennél kevesebb idő is elég erre a feladatra. 10 véletlenszerűen kiválasztott munkás esetében megmérték a feladat elvégzéséhez szükséges időt. A kapott eredmények (perc):

22	17	23	14	19	18	20	21	18	22

Feltételezzük, hogy a minta normális eloszlásból származik. 95%-os szignifikanciaszinten igaza van-e az üzem vezetésének?

Megoldás: Csak a minta elemszáma ismert, a mintaátlagot és a korrigált tapasztalati szórást számolnunk kell a mintaelemekből:

$$\hat{m}_{10} = \frac{22 + 17 + \ldots + 22}{10} = \frac{194}{10} = 19,4$$

$$\hat{s}_{10} = \sqrt{\frac{(22 - 19,4)^2 + (17 - 19,4)^2 + \ldots + (22 - 19,4)^2}{10 - 1}} =$$

$$= \sqrt{\frac{68,4}{9}} = \sqrt{7,6} = 2,7568$$

Az adataink így:

$$n = 10$$

$$\hat{m}_{10} = 19,4$$

$$\hat{s}_{10} = 2,7568$$

I *A null- és ellenhipotézis kimondása.*

A vezetők állítása, hogy a feladat elvégézéséhez szükséges idő várható értéke kevesebb, mint 20 perc. Ezt a H_1 ellenhipotézisben mondjuk ki.

$$\begin{array}{ll} H_0\colon E(X) \geq 20 & \text{(A várható érték legalább 20.)} \\ H_1\colon E(X) < 20 & \text{(A várható érték kisebb, mint 20.)} \end{array}$$

Az ellenhipotézis alapján a próba egyoldali (baloldali).

II *A próba kiválasztása és a minta alapján a próbafüggvény értékének kiszámítása.*

A feltevés a várható értékre vonatkozik, a szórás nem ismert, a minta elemszáma pedig kicsi, így t-próbát alkalmazunk.

A próbafüggvény értéke:

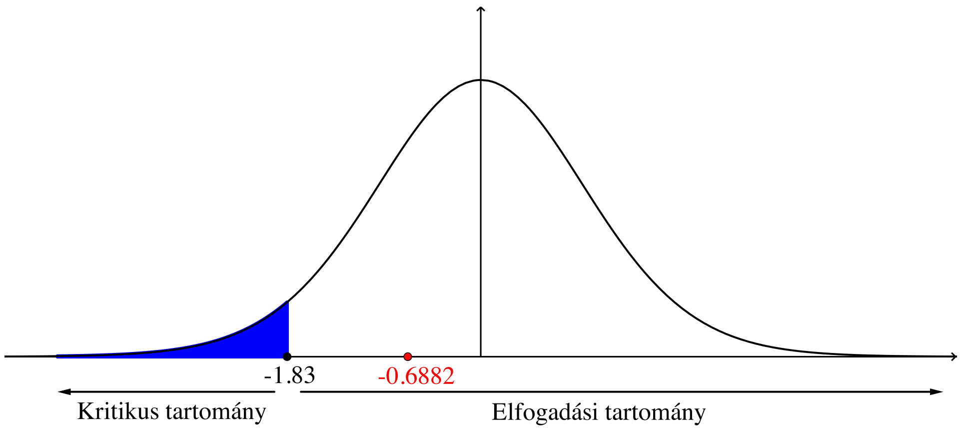
$$\hat{t}_p = \frac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}} = \frac{19,4 - 20}{\frac{2,7568}{\sqrt{10}}} = -0,6882$$

III *A kritikus érték meghatározása.*

A próbafüggvény Student-eloszlású, $f = n - 1 = 10 - 1 = 9$ szabadsági fokkal. Mivel a próba egyoldali, $p = 0,95$. A *Student*-eloszlás táblázatából kiolvasott érték 1,83, az előjelét viszont negatívra kell változtatnunk, mivel a próba baloldali. Igy

$$t_t = -1,83$$

Az alábbi ábrán látható a kritikus, illetve az elfogadási tartomány, melyben elhelyeztük a próbafüggvény értékét is (pirossal).



43. ábra.

IV *A döntés.*

A próbafüggvény számított értéke az elfogadási tartományba esik, H_0 hipotézist elfogadjuk, a H_1-et viszont elutasítjuk. 95%-os szignifikanciaszinten nem fogadható el az üzem vezetésének állítása.

3. feladat Az alábbi minta 7 autó fogyasztási adatait tartalmazza (l/100 km). Az első sorban a szerviz előtti, a másodikban a szerviz utáni értékek találhatók.

Szerviz előtt	7,9	8,1	8,8	7,2	6,0	7,4	8,0
Szerviz után	7,5	7,5	8,1	7,2	5,7	7,6	8,0

Igaz-e 95%-os szignifikancia-szinten, hogy a szerviz csökkenti a fogyasztást?

Megoldás: Figyeljünk arra, hogy a mintából az látszik, hogy az autók többségénél csökkent a fogyasztás a szerviz után. A kérdés azonban az, hogy ez a látható csökkenés ezen a mintán a véletlennek köszönhető-e, vagy a szerviz hatásának tudható be? A mintákat nem tekinthetjük függetlenek, mivel ugyanannak a 7 autónak a fogyasztását mérték, ezért egymintás próbát kell alkalmaznunk. A kérdés valójában a két mérés között bekövetkezett fogyasztásváltozásra vonatkozik. Ehhez a szerviz előtti fogyasztásból kivonjuk mindegyik autó esetén a szerviz utáni fogyasztást, így a hipotézisvizsgálatot az alábbi mintára végezzük el:

fogyasztás csökkenése	0,4	0,6	0,7	0,0	0,3	-0,2	0,0

Az átlagot és a korrigált tapasztalati szórást a mintaelemekből kell számolni:

$$\hat{m}_7 = \frac{0,4 + 0,6 + \ldots + 0,0}{7} = \frac{1,8}{7} = 0,2571$$

$$\hat{s}_7 = \sqrt{\frac{(0,4 - 0,2571-)^2 + (0,6 - 0,2571)^2 + \ldots + (0 - 0,2571)^2}{7-1}} =$$

$$= \sqrt{\frac{0,6771}{6}} = \sqrt{0,1129} = 0,3360$$

Így az adataink:

$$n = 10$$

$$\hat{m}_7 = 0,2571$$

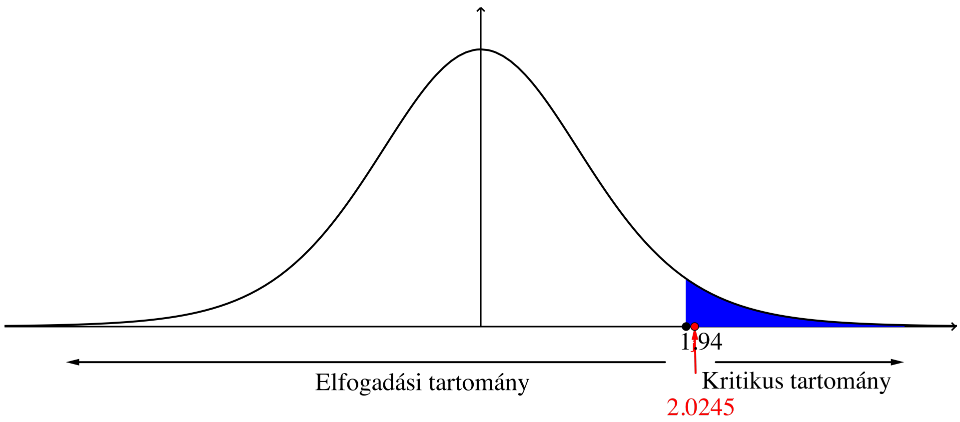
$$\hat{s}_7 = 0,3360$$

I *A null- és ellenhipotézis kimondása.*
A feltevésünk az, hogy a szerviz hatására csökkent a fogyasztás, vagyis a fogyasztás csökkenésének várható értéke nagyobb, mint 0. (Ha a szerviz utáni fogyasztás értékeiből vontuk volna ki a szerviz előttieket, akkor a fogyasztáscsökkenéshez negatív előjel tartozna, a feltevésünk megfordulna.) Az ellenhipotézisben fogalmazhatjuk meg ezt az állításunkat (egyenlőséget nem tartalmaz).

