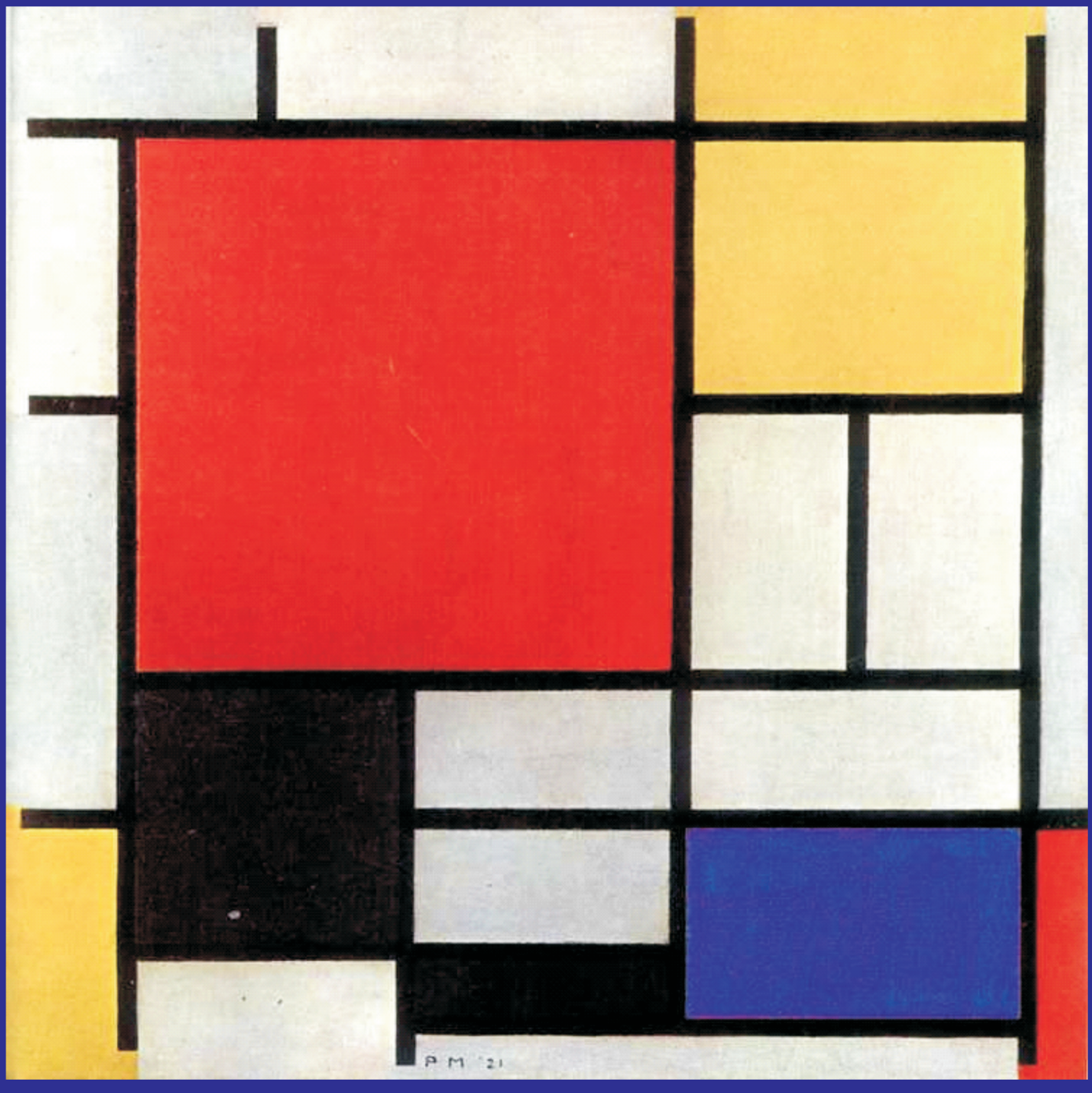


STATISZTIKA és INFORMATIKA

Írta és szerkesztette:
Herényi Levente
a Biofizikai és Sugárbiológiai Intézet
munkatársainak közreműködésével

SEMMELWEIS KIADÓ
2016

Előszó

Ez az alig százoldalas összeállítás abból a célból született, hogy a Semmelweis Egyetemen az orvosképzésben bevezetett „Biostatisztika és informatika” című tantárgy háttéranyagát **röviden** összefoglalja. A „röviden” szót szándékosan emeltem ki, hiszen hasonló témájú könyv akad még egynéhány, akár magyar nyelven is, de ezek lényegesen hosszabbak. A nagyobb terjedelemnek általában az az oka, hogy a tantárgy a matematika tudományának része, illetve arra épül, és a matematika iránt kevésbé fogékony olvasók számára sokkal több magyarázatra, példára van szükség ahhoz, hogy az absztrakt fogalmakat és az alkalmazott módszereket megértsék. A megértés igen fontos, mert csak ebben az esetben kerülhető el a nem megfelelő módszerek szolgai, esetenként teljesen értelmetlen alkalmazása.

Tudva azt, hogy a diákok nem szeretik a vastag tankönyveket, törekedtem a rövidségre. Kompromisszumként az „igazi” matematikai precízséget feladtam, de a világos fogalmazásra azért nagyon törekedtem, szem előtt tartva a fő mondanivalót, nevezetesen azt, hogy **a dolgok hasonlósága, illetve különbözősége sokkal bonyolultabb kérdés, mint amilyennek elsőre látszik**.