$H_0\colon E(X) \leq 0$ (A szerviz hatására nem csökken a fogyasztás.)
$H_1\colon E(X) > 0$ (A szerviz hatására csökken a fogyasztás.)

II *A próba kiválasztása és a minta alapján a próbafüggvény értékének kiszámítása.*
Feltesszük, hogy a minta normális eloszlásból származik. Mivel a szórás nem ismert, a minta elemszáma pedig kicsi, t-próbát alkalmazunk.

$$\hat{t}_p = \frac{\hat{m}_n - m_0}{\frac{\hat{s}_n}{\sqrt{n}}} = \frac{0,2571 - 0}{\frac{0,3360}{\sqrt{7}}} = 2,0245$$



44. ábra.

III *A kritikus érték meghatározása.*
A próba egyoldali (jobboldali), t_t értéke a 6-szabadsági fokú, Student-eloszlás táblázatból $p = 0,95$ mellett:

$$t_t = 1,94$$

A kritikus, illetve az elfogadási tartomány, a próbafüggvény értékével (piros):

IV *A döntés.*
A próbafüggvény számított értéke a kritikus tartományba esik, a H_0 hipotézist elvetjük, a H_1 ellenhipotézist elfogadjuk. 95%-os szignifikanciaszinten elfogadható, hogy a szerviz csökkenti a fogyasztást. (Ugyanakkor a kritikus és a számított érték nagyon közel van egymáshoz, javasolt a mintavételt és a próbát megismételni.)

A kérdés arra vonatkozott, hogy a fogyasztás látható csökkenése ezen a mintán a véletlennek köszönhető-e, vagy a szerviz hatásának tudható be? A megoldásunk azt mutatja, hogy vélhetően nem véletlen a csökkenés, az a szerviz hatására következett be. (Úgy is fogalmazhatunk, hogy szignifikáns kapcsolat van a fogyasztás csökkenése és a szerviz között.)

7.1.2. Kétmintás t-próba

A várható értékek különbségére az alábbi esetben alkalmazhatunk kétmintás t-próbát: Legyenek X és Y normális eloszlású, ismeretlen, $E(X) = m_1$ és $E(Y) = m_2$ várható értékű, ismeretlen, de azonos $D(X) = \sigma_1 = D(Y) = \sigma_2 = \sigma$ szórású valószínűségi változók. Ha a

$$H_0\colon m_1 - m_2 = m_0$$

nullhipotézis fennáll, akkor a

$$\hat{t}_p = \frac{1}{\sqrt{\dfrac{\hat{s}_{n_1}^2 \cdot (n_1 - 1) + \hat{s}_{n_2}^2 \cdot (n_2 - 1)}{n_1 + n_2 - 2}}} \cdot \frac{\hat{m}_{n_1} - \hat{m}_{n_2} - m_0}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

próbafüggvény $(n_1 + n_2 - 2)$ szabadsági fokú Student-eloszlású.
n_1 az X-re vonatkozó, n_2 az Y-ra vonatkozó független minták elemszáma.
Megjegyzés: Ha n_1 és n_2 nagy, akkor az eloszlás közel standard normálisnak tekinthető, megegyezik az u-próba megfelelő alakjával.

Megjegyzés: A szórások azonosságának ellenőrzésére az F-próbát szokás alkalmazni, ezzel ebben a jegyzetben nem foglalkozunk.

4. feladat Tegyük fel, hogy két normális eloszlású valószínűségi változó szórása egyenlő. Megfigyelésükre $n_1 = 10$ és $n_2 = 8$ elemű mintát vettünk. A mérési adatok:

A	101	103	99	96	105	107	105	98	102	101
B	104	101	97	104	99	103	108	104	-	-

Igaz- e 95%-os szignifikanciaszinten, hogy a két valószínűségi változó várható értéke egyenlő?

Megoldás: A mintákból:

$$n_1 = 11, \quad n_2 = 8$$

$$\hat{m}_1 = \frac{1}{n_1} \cdot \sum_{i=1}^{n_1} X_i = \frac{1017}{10} = 101,7$$

$$\hat{s}_1 = \sqrt{\frac{1}{n_1 - 1} \cdot \sum_{i=1}^{n_1} (X_i - \hat{m}_1)^2} = \sqrt{\frac{106,1}{9}} = 3,4335$$

$$\hat{m}_2 = \frac{1}{n_2} \cdot \sum_{i=1}^{n_2} Y_i = \frac{820}{8} = 102,5$$

$$\hat{s}_2 = \sqrt{\frac{1}{n_2 - 1} \cdot \sum_{i=1}^{n_2} (Y_i - \hat{m}_2)^2} = \sqrt{\frac{82}{7}} = 3,4226$$

Így

$$\begin{array}{ll} n_1 = 10 & n_2 = 8 \\ \hat{m}_1 = 101,7 & \hat{m}_2 = 102,5 \\ \hat{s}_1 = 3,4335 & \hat{s}_2 = 3,4226 \end{array}$$

I *A null- és ellenhipotézis kimondása.*
Feltételezzük, hogy a két valószínűségi változó várható értéke azonos ($m_1 = m_2$), ezért a különbségük 0 ($m_1 - m_2 = 0$). A feltevésünk nullhipotézisként mondható ki (egyenlőség).

$$\begin{array}{ll} H_0\colon m_1 - m_2 = 0 & (\text{ A várható értékek egyenlőek.}) \\ H_1\colon m_1 - m_2 \neq 0 & (\text{A várható értékek nem egyenlőek.}) \end{array}$$

A próba kétoldali.

II *A próba kiválasztása és a minta alapján a próbafüggvény értékének kiszámítása.*
A szórások nem ismertek, de a feltételezés szerint azonosak, a minták elemszáma pedig kicsi, kétmintás t-próbát alkalmazunk.

$$\hat{t}_p = \frac{1}{\sqrt{\frac{\hat{s}_{n_1}^2 \cdot (n_1 - 1) + \hat{s}_{n_2}^2 \cdot (n_2 - 1)}{n_1 + n_2 - 2}}} \cdot \frac{\hat{m}_{n_1} - \hat{m}_{n_2} - m_0}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} =$$

$$= \frac{1}{\sqrt{\frac{3,4335 \cdot 9 + 3,4226 \cdot 7}{10+8-2}}} \cdot \frac{101,7 - 102,5 - 0}{\sqrt{\frac{1}{10} + \frac{1}{8}}} = -0,9108$$

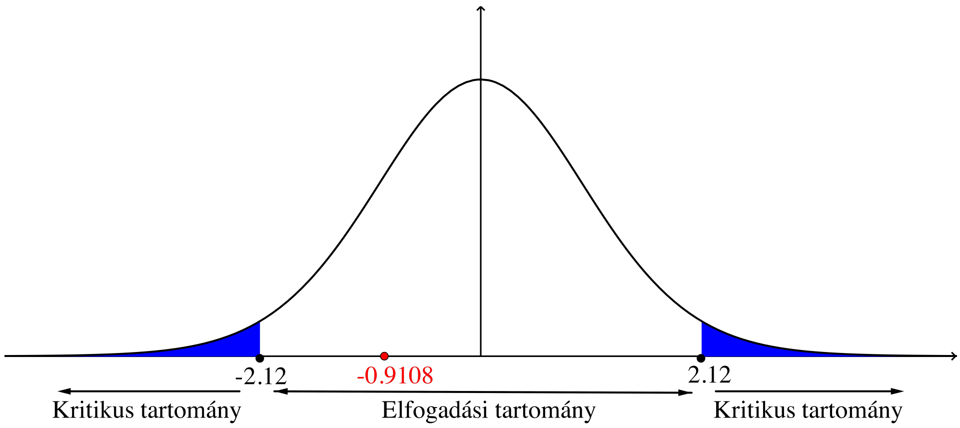
III *A kritikus érték meghatározása.*

A próba kétoldali. A kritikus érték az $n_1 + n_2 - 2 = 10 + 8 - 2 = 16$ szabadsági fokú Student-eloszlás eloszlásfüggvényének inverzének értéke a $\frac{1+0,95}{2} = 0,975$ helyen.

A t-eloszlás táblázatból:

$$t_t = 2,12$$

A következő ábrán látható, hogy a próbafüggvény értéke az elfogadási tartományba esik.



45. ábra.

IV *A döntés.*

A H_0 hipotézist elfogadjuk, 95%-os szignifikanciaszinten a várható értékek egyenlősége elfogadható.

7.2. Nem paraméteres próbák: χ^2-próba

Nem paraméteres próbáról akkor beszélünk, ha a feltevésünk nem a megfigyelt valószínűségeloszlás valamely paraméterére, hanem magára az eloszlásra vonatkozik.

A nem paraméteres próbák előnyei a paraméteresekkel szemben:

- A megfigyelt valószínűségi változó eloszlására nincs kikötés, sőt, nem is kell, hogy egy elméleti eloszlást kövessen. Ezáltal kevesebb feltétel teljesülését várjuk el.
- Nem kell a mintából átlagot, szórást becsülni. Sokszor ez nehezen értelmezhető is lenne, hiszen mit jelent pl. a színvak és nem színvak átlaga?

Hátrány azonban, hogy csak nagyobb eltérések kimutatására alkalmasak. Ha a paraméteres próba feltételei adottak, akkor általában célszerű azt kihasználni.

Az egyik leggyakrabban használt nem paraméteres próba a χ^2-próba. Mielőtt azonban megnézzük az alkalmazását, ejtsünk pár szót a χ^2-**eloszlás**ról:

A *Student*-eloszláshoz hasonlóan standard normális eloszlású valószínűségi változókból (pontosabban azok négyzetösszegéből) származtatjuk a χ^2-eloszlást. Az n szabadsági fokú χ^2-eloszlás sűrűségfüggvénye:

$$f_n(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} & \text{ha } x > 0, \\ 0 & \text{ha } x \leq 0. \end{cases}$$

Egy χ^2-eloszlású valószínűségi változó nem vehet fel negatív értéket. Az (n szabadsági foktól függően felírható) eloszlásfüggvényének értékeit táblázatból olvassuk ki. A statisztikában használatos eloszlás, általában arra vagyunk kíváncsiak, hogy egy adott értéket hol vesz fel az eloszlásfüggvény (azaz az eloszlásfüggvény inverzének értékére). Ábrázolni a sűrűségfüggvény grafikonján fogunk (emlékezzünk megint arra a valószínűségszámítás tanulmányainkból, hogy az eloszlásfüggvény adott helyen felvett értéke megegyezett a sűrűségfüggvény adott helytől balra eső görbe alatti területével).

A χ^2-eloszlás inverzének gyakrabban használt értékei a 8.3 **A χ^2-eloszlásfüggvény inverzének néhány értéke** táblázatban megtalálhatóak. Nézzük, hogyan használjuk a táblázatot!
Példa Adott az f szabadsági fokú χ^2-eloszlású valószínűségi változó. Határozzuk meg, hogy az eloszlásfüggvénye hol veszi fel a p értéket, ha

a) $f = 5$, $p = 0,95$,

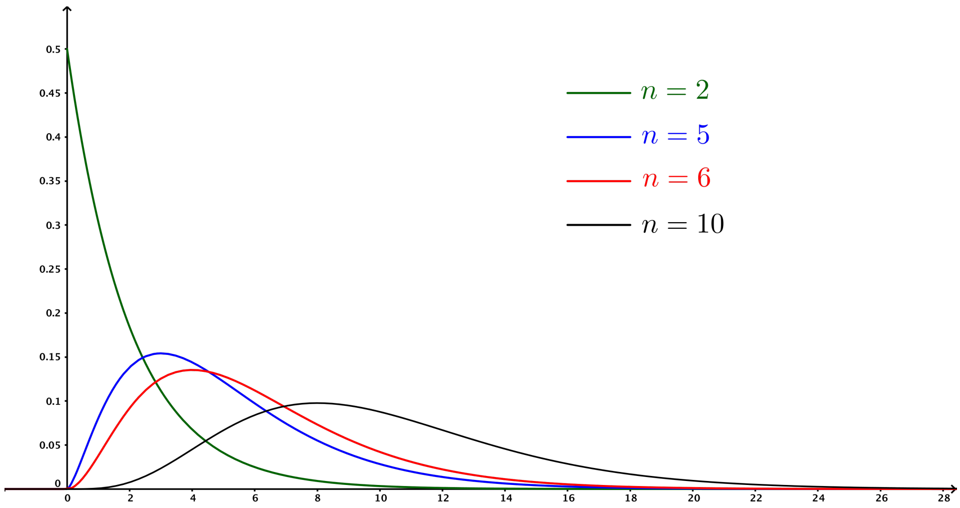
b) $f = 20$, $p = 0,99$!

Megoldás: A 8.3 **A χ^2-eloszlásfüggvény inverzének néhány értéke** táblázat első sorában az eloszlásfüggvény értékei, első oszlopában a szabadsági fok található. Az első oszlopból kikeressük a megadott szabadsági fokot, az első sorból pedig a megadott valószínűséget. A sor és oszlop metszetéből olvassuk le a keresett értéket (amelyet jelöljünk χ_t^2-vel).

a) $\chi_t^2 = 11,07$,

b) $\chi_t^2 = 37,57$.

A sűrűségfüggvény grafikonja különböző szabadsági fokokkal:



46. ábra. χ^2-eloszlás sűrűségfüggvénye különböző szabadsági fokokkal

A χ^2-eloszlás megismerése után nézzünk egy nagyon egyszerű példát a χ^2 próba gondolatmenetének megértéséhez.
Azt szeretnénk eldönteni, hogy egy dobókocka szabályos-e. Ehhez feldobjuk a dobókockát $n = 100$-szor, $r = 6$ különböző kimenetele lehet minden egyes dobásnak. Az egyes dobásértékekhez tartozó, 'megfigyelt' gyakoriságokat jelölje O_i, $i = 1, 2, \ldots 6$. (Ez azt jelenti, hogy O_1 a dobott egyesek száma, O_2 a dobott kettesek száma stb.) Tulajdonképpen a 100 megfigyelésünket $r = 6$ osztályba soroltuk, az egyes osztályokhoz tartozó megfigyelt gyakoriságokat jelöli rendre $O_1, O_2, \ldots, O_6$. Ugyanakkor van egy feltevésünk a kockára vonatkozóan: szabályos a kocka, ami azt jelenti, hogy minden dobásértékhez $\frac{1}{6}$ valószínűség tartozik. Ebből a feltételezett eloszlásból kiszámolhatjuk az egyes dobásértékekhez tartozó, úgynevezett 'elvárt' gyakoriságokat (E_i, $i = 1, 2, \ldots 6$). Ebben az esetben azt, hogy ha a kocka szabályos, akkor a dobások $\frac{1}{6}$ részét várjuk egyesnek ($E_1 = 100 \cdot \frac{1}{6} = 16,6667$), ugyanekkora részét kettesnek ($E_2 = 100 \cdot \frac{1}{6} = 16,6667$) és így tovább. Ezután azt kell eldöntenünk, hogy a megfigyelt gyakoriságok mennyire illeszkednek az elvárt gyakoriságokra. A köztük lévő különbség betudható-e nagy valószínűséggel a véletlen hatásának?

A χ^2**-próba**:
Legyen a H_0 hipotézis az, hogy a megfigyelt és az elvárt értékek illeszkednek. A megfigyelt gyakoriságok (O_i) és elvárt gyakoriságok (E_i) (i=1,2,...,r) összehasonlítására szolgáló

$$\hat{\chi}_p^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}$$

próbafüggvény közelítőleg χ^2-eloszlású, ha az O_i és E_i gyakoriságok nagyok és H_0 fennáll.
A χ^2-próba csak nagy mintaelemszám esetén ad megbízható eredményt, kis- és közepes minták esetén ne alkalmazzuk!

Fontos megjegyezni, hogy a χ^2-próba mindig jobboldali próba (nézzük meg, hogy a próbafüggvény soha nem vehet fel negatív értéket). Ha $\hat{\chi}_p^2 = 0$, akkor a megfigyelt és az elvárt értékek pontosan illeszkednek, ha $\hat{\chi}_p^2 > 0$, akkor nem pontos az illeszkedés. Ekkor minél nagyobb a számított érték, annál nagyobb az eltérés.