Úgy gondolom, hogy ebből az összeállításból sok minden megérthető, megtanulható, de azt is töredelmesen bevallhatom, hogy nem könnyű olvasmány. A használatáról csak annyit, hogy amikor tanulunk, egy-egy magyarázatot nem szabad csupán emlékezetünkbe vésni, ezen a területen ugyanis a logikai lánc követése a lényeg, mindenre kiterjedő, kőbe vésett szabályokat úgy sem tudunk megfogalmazni. Ha egy magyarázat nem eléggé érthető, kérjünk segítséget gyakorlatvezetőnktől, de ne próbáljuk memorizálni a levezetés lépéseit, annak semmi értelme, inkább ugorjuk át. Bár a bevezetett új fogalmak a legtöbb esetben egymásra épülnek, így célszerű az olvasást az elején elkezdeni, de a tanulást, illetve az eligazodást segítheti a kiadvány elején található „Tartalomjegyzék”, továbbá a végén található „Név- és tárgymutató”.

Terveink szerint ezt a kiadványt követni fogja egy gyakorlati jegyzet, amely sok példával, sok gyakorlati alkalmazással segíti elő a leglényegesebb fogalmak és módszerek jobb megértését. Természetesen addig is támaszkodjunk a gyakorlatok anyagára és a gyakorlatvezető kérdéseinkre adott válaszaira.

A kiadvány jellegét a könyv vagy jegyzet helyett az „összeállítás” szóval próbáltam kifejezni, ugyanis a „Felhasznált irodalmi források”-on túlmenően, amelyeket az általam vélelmezett fontosságuknak megfelelően állítottam sorba, a legfőbb forrás a Biofizikai és Sugárbiológiai Intézetben folyó több évtizedes oktató munka eredményeként létrejött tudás, amely korábbi és jelenlegi kollégáimnak köszönhető.

A korábbi kollégáim közül meg kell említenem Hajtman Béla és Berkes László nevét. (Hajtman Béla jegyzetét hosszú éveken keresztül használták az orvostanhallgatók.) A jelenlegi kollégáim közül első helyen Módos Károlyt emelem ki, aki a tantárgy keretében tartott előadásainak anyagát, illetve annak mintegy harmincöt oldalas írott verzióját rendelkezésemre bocsátotta. Felhasználtam továbbá Gróf Pál, Kaposi András, Kellermayer Miklós valamint Schay Gusztáv előadásainak összefoglalóit. Köszönetet kell mondanom azoknak a kollégáknak is akik hasznos tanácsaikkal elősegítették a kiadvány megszületését, így Agócs Gergelynek, Csík Gabriellának, Derka Istvánnak, Osváth Szabolcsnak, Smeller Lászlónak, Tölgyesi Ferencnek, Veres Dánielnek és Voszka Istvánnak. Ezen kollégáim valamennyien olvasták a kézirat valamelyik részét és konkrét javaslataikkal segítették a lektorálást. Természetesen a jelenlegi változat kialakításakor az összes véleményt nem tudtam figyelembe venni, de igyekeztem minden hasznos tanácsot megfogadni.

Végül, de nem utolsósorban köszönetet kell mondanom intézetünk igazgatójának, Kellermayer Miklós professzor úrnak, aki támogatta tevékenységemet és egyszer sem tette fel azt a frusztráló kérdést, hogy mikor leszek kész a munkával.

Budapest, 2016.

Herényi Levente

Tartalomjegyzék

1.0. Bevezetés 2
1.1. Statisztika, statisztikus törvényszerűség 2
2.0. Leíró és induktív statisztika 4
2.1. Az adatok áttekinthetővé tétele, gyakoriság 4
2.2. Halmazok és események 5
2.3. Összegzési és szorzási szabály 6
3.0. Többszöri megfigyelés, „kísérletsorozat”, valószínűség 6
3.1. Eseménytér 7
3.2. Kapcsolatok események között 7
3.3. A relatív gyakoriság és a valószínűség alaptulajdonságai 8
3.4. További skálák az esély mértékének megadására 9
4.0. Összetett megfigyelés, függetlenség 9
4.1. Feltételes valószínűség és függetlensé 11
5.0. Az eloszlás szemléletes jelentése 12
5.1. Az eloszlás matematikai megfogalmazása 12
6.0. Valószínűségi változó 12
6.1. A valószínűségi változó és a számszerű adatok kapcsolata 13
7.0. Folytonos valószínűségi változó jellemzése, eloszlásfüggvénye 16
7.1. Adatsűrűség, gyakorisági eloszlás, relatív gyakorisági eloszlás, hisztogram, gyakoriságsűrűség, relatív gyakoriságsűrűség 17
7.2. Sűrűségfüggvény 19
8.0. Diszkrét valószínűségi változó jellemzése, eloszlásfüggvénye 20
9.0. A valószínűségi változók további tulajdonságai 21
10.0. Adatrendszer és valószínűségi változó számszerű jellemzői 22
10.1. A számszerű jellemzők áttekintése és összehasonlítása 25
11.0. Nevezetes eloszlások, matematikai modellek 25
11.1. A normális eloszlás kitüntetett szerepe 28
12.0. Két valószínűségi változó együttes eloszlása, feltételes eloszlása, függetlensége 30
12.1. A valószínűségi változók transzformációi 31
12.2. Két valószínűségi változó függősége, korreláció, regresszió 32
13.0. A statisztika alapfogalmai; alapsokaság, minta, változó 33
13.1. Mintavételi módszerek 34
13.2. A minta és az alapsokaság hasonlósága, a statisztika alaptétele 35
13.3. A leíró és az induktív statisztika kapcsolata 36
13.4. Adat típusok, a változók osztályozása 37
14.0. Becslés, statisztikai becslés, jó becslés 38
14.1. A becslés pontossága, hibája 39
14.2. A mintaátlag és a mintaszórás néhány fontos tulajdonsága 40
14.3. éseloszlása normális eloszlású sokaság esetén 42
14.4. Egy valószínűség becslése 42
14.5. Konfidencia intervallum (tartomány) 43
14.6. Referencia tartomány (intervallum) 45