Megjegyzés: Amikor a χ^2-eloszlást alkalmazzuk, egy folytonos eloszlással közelítünk diszkrét eloszlást. Ha az E_i elvárt értékek bármelyike 5-nél kisebb, akkor a próba nem alkalmazható. Ezt azonban az osztályok összevonásával kiküszöbölhetjük. A feladatok megoldásánál erre is nézünk példát. Ugyanígy megnézzük, hogy hogyan számolhatjuk ki a χ^2-eloszlás szabadsági fokát a különböző feladatoknál. (Általánosságban a független események számából vonjuk ki a becsült paraméter(ek) számát (ha van ilyen).)

A χ^2-próba alkalmazása:

7.2.1. Illeszkedésvizsgálat

Azt akarjuk ellenőrizni, hogy a megfigyelt minta illeszkedik-e egy adott eloszlásra. Ha az eloszlás paramétereit ismerjük, akkor **tiszta illeszkedésvizsgálat**ról, ha az eloszlás egy- vagy több paraméterét a mintából becsüljük, akkor **becsléses illeszkedésvizsgálatról** beszélünk.

5. feladat (**tiszta illeszkedésvizsgálat**) Egy dobókockáról szeretnénk eldönteni, hogy szabályos-e. A vizsgálathoz 300-szor feldobtuk a dobókockát. Az eredményeket az alábbi táblázat tartalmazza:

dobás értéke	1	2	3	4	5	6	Összes
gyakoriság	41	55	42	43	54	65	300

Elfogadható-e 95%-os szignifikancia szinten, hogy a dobókocka szabályos?

Megoldás: Feltételezésünk az, hogy a dobókocka szabályos. Ez azt jelenti, hogy mindegyik számot azonos valószínűséggel dobjuk, azaz egyes számok dobásának valószínűsége $\frac{1}{6}$.

A táblázat az $n = 300$ kísérlet eredményét $r = 6$ osztályba sorolta, a dobás értéke alapján. Az egyes osztályokhoz tartozó megfigyelt gyakoriság $(O_i, i = 1 \ldots r)$ a táblázat második sorában található. (Pl. $O_2 = 55$, azaz a kísérlet során 55-ször dobtunk kettest. Ez tulajdonképpen azt jelenti, hogy az az esemény, hogy kettest dobunk, a kísérlet során 55-ször következett be.)

A feladatunk most, hogy számítsuk ki az elvárt gyakoriságokat $(E_i, i = 1 \ldots r,)$, azaz azt, hogy ha helyes a feltételezésünk, akkor az adott osztályba a kísérlet hány kimenetelét várjuk. (Pl. E_2 azt jelenti, hogy hány esetben várjuk, hogy kettest dobunk, ha a kocka szabályos.) A dobás értékét tekintjük az X valószínűségi változónak. Ha helyes a feltételezésünk, akkor X valószínűségi változó minden lehetséges értékéhez $p = \frac{1}{6}$ valószínűség tartozik. A hipotézisvizsgálat során azt kell eldöntenünk, hogy a megfigyelt valószínűségi változó (a dobás értéke), a minta alapján tekinthető-e ugyanilyen eloszlásúnak. Ennek alapján a próba illeszkedésvizsgálat. Mivel az eloszlás paramétereit ismerjük, nem a mintából kell becsülnünk, ezért tiszta illeszkedésvizsgálatot hajtunk végre.

Lássuk tehát a várt gyakoriságokat! Számítsuk ki a dobás értékéhez tartozó valószínűségeket (azaz írjuk fel X eloszlását)!

$$p_1 = P(X = 1) = \frac{1}{6}$$
$$p_2 = P(X = 2) = \frac{1}{6}$$
$$p_3 = P(X = 3) = \frac{1}{6}$$
$$p_4 = P(X = 4) = \frac{1}{6}$$
$$p_5 = P(X = 5) = \frac{1}{6}$$
$$p_6 = P(X = 6) = \frac{1}{6}$$

A kiszámolt valószínűségek alapján meg tudjuk mondani, hogy mennyi az egyes osztályokba eső mintaelemek várt gyakorisága. Ehhez nincs más dolgunk, mint a fenti valószínűségekkel szorozni a mintaelemek számát ($E_i = n \cdot p_i, i = 1 \dots r$). (Tehát pl. a 300 kísérletből $E_2 = n \cdot p_2 = \frac{1}{6} \cdot 300 = 50$ esetben várunk kettest.)

Az áttekinthetőbb számoláshoz az alábbi táblázatot hozzuk létre:

dobás értéke	megfigyelt gyakoriság (O_i)	várt gyakoriság (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	41	50	81	1,62
2	55	50	25	0,5
3	42	50	64	1,28
4	43	50	49	0,98
5	54	50	16	0,32
6	65	50	225	4,5
összeg	300	300		**8,2**

A hipotézisvizsgálat lépései ezután az alábbiak:

I *A nullhipotézis kimondása.*

A feltételezésünk most az, hogy a dobókocka szabályos, a dobás értékének eloszlása: $P(X = k) = \frac{1}{6}$, $k = 1, 2, \dots, 6$.

II *A próbafüggvény kiválasztása és értékének kiszámítása.*

A próba tiszta illeszkedésvizsgálat, χ^2-próbával. A próbafüggvény:

$$\hat{\chi}_p^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

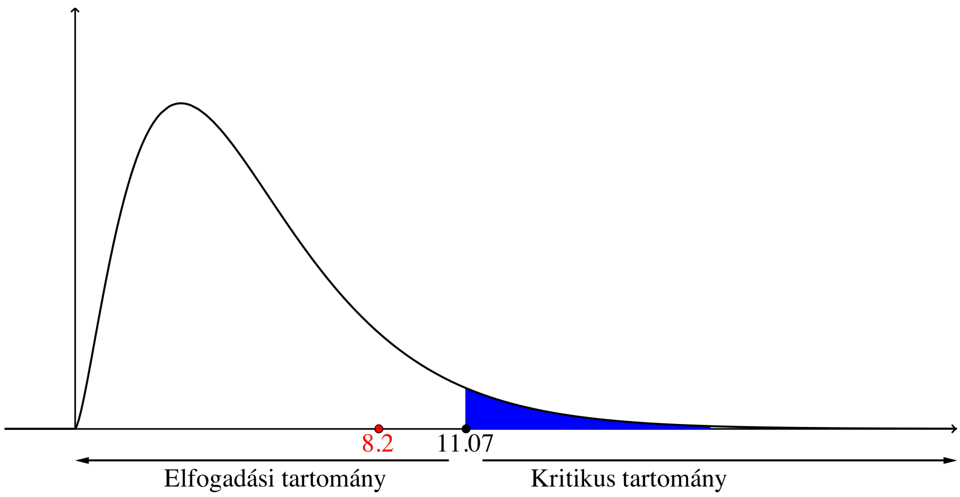
melynek mintásból számított értéke a táblázatunk utolsó sorában szereplő értékek összege, azaz $8,2$. Tehát

$$\hat{\chi}_p^2 = 8,2$$

III *A kritikus tartomány kijelölése.*
A próba jobboldali χ^2-próba. A kritikus értéket a χ^2-eloszlás táblázatából olvashatjuk ki. A szabadsági fok az osztályok száma mínusz 1, vagyis $f = r - 1 = 6 - 1 = 5$, $0,95$ valószínűség mellett. Ez a jegyzet χ^2-eloszlás táblázatában az ötödik sor, első oszlopban található

$$\chi_t^2 = 11,07$$

érték. (A kritikus tartomány ábráján a korábban megszokott jelölésekkel látható, hogy a próba-



47. ábra.

függvény értéke az elfogadási tartományba esik.)

IV *A döntés.*
$8,2 < 11,07$, a próbafüggvény értéke az elfogadási tartományba esik, a nullhipotézist elfogadjuk 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján elfogadható, hogy a dobókocka szabályos.

6. feladat (**tiszta illeszkedésvizsgálat**) Négy pénzérmét 160-szor feldobunk. A kapott gyakoriságok:

fejek száma	0	1	2	3	4	Összes
gyakoriság	5	35	67	41	12	160

Elfogadható-e 95%-os szignifikancia szinten, hogy az érmék szabályosak?