15.0. Statisztikai hipotézisvizsgálat (feltevésvizsgálat) 46
15.1. A hipotézisvizsgálat főbb lépései, döntés a konfidencia intervallum ismeretében 48
15.2. Statisztikai próbák 49
15.3. Hibalehetőségek, alternatív hipotézisek 51
15.4. A statisztikai próbák elvégzésének gyakorlati kérdései 52
15.5. A statisztikai próbák fajtái és az ezzel kapcsolatos tudnivalók, illetve félreértések 53
15.6. A sokaság várhatóértékére vonatkozó statisztikai próbák 54
15.7. A sokaság varianciájára vonatkozó statisztikai próbák 56
15.8. Eloszlásokra vonatkozó statisztikai próbák 56
15.9. A sokaság mediánjára vonatkozó statisztikai próba (előjel próba) 60
15.10. Rangpróbák 60
16.0. Korreláció és regresszió számítás 63
16.1. A korreláció mértéke, korrelációs t-próba 63
16.2. Rangkorreláció 64
16.3. A regresszió jelentése a gyakorlatban 64
16.4. Lineáris regresszió 65
16.5. A lineáris regresszióval kapcsolatos hipotézisvizsgálatok 66
16.6. A lineáris regresszió bonyolultabb esetei 67
16.7. Lineárisra visszavezethető nem lineáris regressziók 68
17.0. Varianciaelemzés (varianciaanalízis) 69
18.0. Néhány módszer az orvosi statisztika köréből 71
18.1. Visszatekintő (eset-kontroll) vizsgálat 72
18.2. Előretekintő, követéses (kohort) vizsgálat 72
18.3. Diagnosztikai tesztek jellemzésére szolgáló statisztikai módszerek 73
18.4. A diagnosztikai tesztek összehasonlításának szempontjai, a hatékonyság jellemzése, ROC elemzés 76
19.0. Néhány összegző megjegyzés 76
20.0. Informatikai alapfogalmak 78
20.1. A megfigyelések, „kísérletek” határozatlansága 78
20.2. Információmennyiség 79
20.3. Üzenetek információmennyisége 80
20.4. Redundancia 82
20.5. Üzenetek továbbítása, hírközlési rendszer 82
20.6. A kódolással kapcsolatban felmerülő problémák 83
20.7. Analóg jelek digitális jelekké alakítása 84
21.0. Az irányítás alapelvei 85
21.1. Alapismeretek a számítógépekről 86
21.2. Matematikai logikai alapok 88
22.0. Adatbázisok 89
A legfontosabb irodalmi források 91
Név- és tárgymutató 92
Statisztikai táblázatok 94

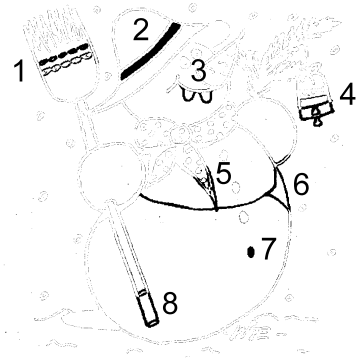
STATISZTIKA ÉS INFORMATIKA

1.0. Bevezetés

1. példa

Hasonítsunk össze két dolgot, például az alábbi két képet. Jól ismert a rejtvény: találd meg a 8 különbséget!

A különbség a kivonás műveletének eredménye. Nosza, vonjuk ki a két képet egymásból, (ami a kép digitalizálása után számítógép segítségével meg is oldható vö. 20.7. rész).



Az eredmény meglepő. Bár a rajzoló és a nyomdász valószínűleg nagyon figyelt arra, hogy 8-nál több különbség ne legyen, a kivonás eredményeként mi mégis többet találtunk. Tudjuk azonban, hogy a 8 feletti meglévő kis különbségek nagy részét még a nagyon alapos szemlélő sem vette volna észre, azok tehát az eredeti körülmények között **számba nem vehetők**.

Az **azonosítás, hasonlítás, különbségtétel** mindennapi életünk állandó része. Születésünktől kezdve ezek alapján tanuljuk meg az újat, találunk megoldásokat a különböző élethelyzetekben, így nem véletlen, hogy igen jelentős a szerepük a gyógyításban, illetve a megelőzésben is. A gyakorló orvos rendszeresen összeveti a beteg vizsgálata során nyert tünetcsoportot a tanult „ideális" betegségtípusok jellemzőivel, illetve saját korábbi tapasztalataival. Döntéseinek meghozatala előtt **bizonyosságot** szeretne. Ehelyett azonban csak a kisebb vagy nagyobb **bizonytalanság** jut osztályrészéül. Ennek az az oka, hogy ***ismereteink sohasem teljes körűek, így mindig lesznek olyan körülmények, amelyeket nem tudunk (vagy nem akarunk) figyelembe venni.***

Így a leendő orvosok számára a **statisztika és informatika** legfőbb célja talán éppen az, hogy **mennyiségileg** is jellemezhető módon adjon iránymutatást arra nézve, hogy a körülöttünk levő világban ***két vagy több* dolog *mennyire hasonlít, mennyire felel meg egymásnak, illetve mennyire különbözik egymástól*** (1. példa).

E két egymással szorosan összefüggő tudományág legfontosabb alapfogalmai az adatok és a jelek. A statisztika elsősorban az adatokkal, az informatika elsősorban a jelekkel foglalkozó terület.

Az **adatok** valakinek, vagy valaminek a megismeréséhez, jellemzéséhez hozzásegítő tények, **a környező világ minőségi és mennyiségi jellemzői**. A hasonlóságokat illetve különbözőségeket a rendelkezésünkre álló adatok alapján állapíthatjuk meg. A **jelek az adatok közvetítői**; az adatok megismerésére, leírására, közzétételére, bemutatására, továbbítására szolgálnak.

1.1. Statisztika, statisztikus törvényszerűség

Maga a szó több jelentéssel bír. A tudományterület megnevezésére vonatkozó értelmét a latin eredetre visszanyúlva a „status" szóból fejthetjük meg. A status eredeti jelentése: állapot. Ennek megismeréséhez szükségesek az adatok. Az állapot, illetve annak jellemzői azonban térben és időben változhatnak. A „status idem" gyakori megállapítás épp a beteg állapotának változatlanságára utaló kifejezés az orvosi gyakorlatban.

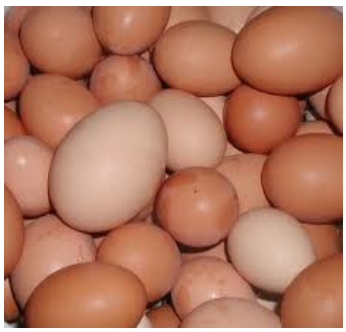
Mindennapi életünkben leggyakrabban előforduló adatok például a személyi adataink: nevünk, születési helyünk, születésünk időpontja; vagy akármilyen boltban az árucikkek neve, az árucikkek ára; de egészségi állapotunkkal kapcsolatban is mondhatunk példát: arcunk sápadtsága, vérnyomásunk, hőmérsékletünk, de akármilyen laboratóriumi diagnosztikai vizsgálat eredménye is adat.

Két dolog egyes adatainak **teljes egyezőségére** rengeteg szavunk van: *ugyanekkora, ugyanakkora, ugyanennyi, ugyanannyi, ugyanilyen, ugyanolyan* stb., de teljes egyezőség valójában csak speciális esetekben fordulhat elő.

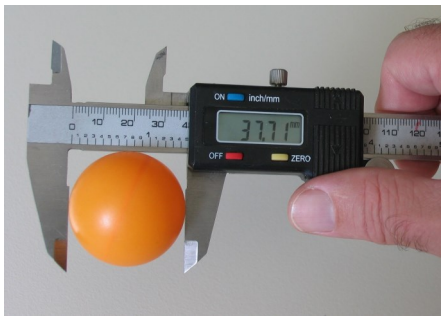
Léteznek például azonos nevű emberek, ennek igazolására csak ki kell nyitnunk egy telefonkönyvet, amelyben biztosan találunk erre példát. Ugyanakkor gyakran mondják, hogy „olyan, mint két tojás", ami köznapi értelemben azt jelenti, hogy a két dolog „tökéletesen egyforma", pedig tudjuk, hogy valójában nem létezik két „tökéletesen egyforma" tojás (2. példa).

Egyszerűsítve az előző példát, tojás helyett vegyünk két pingponglabdát, és fontoljuk meg, hogy mennyire egyformák, mennyire hasonlítanak egymáshoz, illetve az absztrakt „elméleti" labdához. Az előírás szerint a versenyeken engedélyezett „elméleti" labda gömb alakú, 38 mm átmérőjű, matt fehér színű (a márkajelzésektől most tekintsünk el) és 2,5 g tömegű. Minden „gyakorlati" labda igyekszik követni ezeket az adatokat, de hogy valójában milyenre sikerült, csak **megfigyelésekkel**, **mérésekkel** állapíthatjuk meg (3. példa).

Itt rögtön egy nagyon fontos észrevételt kell tennünk. Meg kell ugyanis különböztetnünk a „valóságot" a modelljétől, illetve a matematikai modelljétől. Minden-



2. példa
Mennyire egyformák a tojások?



3. példa
Egy „valóságos”, „gyakorlati” pingponglabda.

ki tisztában van a Föld, egy földgömb és a geometriai értelemben vett gömb viszonyával. A Föld sok szempontból hasonlít azokra a földgömbökre, melyek a térképboltokban kaphatók, ezért ezek a földgömbök mind modelljei a Földnek. A matematikai gömb csak absztrakció, a valóságban nem létezik, mégis jól használható a Föld vagy akár a pingponglabda modellezésére. Egy matematikai objektum akkor tekinthető egy valóságos dolog matematikai modelljének, ha azokból a szempontokból, amelyek bennünket érdekelnek, hasonlítanak egymásra. (Az, hogy egy modell jó-e, vagy hogy több modell közül melyiket használjuk, tapasztalatokon alapuló szubjektív döntésen múlik. Pusztán matematikai okoskodással nem lehet „bebizonyítani”, hogy az egyik modell jobb, mint a másik.)

A „valóság” leírásához tehát sok esetben **modelleket használunk**. A természettudományok törvényei is mind-mind modellekre vonatkoznak, és nem a „valóságra”. Gondoljunk csak olyan fogalmakra, mint a „merev test”, a „tömegpont” vagy az „ideális gáz”. Mivel a statisztikus törvényszerűségek matematikai modelljeit a valószínűségszámítás szolgáltatja, ezért erről is szót kell ejtenünk.

Kezdjük három, a valószínűségszámítás körében használt alapfogalommal:

Jelenség: minden, ami ***lényegében azonos feltételek mellett megismétlődhet***, amivel kapcsolatban megfigyeléseket lehet végezni, lehet vele „kísérletezni”.

Megfigyelés, „kísérlet”: az a tevékenység, amikor először megadjuk, hogy a ***jelenséggel kapcsolatban mire vagyunk kíváncsiak***, illetve hogy azt hogyan érzékeljük vagy hogyan mérjük, majd erre vonatkozóan ***adatokat szerzünk.*** Egy-egy ilyen adat a megfigyelés eredménye vagy a kísérlet kimenetele.

Esemény: a ***kapott adatokra, kimenetelekre vonatkozó állítás***, amelyről eldönthető, hogy bekövetkezett-e vagy sem (4. táblázat).