Megoldás: Feltételezésünk az, hogy az érmék szabályosak (mindegyiken $\frac{1}{2}$ valószínűséggel dobunk fejet). A mintaelemek száma $n = 160$ (160 kísérletet hajtottunk végre), a mintaelemeket a dobott fejek száma szerint osztályoztuk, $r = 5$ csoportba soroltuk.
A dobott fejek számát tekintjük az X valószínűségi változónak. Ha helyes a feltételezésünk, akkor X binomiális eloszlású valószínűségi változó $n^* = 4$, $p = \frac{1}{2}$ paraméterekkel. (n^* a dobások száma egy kísérletben.) A paraméterek elméleti értékek, ebben a feladatban nem a mintából becsüljük a várható értéket, így tiszta illeszkedésvizsgálatot hajtunk végre.
Először meg szeretnénk határozni azt, hogy ha az érmék szabályosak, akkor az egyes osztályokba hány mintaelemet várunk.
Számítsuk ki tehát a várt gyakoriságokat ($E_i, i = 1 \dots r,$), vagyis azt, hogy ha helyes az eloszlásra vonatkozó feltételezésünk, akkor az adott osztályba a kísérlet hány kimenetelét várjuk. (Pl. E_1 azt

jelenti, hogy az első osztályba a kísérlet hány kimenetelére számítunk, azaz hány esetben várjuk, hogy a négy érméből 0 fejet dobunk.)

Számítsuk ki a dobott fejek számához tartozó valószínűségeket (azaz írjuk fel X eloszlását)!

$$p_1 = P(X = 0) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$
$$p_2 = P(X = 1) = \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 \binom{4}{1} = \frac{4}{16}$$
$$p_3 = P(X = 2) = \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 \binom{4}{2} = \frac{6}{16}$$
$$p_4 = P(X = 3) = \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 \binom{4}{3} = \frac{4}{16}$$
$$p_5 = P(X = 4) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

A fenti valószínűségekkel szorozva a mintaelemek számát (n), megkapjuk a várt gyakoriságokat ($E_i = n \cdot p_i$, $i = 1 \ldots r$). (Tehát pl. a 160 kísérletből $E_1 = n \cdot p_1 = \frac{1}{16} \cdot 160 = 10$ esetben várunk olyan eredményt, ahol a négy dobás egyike sem fej).

Nézzük a táblázatot!

dobott fejek száma	megfigyelt gyakoriság (O_i)	várt gyakoriság (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	5	10	25	2,62
1	35	40	25	0,625
2	67	60	49	0,8167
3	41	40	1	0,025
4	12	10	4	0,4
összeg	160	160		**4,3667**

A hipotézisvizsgálat lépései ezután az alábbiak:

I *A nullhipotézis kimondása.*
A feltételezésünk most az, hogy az érmék szabályosak, a dobott fejek száma binomiális eloszlású, $(4, \frac{1}{2})$ paraméterekkel.

II *A próbafüggvény kiválasztása és értékének kiszámítása.*
A próba tiszta illeszkedésvizsgálat, χ^2-próbával. A próbafüggvény:

$$\hat{\chi}_p^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i}$$

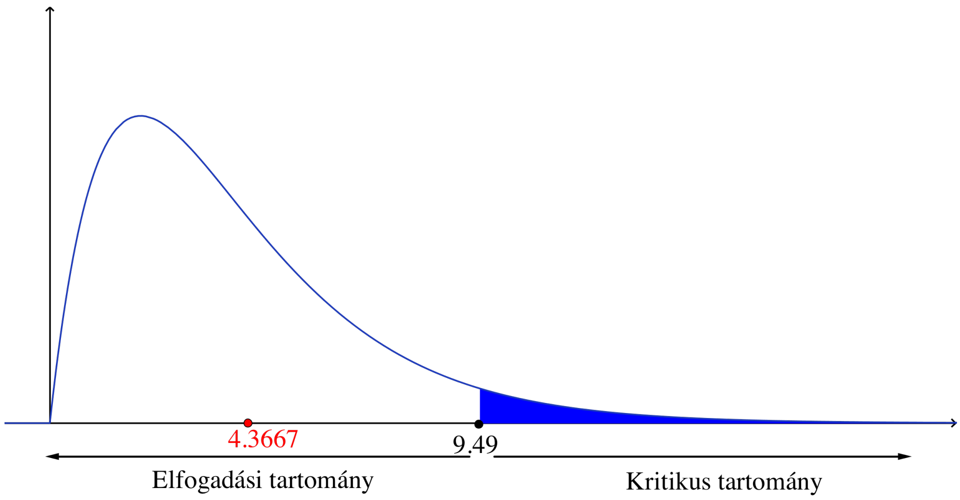
melynek mintából számított értéke a táblázatunk utolsó sorában szereplő értékek összege, azaz $4,3667$:

$$\hat{\chi}_p^2 = 4,3667$$

III *A kritikus tartomány kijelölése.*
A próba egyoldali χ^2-próba. A kritikus értéket a χ^2-eloszlás táblázatából olvashatjuk ki, $r-1=5-1=4$ szabadsági fok (osztályok száma mínusz 1) és $0,95$ valószínűség mellett. Ez a jegyzet χ^2-eloszlás táblázatában a negyedik sor, első oszlopban található

$$\chi_t^2 = 9,49$$

érték.



48. ábra.

(A kritikus tartomány ábráján a korábban megszokott jelölésekkel látható, hogy a próbafüggvény értéke az elfogadási tartományba esik.)

IV *A döntés.*
$4,3667 < 9,49$, a próbafüggvény értéke az elfogadási tartományba esik, a nullhipotézist elfogadjuk 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján elfogadható, hogy az érmék szabályosak.

7. feladat (**becsléses illeszkedésvizsgálat**) Egy gyártósornál rendszeresen 5 elemű mintát vesznek a termékekből. Egy hét alatt 500 mintát vettek. A mintákban talált selejtek gyakorisága az alábbi volt:

selejtek száma	0	1	2	3	4	5
gyakoriság	170	180	120	20	8	2

Modellezhető-e a mintában levő selejtek száma olyan binomiális eloszlással, melynek várható értéke a fentiekből számolt átlag?

Megoldás:
Az egyes mintákban levő selejtek számát tekintjük az X valószínűségi változónak. Azt feltételezzük, hogy X binomiális eloszlású valószínűségi változó, melynek **várható értékét** a mintából(!) számított **átlaggal becsüljük** (becsléses illeszkedésvizsgálatot hajtunk végre). Tehát a várható érték kiszámításához az összes kihúzott selejt számát osztjuk a kísérletek számával (kiszámoljuk az átlagot):

$$\overline{x} = \frac{0\cdot 14 + 1\cdot 180 + 2\cdot 120 + 3\cdot 20 + 4\cdot 8 + 5\cdot 2}{170+180+120+20+8+2} = \frac{522}{500} = 1,044$$

A mintaelemek száma $n = 500$ (500 kísérletet hajtottunk végre), a mintaelemeket a kihúzott selejtek

száma szerint osztályoztuk, ezzel $r = 6$ osztályba soroltuk.
Ha X-t binomiális eloszlással akarjuk modellezni, ismernünk kell a binomiális eloszlás két paraméterét, az n kísérletszámot, amely 5, illetve a megfigyelt esemény (selejthúzás) p valószínűségét. Ez utóbbi nem ismert, de a binomiális eloszlás várható értékéből becsülhetjük: az $1,044 = E(X) = n \cdot p$ egyenletből

$$p = \frac{E(X)}{n} = \frac{1,044}{5} = 0,2088$$

Az illeszkedésvizsgálathoz először meghatározzuk, hogy ha X binomiális eloszlású, $n = 5$, $p = 0,2088$ paraméterekkel, akkor az egyes osztályokba átlagosan hány mintaelemet várunk, azaz kiszámítjuk a várt gyakoriságokat ($E_i, i = 1, \dots, 6$). Számítsuk ki a kihúzott selejtek számának lehetséges értékeihez tartozó valószínűségeket! (Kiszámoljuk, hogy mekkora annak a valószínűsége, hogy az 5 elemű mintában éppen k darab selejt van, ahol $k = 0, 1, 2, 3, 4, 5$. Célszerű minél több tizedes pontossággal számolni!).