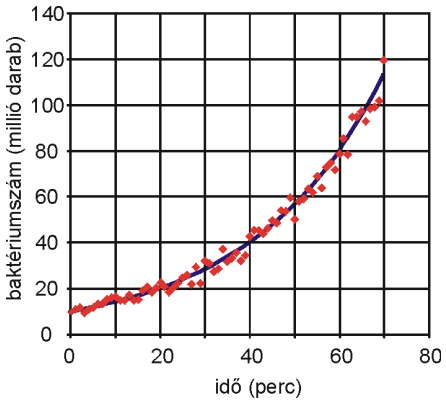
	példák			
Jelenség	orvosi vizsgálat	pénzfeldobás (1)	várakozás a buszra	pénzfeldobás (2)
Megfigyelés	a beteg bőrének színe	az érme repülési ideje	a várakozók száma	az érme melyik oldalára esik
Esemény	sárga vagy piros	0,5 és 1,5 s között van	tíz	fej

4. táblázat
Néhány példa a fogalmak tisztázása érdekében.

Gyakran előfordul, hogy egy jelenséggel kapcsolatosan már néhány megfigyelés után valamilyen törvényszerűséget fedezünk fel. A lámpa elalszik, ha lekapcsolják, a cukor feloldódik a forró teában. Az ilyen törvényszerűségeket determinisztikus törvényszerűségeknek szokás nevezni, hiszen a jelenség körülményei meghatározzák, determinálják a megfigyelés eredményét. Na, de milyen az a törvényszerűség, amelyik nem determinisztikus. Azt gondolhatnánk, hogy a törvényszerűség éppen azért törvényszerűség, mert determinisztikus.

Ha munkába menet a buszra várakozók számát figyeljük, és mondjuk három egymás utáni napon 8-an, 18-an, illetve 13-an várakoztak, ebből különösebb törvényszerűséget még nem vonhatunk le. De, ha például egy hónapon keresztül egyszer sem fordul elő, hogy a zsúfoltság miatt lemaradunk, akkor azt mondhatjuk, hogy ez a járat nem zsúfolt, bátran ajánlhatjuk a munkába igyekvőknek. Ebből persze nem következik az, hogy a továbbiakban már sohasem maradunk le róla. Ilyen az úgynevezett statisztikus törvényszerűség. Szembetűnő és lényeges jellemzője, hogy csupán néhány megfigyelés alapján nem jut érvényre.

Ha ezek után jobban szemügyre vesszük a determinisztikusnak mondott törvényszerűségeket, akkor ott is megfigyelhetünk ehhez hasonlókat. Vegyük például azt az esetet, amikor egy céltáblára lövünk. A megfelelő fizikai törvények alapján egyértelműen meg tudjuk mondani, hogy hova érkezik a lövedék, ha ismerjük a lövés körülményeit. A helyzet azonban az, hogy ***nem tudunk minden körülményt számba venni***. Így a kimenetel sem egyértelmű, és emiatt az eredetileg determinisztikusnak vélt törvényszerűség „statisztikussá válik”.



5 ábra
A baktérium kolónia szaporodása elméletben, a megfelelő matematikai modell szerint (kék görbe) és gyakorlatban, a mérések alapján (piros szimbólumok).

Annak érdekében, hogy a problémát még jobban érzékeljük, kövessük nyomon egy baktérium kolónia szaporodását. Először tekintsük azt a matematikai modellt, amelyet a jelenség determinisztikus leírására szokás használni. Amennyiben N_0 a baktériumok száma a vizsgálat kezdetén (t = 0 időpontban), akkor egy későbbi t időpontban $N(t)$, amit az

$$N(t) = N_0 2^{\frac{t}{T}} \qquad (1)$$

függvénnyel adhatunk meg, ahol T a duplázódási idő, ennyi idő alatt kétszereződik meg a baktérium populáció egyedszáma. Ha ugyanis t éppen T-vel egyenlő, akkor a kitevő 1, tehát $N(T) = 2N_0$.

Vegyünk egy konkrét esetet, legyen N_0 = 10 millió és T = 20 perc. Ekkor a 5. ábrán a kék görbe mutatja az $N(t)$ függvényt, a piros szimbólumok pedig egy képzeletbeli mérési sorozat eredményeit. Láthatjuk, hogy a méréskor **a „változások" statisztikus és determinisztikus része együtt van jelen**: kisebb ingadozások mellett „kvázi monoton" a növekedés, de ezek szétválasztása nem egyértelmű. A kérdést itt is úgy lehet feltenni: mennyire hasonlít a két reprezentáció egymásra, vagy mennyire felel meg a matematikai modell a mérési eredményeknek.

A fordított példát követve a feldobott pénzérmét sem a „vak véletlen" vezérli, amikor egyik vagy másik oldalára esik. Egyszerűen arról van szó, hogy nem ismerjük kellő pontossággal azokat az adatokat, amelyek egyértelműen meghatároznák a pénzérme végső állapotát, nevezetesen azt, hogy a dobás „eredménye" fej vagy írás. Mivel nem tudunk minden körülményt figyelembe venni, ezért nem tudunk egyértelmű választ sem adni, tehát csak azt mondhatjuk, hogy **a jelenség véletlenszerű**, ahol **a „véletlen"** szó csak **ismereteink hiányát fejezi ki**.

1.	adatgyűjtés	leíró statisztika
2.	az adatok áttekinthetővé tétele	
3.	elemzés	induktív statisztika
4.	következtetés	

6. táblázat
A statisztikai tevékenység legfontosabb lépései és két fő csoportja.