$$\begin{aligned}
p_0 &= P(X=0) = 0,7912^5 = 0,31005 \\
p_1 &= P(X=1) = 0,2088^1 \cdot 0,7912^4 \binom{5}{1} = 0,40912 \\
p_2 &= P(X=2) = 0,2088^2 \cdot 0,7912^3 \binom{5}{2} = 0,21593 \\
p_3 &= P(X=3) = 0,2088^3 \cdot 0,7912^2 \binom{5}{3} = 0,05699 \\
p_4 &= P(X=4) = 0,2088^4 \cdot 0,7912^1 \binom{5}{4} = 0,00752 \\
p_5 &= P(X=5) = 0,2088^5 = 0,000397
\end{aligned}$$

A fenti valószínűségekkel szorozzuk a mintaelemek számát, így megkapjuk az i-edik osztályhoz tartozó elvárt gyakoriságot várt gyakoriságok ($E_i, i = 1, \dots, 6$) értékét (így pl. az 500 kísérletből átlagosan $E_2 = 500 \cdot 0,40912 = 204,56$ esetben kapunk olyan eredményt, ahol az 5 kihúzott termékből pontosan egy selejtes). Az elvárt gyakoriságokat beírjuk a táblázatba:

selejtek száma	megfigyelt gyakoriság (O_i)	várt gyakoriság (E_i)
0	170	155,03
1	180	204,56
2	120	107,97
3	20	28,49
4	8	3,75
5	2	0,2
összeg	500	500

Az elvárt gyakoriságok között azonban két 5-nél kisebb érték is szerepel, ami illeszkedésvizsgálatnál nem megengedett! Ahhoz, hogy az illeszkedésvizsgálatot megfelelően el tudjuk végezni, **cellákat kell összevonnunk**. Mivel az utolsó két osztályban a várt gyakoriságok összege $(3,75 + 0,2 = 3,95)$ is kisebb 5-nél, ezért az utolsó három osztályt vonjuk össze. Ezzel az osztályok száma 4-re csökkent. Ennek megfelelően a táblázatot az alábbiak szerint töltjük ki:

selejtek száma	megfigyelt gyakoriság (O_i)	várt gyakoriság (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	170	155,03	224,1009	1,4455
1	180	204,56	603,1936	2,9487
2	120	107,97	144,7209	51,3404
3 v. 4 v. 5	30	32,44	5,9536	0,1835
összeg	500	500		**5,9181**

A hipotézisvizsgálat lépései ezután az alábbiak:

I *A nullhipotézis kimondása.*
A feltételezésünk most az, hogy X binomiális eloszlású az $n = 5$ és $p = 0,2088$ paraméterekkel.

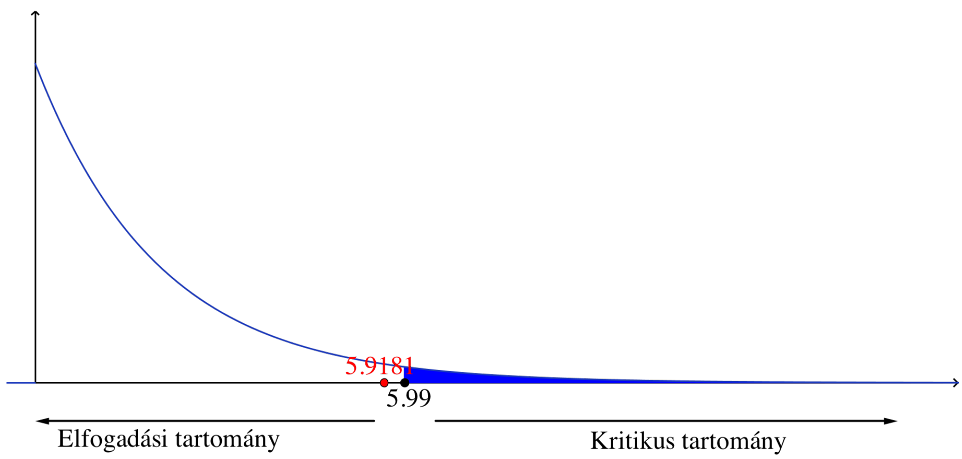
II *A próbafüggvény kiválasztása és értékének kiszámítása.*
A próba illeszkedésvizsgálat, χ^2-próbával. A próbafüggvény mintából számított értéke a táblázatunk utolsó sorában szereplő értékek összege, azaz $5,9181$. Így

$$\hat{\chi}_p^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 5,9181$$

III *A kritikus tartomány kijelölése.*
A próba egyoldali χ^2-próba. A kritikus értéket a χ^2-eloszlás táblázatból olvashatjuk ki. A **szabadsági fok a várható érték mintából való becslése miatt** osztályok száma -2, azaz $4 - 2 = 2$. (Ahány paramétert becsülünk a mintából, annyival csökken a szabadsági fok!) A szignifikancia szint nem volt megadva, ezért a valószínűséget a szokásos $0,95$-nek tekintjük. Ez a χ^2-eloszlás táblázatában a második sor, első oszlopban található

$$\chi_t^2 = 5,99$$

érték.



49. ábra.

(A kritikus tartomány ábráját felrajzolva látható, hogy a kritikus érték az elfogadási tartományba esik.)

IV *A döntés.*

$5.9181 < 5.99$, a próbafüggvény értéke az elfogadási tartományba esik, a nullhipotézist elfogadjuk 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján elfogadható, hogy dobott fejek száma binomiális eloszlású, $n = 5$, $p = 0,2088$ paraméterekkel.
Ugyanakkor vegyük észre, hogy a számított és a kritikus érték nagyon közel van egymáshoz!

7.2.2. Függetlenségvizsgálat

Azt akarjuk eldönteni, hogy két valószínűségi változó, X és Y függetlenek-e. (Ehhez tulajdonképpen az események függetlenségének definícióját használjuk fel: két esemény független, ha az együttes bekövetkezésük valószínűsége megegyezik a valószínűségeik szorzatával. Ezt a gyakoriságokra, illetve relatív gyakoriságokra alkalmazzuk.) A számítás áttekinthetőségéhez úgynevezett **kontingencia-táblákat** használunk.

8. feladat (**függetlenségvizsgálat**) Egy nagyvállalatnál azt vizsgálták, van-e összefüggés a dolgozók neme és a beosztása között. Az eredményeket az alábbi táblázat tartalmazza:

	vezető	beosztott
nő	24	158
férfi	57	321

Elfogadható-e 95%-os szignifikancia szinten az az állítás, hogy a dolgozók neme és beosztása között nincs összefüggés?

Megoldás: Feltételezésünk az, hogy a nem és a beosztás független egymástól, ezért függetlenségvizsgálatot hajtunk végre. Az adatokat a nem szerint $r = 2$, a beosztás szerint $s = 2$ csoportba osztották, a mintaelemek száma összesen $n = 540$.
A fenti, megfigyelt gyakoriságokat tartalmazó táblázatot (2×2-es kontingencia-tábla) egészítsük ki a peremgyakoriságokkal. Ez azt jelenti, hogy meghatározzuk a nemek szerinti gyakoriságokat (a mintában összesen hány nő és hány férfi szerepel), illetve a beosztás szerinti gyakoriságokat (a mintában összesen hány vezető és hány beosztott szerepel). Ehhez összeadjuk az egyes sorokban, illetve oszlopokban található értékeket. Az alábbi táblázatot kapjuk:

	vezető	beosztott	összes
nő	24	158	182
férfi	57	321	378
összes	81	479	560

Most nézzük meg, hogy függetlenség esetén milyen eloszlást várunk!
Az elvárt gyakoriságok kiszámításához minden esetben az alábbiak szerint kell eljárnunk: az adott megfigyelt gyakorisághoz tartozó sor- és oszlopösszeget összeszorozzuk, majd elosztjuk a minta elemszámával. Például a női vezetők száma (megfigyelt gyakorisága) a megfigyelés szerint 24. A női vezetők számának elvárt gyakoriságát úgy számolhatjuk ki, hogy a 24-hez tartozó sorösszeget (182) szorozzuk az oszlopösszeggel (81), majd a szorzatot elosztjuk a minta elemszámával (560).
Így a női vezetők számának elvárt gyakorisága: $\dfrac{182 \cdot 81}{560} = 26,3250$.
Hasonlóan számítjuk ki a többi értéket is:
A férfi vezetők számának elvárt gyakorisága: $\dfrac{378 \cdot 81}{560} = 54,6750$.