7. megjegyzés
Az **induktív** gondolkodás lényege az, hogy egyes adatokból következtetünk általános törvényszerűségekre. Az induktív gondolkodás révén alapvetően új tudás keletkezhet, de az így nyert tudás igazsága nem mindig bizonyítható, sok esetben csak valószínűsíthető. (Lásd például a baktérium kolónia szaporodását és annak matematikai modelljét.)
A matematikai feladatok megoldásánál is gyakran alkalmazzuk a logikai következtetésnek ezt a formáját, például, amikor számsorozatoknál ki kell következtetni azt az általános szabályt, ami szerint a sorozat felépül. Például a Fibonacci–számok (F_n = 0, 1, 1, 2, 3, 5, 8, 13,…) esetén az általános szabály:

$$F_n = \begin{cases} 0, & \text{ha } n = 0 \\ 1, & \text{ha } n = 1 \\ F_{n-1} + F_{n-2}, & \text{ha } n > 1 \end{cases}$$

2.0. Leíró és induktív statisztika

Első közelítésben a statisztikai tevékenységet négy csoportba sorolhatjuk, bár ezek között nincs éles határ: **adatgyűjtés**, **az adatok áttekinthetővé tétele**, az adatok **elemzése**, valamint a **következtetések levonása** az adatok alapján (6. táblázat). Az első kettő a leíró statisztika körébe tartozik, ahol az adatrendszerekkel kapcsolatban olyan fogalmakat ismerünk meg, melyek ***statisztikus törvényszerűségek feltételezése nélkül is értelmezhetők***. Ennek ellenére célszerű ezeket a fogalmakat a nekik megfelelő valószínűségszámítási fogalmakkal együtt bevezetni. A másik kettő az induktív statisztika körébe tartozik, ahol ***a valószínűségszámítási alapok nélkülözhetetlen*** kellékek (7. megjegyzés).

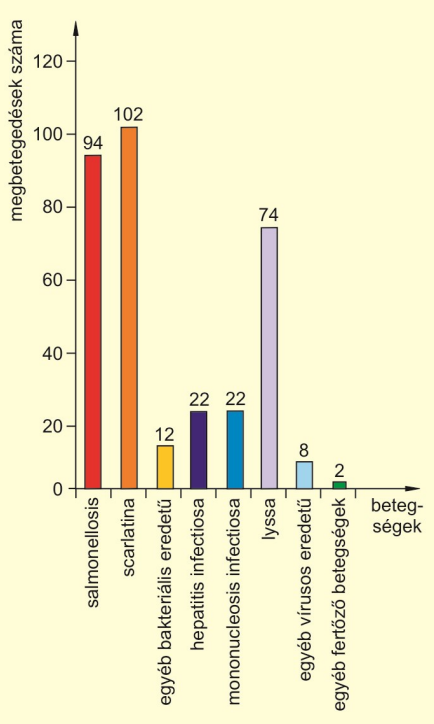
Normális esetben adatokat csak valamilyen cél érdekében gyűjtünk. Például, azért kérdezzük meg valakinek a telefonszámát, hogy később fel tudjuk hívni. Nem szerencsés az a hozzáállás, hogy gyűjtsünk adatokat, majd csak jó lesz valamire, és utólag próbálunk célokat kitalálni. Nagyobb számban adatokat akkor gyűjtünk, ha azt reméljük, hogy ezek segítségével valamilyen korábban feltett ***kérdésünkre*** feleletet kapunk. Az adatok egy része ismert, csak meg kell kérdezni valakitől, másik részét meg kell mérni valahogy, de lehet, hogy csak meg kell figyelni valamit, és ezáltal juthatunk hozzájuk. Az ***adatgyűjtés*** kérdésére, fontossága miatt, a későbbiek során még visszatérünk (13.0. rész).

2.1. Az adatok áttekinthetővé tétele, gyakoriság

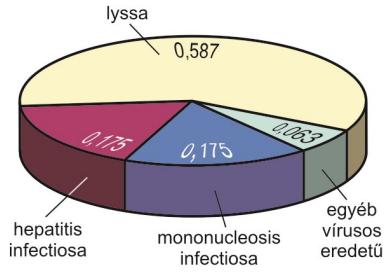
Az összegyűjtött, de rendezetlen adatok önmagukban sok esetben teljesen használhatatlanok. A mindennapi életben is gyakran előfordul, hogy egy probléma kapcsán viszonylag sok adat áll rendelkezésünkre. Ilyen esetekben **szükséges, hogy az adatokról valamilyen áttekintésünk legyen**.

A 8. táblázat a Budapesten, 2000 októberében bejelentett fertőző megbetegedések összesítését mutatja. A táblázat első számoszlopában látható számok (94, 102 stb.) azt mutatják, hogy az egyes betegségtípusokból (salmonellosis, scarlatina stb.) hányat észleltek az adott időszakban. Ezeket a számokat abszolút gyakoriságoknak nevezzük. A következő oszlopban a részösszegek (208, 126, 2) szerepelnek, tehát az, hogy az észlelt betegségek közül hány volt bakteriális, vírusos, vagy egyéb eredetű. Ezek szintén abszolút gyakoriságok.

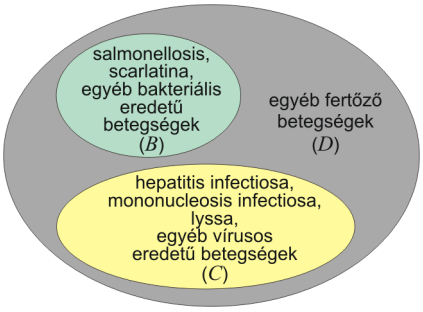
Ha az abszolút gyakoriságokat elosztjuk az adott területen, adott időszakban



9. ábra
Oszlop diagram. Abszolút gyakoriságok a betegségek függvényében.



10. ábra
Torta diagram. Relatív gyakoriságok a vírusos eredetű fertőző betegségek megoszlásáról.