A női beosztottak számának elvárt gyakorisága: $\dfrac{182 \cdot 479}{560} = 155,6750.$

A férfi beosztottak számának elvárt gyakorisága: $\dfrac{378 \cdot 479}{560} = 323,3250.$

A fenti értékekkel elkészíthetjük a várt gyakoriságoknak megfelelő táblázatot (ügyelve arra, hogy minden értéket a megfelelő helyre írjunk):

	vezető	beosztott	összes
nő	26,3250	155,6750	182
férfi	54,675	323,3250	378
összes	81	479	560

Az eredményeinket az illeszkedésvizsgálatnál látott táblázatos alakba is írhatjuk, ekkor könnyebben áttekinthető a számolás:

	megfigyelt gyakoriság (O_i)	várt gyakoriság (E_i)	$(O_i - E_i)^2$	$\dfrac{(O_i - E_i)^2}{E_i}$
	24	26,3250	5,405625	0,2053
	158	155,6750	5,405625	0,0347
	57	54,6750	5,405625	0,0989
	321	323,3250	5,405625	0,0167
összeg	560	560		**0,3556**

A hipotézisvizsgálat lépései ezután az alábbiak:

I *A nullhipotézis kimondása.*
A feltételezésünk az, hogy ennél a vállalatnál a nem és a beosztás függetlenek egymástól.

II *A próbafüggvény kiválasztása és értékének kiszámítása.*
A próba függetlenségvizsgálat χ^2-próbával. A próbafüggvény

$$\hat{\chi}_p^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

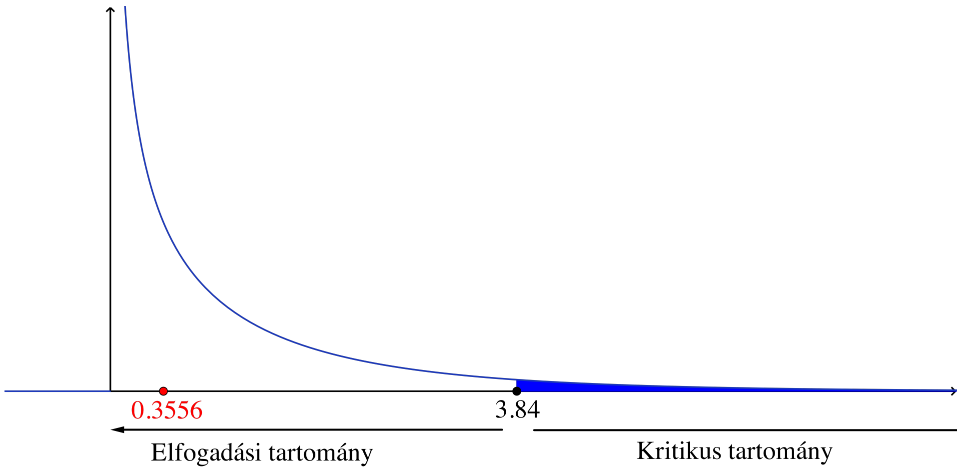
mintából számított értéke a korábban kiszámolt

$$\hat{\chi}_p^2 = 0,3556$$

III *A kritikus tartomány kijelölése.*
A próba egyoldali χ^2-próba, az elfogadási tartomány $(0, \chi_t^2)$. A kritikus értéket a χ^2-eloszlás táblázatából olvashatjuk ki. Ügyeljünk arra, hogy függetlenségvizsgálatnál $(r-1)\cdot(s-1)$ a szabadsági fok, ami jelen példa esetében $(2-1)\cdot(2-1) = 1$, 0,95 valószínűség mellett. Ez a jegyzet χ^2-eloszlás táblázatában az első sor, első oszlopban található

$$\chi_t^2 = 3,84$$

érték.



50. ábra.

IV *A döntés.*
$0,3556 < 3,84$, a próbafüggvény értéke az elfogadási tartományba esik, a nullhipotézist elfogadjuk 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján elfogadható, hogy a vállalatnál független a dolgozók beosztása a nemüktől.

9. feladat (**függetlenségvizsgálat**) Egy kutatás során azt vizsgálták, hogy összefügg-e a néző életkora és a legkedveltebb film típusa. A felmérés eredményeit az alábbi táblázat tartalmazza:

	ismeretterjesztő film	kalandfilm	vígjáték	romantikus film
0-18 éves	12	52	34	25
19-49 éves	28	35	43	24
49 év felett	58	14	43	32

Elfogadható-e 95%-os szignifikancia szinten az az állítás, hogy az életkor és a legkedveltebb film típusa között nincs összefüggés?

Megoldás: Az adatokat az életkor szerint $r = 3$, a film típusa szerint $s = 4$ csoportba osztották, a mintaelemek száma összesen $n = 400$.
A megfigyelt gyakoriságokat tartalmazó táblázatban (3×4-es kontingencia-tábla) az egyes sorokban, illetve az egyes oszlopokban található gyakoriságokat összeadva megkapjuk a peremgyakoriságokat. Foglaljuk ezt össze az alábbi táblázatban:

	ismeretterjesztő film	kalandfilm	vígjáték	romantikus film	összes
0-18 éves	12	52	34	25	123
19-49 éves	28	35	43	24	130
49 év felett	58	14	43	32	147
összes	98	101	120	81	400

A következő lépésben számítsuk ki, hogy függetlenség esetén milyen értékeket várunk az egyes cellákba!
Az előző feladatban látottak szerint az adott megfigyelt gyakorisághoz tartozó sor- és oszlopösszeget összeszorozzuk, majd elosztjuk a minta elemszámával.
Vegyük észre, hogy az algoritmus a következő számítást adja: a kiválasztott cella és az utolsó sor és utolsó oszlopban található, összes mintaszámnak megfelelő cella, tekinthető egy téglalap két szemközti

csúcsának. Pl. tekintsük a 19-49 évesek közül a kalandfilmeket kedvelőket. A cellában a gyakoriság 35, az minta elemszáma 400. Egészítsük ki a táblázatot a téglalap másik két csúcsában található cellák elemszámával (101, 130):

	ismeretterjesztő film	kalandfilm	vígjáték	romantikus film	összes
0-18 éves					
19-49 éves		**35**			*130*
49 év felett					
összes		*101*			**400**

Ennek segítségével a 19-49 évesek között a kalandfilmeket kedvelők cellájába a várt gyakoriság úgy számolható, hogy a téglalapon a kiválasztott cellával szomszédos csúcsokon található cellák értékeinek szorzatát osztjuk a szemközti csúcs cellájának értékével:

$$\frac{130 \cdot 101}{400} = 32,8250$$

Hasonlóan a többi cellára is kiszámítva a várt gyakoriságot, az alábbi táblázatot kapjuk:

	ismeretterjesztő film	kalandfilm	vígjáték	romantikus film	összes
0-18 éves	30,1350	31,0575	36,9000	24,9075	123
19-49 éves	31,8500	32,8250	39,0000	26,3250	130
49 év felett	36,0150	37,1175	44,1000	29,7675	147
összes	98	101	120	81	400

Táblázatos formában számoljuk a próbafüggvény értékét:

	megfigyelt gyakoriság (O_i)	várt gyakoriság (E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
	12	30,1350	328,8782 25	10,9135
	52	31,0575	438,588306	14,1218
	34	36,9000	8,41	0,2279
	25	24,9075	0,008556	0,0003
	28	31,8500	14,8225	0,4654
	35	32,8250	4,730625	0,1441
	43	39,0000	16	0,4103
	24	26,3250	5,405625	0,2053
	58	36,0150	483,340225	13,4205
	14	37,1175	534,418806	14,3980
	43	44,1000	1,21	0,0274
	32	29,7675	4,984056	0,1674
összeg	300	300		**54,5019**

A hipotézisvizsgálat :

I *A nullhipotézis kimondása.*
A feltételezésünk az, hogy az életkor és preferált film típusa függetlenek egymástól.