11. ábra
A 7. táblázatban szereplő fertőző betegségek mint részhalmazok.(Itt az általános elem, vagyis a változó maga a betegség.)

KÓROKOZÓ	BETEGSÉG	abszolút gyakoriság		relatív gyakoriság		feltételes relatív gyakoriság	
baktérium	salmonellosis (szalmonella fertőzés)	94	208	0,280	0,619	0,452	1,000
	scarlatina (skarlát)	102		0,303		0,490	
	egyéb bakteriális eredetű	12		0,036		0,058	
vírus	hepatitis infectiosa (fertőző májgyulladás)	22	126	0,065	0,375	0,175	1,000
	mononucleosis infectiosa (mirigyláz)	22		0,065		0,175	
	lyssa (veszettség)	74		0,220		0,587	
	egyéb vírusos eredetű	8		0,025		0,063	
egyéb	egyéb fertőző betegségek	2	2	0,006	0,006	1,000	1,000
összesen:		336	336	1,000	1,000		

8. táblázat
Összesítés a fertőző megbetegedésekről.

előforduló összes fertőző betegség számával (336), akkor megkapjuk a viszonylagos értékeket, a relatív gyakoriságokat. Ezek mindig 0 és 1 közé eső számok, és a táblázat következő két oszlopa tartalmazza őket, de %-ban is kifejezhetők. A relatív gyakoriság, értelmezéséből kifolyólag, egy hányados. Ezért **nem csak azt kell tisztázni, hogy minek a relatív gyakoriságáról beszélünk, hanem azt is, hogy mihez viszonyítunk**.

Ha például arra vagyunk kíváncsiak, hogy a bakteriális eredetű betegségeken belül milyen gyakori a szalmonella fertőzés, akkor a szalmonella fertőzések számát (94) az összes bakteriális eredetű betegség számával (208) kell elosztani. Az így kapott hányados (0,452) is relatív gyakoriság, de most nem az összes fertőző betegséghez (336), hanem csak a bakteriális eredetű betegségekhez (208) viszonyítottunk. Ez a feltételes relatív gyakoriság, ahol a „feltétel” azt jelenti, hogy az összes fertőző betegség helyett csak a bakteriális eredetű betegségek között vizsgálódunk. A „feltételes” szó mindig arra utal, hogy szűkebb összességhez viszonyítunk, esetünkben: **„feltéve, hogy”** a betegségek bakteriális eredetűek.

Az abszolút és relatív gyakoriságok ábrázolására sok lehetőség kínálkozik. Ezek közül mutatunk be kettőt a 9. és 10. ábrán.

2.2. Halmazok és események

A 8. táblázatban rendszereztük, csoportosítottuk a megbetegedéseket, de úgy is mondhatjuk, hogy különböző halmazokat adtunk meg. A halmaz tetszőleges természetű dolgoknak valamilyen módon **egyértelműen** jellemzett összessége. A halmazhoz tartozó dolgok a halmaz elemei. Valamely halmaz általános elemét gyakran változónak nevezik. Legegyszerűbben úgy képzelhetjük el a halmazokat, hogy azok „absztrakt zsákok”, amelyeket akkor ismerünk, ha tudjuk, hogy mi van bennük. A legfontosabb éppen az, hogy **el tudjuk dönteni valamiről, hogy az eleme vagy nem eleme az adott halmaznak**, azaz benne van a „zsákban” vagy nincs benne.

A 11. ábrán ellipszis alakú területek az „absztrakt zsákok”. A nagy ellipszis szemlélteti az összes fertőző betegséget (jelöljük ezt a halmazt A-val), ezen belül a világoskék terület mutatja a bakteriális eredetű betegségek részhalmazát (B), a világossárga terület mutatja a vírusos eredetű betegségek részhalmazát (C), a maradék szürke terület pedig az egyéb fertőző betegségek részhalmazát (D). Ha egyesítjük a B, C, D halmazokat, megkapjuk A-t. A matematikában szokásos jelöléssel:

$$B \cup C \cup D = A\,, \qquad (2)$$

ahol az $\cup$ jel uniót, egyesítést jelent.

Azt, hogy a B és C halmazoknak nincs közös részük (egy vírusos betegség nem lehet egyben bakteriális eredetű is), másként mondva a halmazok diszjunktak, azaz

	példa
Jelenség	orvosi vizsgálat
Megfigyelés	a fertőző betegség eredete
Esemény (C)	vírusos eredetű

12. táblázat
A vírusos eredetű megbetegedés, mint esemény.

közös részük a üres halmaz, a következőképpen jelöljük:

$$B \cap C = \emptyset \ , \qquad (3)$$

ahol a $\cap$ szimbólum a közös rész vagy metszet, a Ø pedig az üres halmaz jele. Üres halmaznak nevezzük azt az „absztrakt zsákot”, amelyik üres, tehát nincs eleme.

A 10. ábrán feltüntetett B, C, D részhalmazok a valószínűségszámítás fogalomkörével is leírhatók. Így, ha az orvos vírusos eredetű megbetegedést észlel az orvosi vizsgálat során, akkor azt is mondhatjuk, hogy a C esemény bekövetkezett (12. táblázat). Ennek megfelelően, az eseményekhez is hozzárendelhetők gyakoriságok, illetve relatív gyakoriságok. A 8. táblázat adatai szerint a C esemény bekövetkezésének gyakorisága 126, relatív gyakorisága 0,375. **Ily módon a halmazok és az események megfeleltethetők egymásnak.**

2.3. Összegzési és szorzási szabály

Mintafeladat

Egyetemünkön a tavalyi vizsgákon az elégtelen osztályzatok relatív gyakorisága 0,15, a sikeres vizsgák között a jelesek relatív gyakorisága 0,2 volt. Az összes vizsgajegy között mennyi volt a jeles relatív gyakorisága?