II *A próbafüggvény kiválasztása és értékének kiszámítása.*
A próba függetlenségvizsgálat χ^2-próbával. A próbafüggvény

$$\hat{\chi}_p^2 = \sum_{i=1}^{12} \frac{(O_i - E_i)^2}{E_i}$$

mintából számított értékét kiolvassuk a táblázatból:

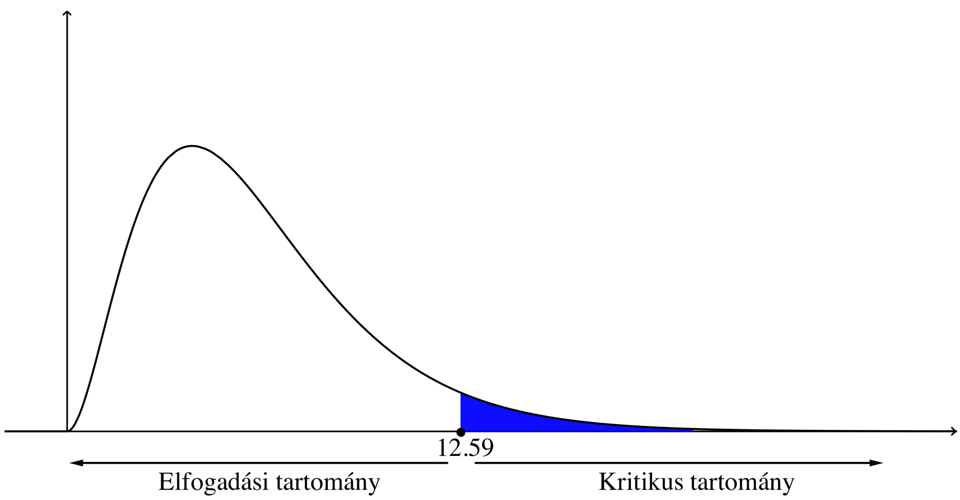
$$\hat{\chi}_p^2 = 54,5019$$

III *A kritikus tartomány kijelölése.*
A próba egyoldali χ^2-próba, az elfogadási tartomány $(0, \chi_t^2)$. A kritikus értéket olvassuk ki a χ^2-eloszlás táblázatából. A szabadsági fok $(r-1)\cdot(s-1) = (3-1)\cdot(4-1) = 6$, $0,95$ valószínűség mellett. Ez a jegyzet χ^2-eloszlás táblázatában a hatodik sor, első oszlopban található

$$\chi_t^2 = 12,59$$

érték.

(A számított érték az x-tengelyen nem is fér fel az ábrára.)



51. ábra.

IV *A döntés.*
$54,5019 > 12.59$, a próbafüggvény értéke az elutasítási tartományba esik, a nullhipotézist elutasítjuk, 95%-os szignifikanciaszinten. Ez azt jelenti, hogy az adott minta alapján van összefüggés az életkor és a legkedveltebb film típusa között.

8. Táblázatok

8.1. A standard normális eloszlásfüggvény ($\Phi(x)$) értékei

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

8.2. A Student-féle eloszlásfüggvény inverzének értékei

f	0,90	0,95	0,975	0,98	0,99	0,995
1	3,08	6,31	12,71	15,89	31,82	63,66
2	1,89	2,92	4,30	4,85	6,96	9,92
3	1,64	2,35	3,18	3,48	4,54	5,84
4	1,53	2,13	2,78	3,00	3,75	4,60
5	1,48	2,02	2,57	2,76	3,36	4,03
6	1,44	1,94	2,45	2,61	3,14	3,71
7	1,41	1,89	2,36	2,52	3,00	3,50
8	1,40	1,86	2,31	2,45	2,90	3,36
9	1,38	1,83	2,26	2,40	2,82	3,25
10	1,37	1,81	2,23	2,36	2,76	3,17
11	1,36	1,80	2,20	2,33	2,72	3,11
12	1,36	1,78	2,18	2,30	2,68	3,05
13	1,35	1,77	2,16	2,28	2,65	3,01
14	1,35	1,76	2,14	2,26	2,62	2,98
15	1,34	1,75	2,13	2,25	2,60	2,95
16	1,34	1,75	2,12	2,24	2,58	2,92
17	1,33	1,74	2,11	2,22	2,57	2,90
18	1,33	1,73	2,10	2,21	2,55	2,88
19	1,33	1,73	2,09	2,20	2,54	2,86
20	1,33	1,72	2,09	2,20	2,53	2,85

f	0,90	0,95	0,975	0,98	0,99	0,995
21	1,32	1,72	2,08	2,19	2,52	2,83
22	1,32	1,72	2,07	2,18	2,51	2,82
23	1,32	1,71	2,07	2,18	2,50	2,81
24	1,32	1,71	2,06	2,17	2,49	2,80
25	1,32	1,71	2,06	2,17	2,49	2,79
26	1,31	1,71	2,06	2,16	2,48	2,78
27	1,31	1,70	2,05	2,16	2,47	2,77
28	1,31	1,70	2,05	2,15	2,47	2,76
29	1,31	1,70	2,05	2,15	2,46	2,76
30	1,31	1,70	2,04	2,15	2,46	2,75
35	1,31	1,69	2,03	2,13	2,44	2,72
40	1,30	1,68	2,02	2,12	2,42	2,70
45	1,30	1,68	2,01	2,12	2,41	2,69
50	1,30	1,68	2,01	2,11	2,40	2,68
60	1,30	1,67	2,00	2,10	2,39	2,66
70	1,29	1,67	1,99	2,09	2,38	2,65
80	1,29	1,66	1,99	2,09	2,37	2,64
90	1,29	1,66	1,99	2,08	2,37	2,63
100	1,29	1,66	1,98	2,08	2,36	2,63
200	1,29	1,65	1,97	2,07	2,35	2,60

8.3. A χ^2-eloszlásfüggvény inverzének néhány értéke

f	0,95	0,98	0,99
1	3,84	5,41	6,63
2	5,99	7,82	9,21
3	7,81	9,84	11,34
4	9,49	11,67	13,28
5	11,07	13,39	15,09
6	12,59	15,03	16,81
7	14,07	16,62	18,48
8	15,51	18,17	20,09
9	16,92	19,68	21,67
10	18,31	21,16	23,21
11	19,68	22,62	24,72
12	21,03	24,05	26,22
13	22,36	25,47	27,69
14	23,68	26,87	29,14
15	25,00	28,26	30,58
16	26,30	29,63	32,00
17	27,59	31,00	33,41
18	28,87	32,35	34,81
19	30,14	33,69	36,19
20	31,41	35,02	37,57

f	0,95	0,98	0,99
21	32,67	36,34	38,93
22	33,92	37,66	40,29
23	35,17	38,97	41,64
24	36,42	40,27	42,98
25	37,65	41,57	44,31
26	38,89	42,86	45,64
27	40,11	44,14	46,96
28	41,34	45,42	48,28
29	42,56	46,69	49,59
30	43,77	47,96	50,89
31	44,99	49,23	52,19
32	46,19	50,49	53,49
33	47,40	51,74	54,78
34	48,60	53,00	56,06
35	49,80	54,24	57,34
36	51,00	55,49	58,62
37	52,19	56,73	59,89
38	53,38	57,97	61,16
39	54,57	59,20	62,43
40	55,76	60,44	63,69

f	0,95	0,98	0,99
50	67,50	72,61	76,15
60	79,08	84,58	88,38
70	90,53	96,39	100,43
80	101,88	108,07	112,33
90	113,15	119,65	124,12
100	124,34	131,14	135,81
110	135,48	142,56	147,41
120	146,57	153,92	158,95
130	157,61	165,22	170,42
140	168,61	176,47	181,84
150	179,58	187,68	193,21
160	190,52	198,85	204,53
170	201,42	209,98	215,81
180	212,30	221,08	227,06
190	223,16	232,15	238,27
200	233,99	243,19	249,45
300	341,40	352,42	359,91
400	447,63	460,21	468,72
500	553,13	567,07	576,49
1000	1074,68	1093,98	1106,97