Megoldás: A feladat szövege szerint a 0,15 relatív gyakoriság, a 0,2 feltételes relatív gyakoriság, hiszen itt csak egy szűkebb összességen belül, nevezetesen a sikeresen vizsgázók között vizsgálódunk. Ez a szűkebb összesség a vizsgázók 85%-a, ugyanis 15% megbukott. A kérdés tehát az, hogy a 85%-nak a 20%-a az eredetinek hányad rész. 0,85-nek az egyötöde (20%-a) 0,17. Tehát az összes vizsgajegy között 0,17 volt a jeles relatív gyakorisága (13. példa).

A mintafeladat megoldása két fontos szabályon alapszik:

$$1.\quad \underset{\text{elégtelenek száma}}{\overset{0,15}{\frac{\text{elégtelenek száma}}{\text{összes hallgató száma}}}} + \underset{\text{összes hallgató száma}}{\overset{0,85}{\frac{\text{sikeres vizsgák száma}}{\text{összes hallgató száma}}}} = 1$$

$$\downarrow$$

$$2.\quad \underset{\cancel{\text{sikeres vizsgák száma}}}{\overset{0,2}{\frac{\text{jelesek száma}}{\cancel{\text{sikeres vizsgák száma}}}}} \cdot \underset{\text{összes hallgató száma}}{\overset{0,85}{\frac{\cancel{\text{sikeres vizsgák száma}}}{\text{összes hallgató száma}}}} = \underset{\text{összes hallgató száma}}{\overset{0,17}{\frac{\text{jelesek száma}}{\text{összes hallgató száma}}}}$$

13. példa
Példa az összegzési és szorzási szabályra a mintafeladat alapján.

1. Összegzési szabály:

a. Az abszolút gyakoriságok mindig összeadhatók.

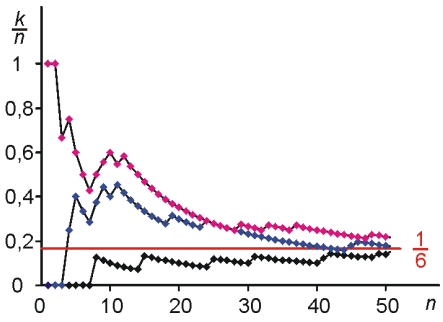
b. A relatív gyakoriságok is összeadhatók abban az esetben, ha ugyanahhoz az összességhez viszonyítjuk őket. (Közös nevező!)

2. Szorzási szabály:

Egy feltételes relatív gyakoriság és egy (feltétel nélküli) relatív gyakoriság összeszorozható, amennyiben a fentiek szerint tudunk egyszerűsíteni (lásd 13. példa). Ez a fajta szorzás akár több tényezőre is alkalmazható.

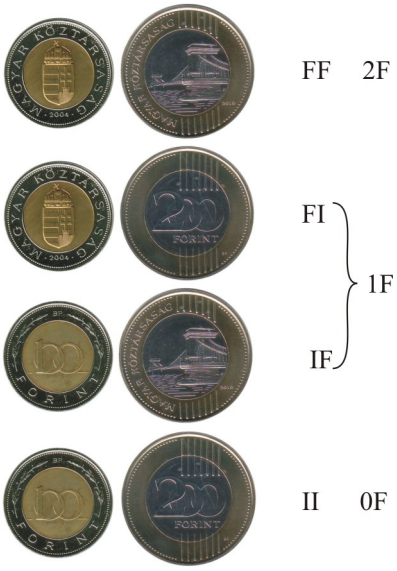
3.0. Többszöri megfigyelés, „kísérletsorozat”, valószínűség

Amikor egy jelenséget egymás után többször, mondjuk n-szer megfigyelünk – a megfigyelést meghatározó körülmények lényegi megváltoztatása nélkül – , akkor azt is mondhatjuk, hogy n hosszúságú „kísérletsorozatot” hajtunk végre. Például, ha 50-szer dobunk egy dobókockával, és az minden alkalommal megáll az asztalon, mutatva a dobás eredményét (mondjuk, ezzel a feltétellel határozzuk meg a megfigyelést), akkor ez egy 50 hosszúságú „kísérletsorozat”. Megmaradva ennél a példánál vizsgáljuk meg a 6-os dobás eseményét. Végezzük el a „kísérletsorozatot” úgy, hogy minden egyes dobás után számítsuk ki az esemény relatív gyakoriságát az addig végrehajtott kísérletekből. Ha n a „kísérletsorozat” aktuális hosszúsága és k a kedvező kimenetelek, azaz a 6-os dobások száma (másként mondva az esemény bekövetkezésének abszolút gyakorisága), akkor a



14. ábra
A 6-os dobás eseményének relatív gyakoriságai 3 „kísérletsorozatban”.

15. megjegyzés
A relatív gyakoriságokra vonatkozó nagy számok törvénye statisztikus törvény. Hiába ismerjük egy esemény valószínűségét, nem lehet előre megjósolni, hogy egy „kísérlet” során az esemény bekövetkezik-e vagy sem. A valószínűség ismerete csupán arról ad felvilágosítást, hogy nagy számú „kísérlet” esetén milyen arányban következik be az esemény.



16. példa
Példák eseménytérre egy 100 Ft-os és egy 200 Ft-os érme egyidejű feldobásával kapcsolatban.

szóban forgó „kísérletsorozatban” a $\frac{k}{n}$ hányados az esemény relatív gyakorisága. A 14. ábrán 3 ilyen 50-es sorozat eredménye látható.

Tapasztalati tény, hogy a relatív gyakoriságok ilyen sorozatai – bár ingadozásokat mindig mutatnak – a „kísérletsorozat” hosszának növekedtével egyre inkább stabilizálódnak valamilyen érték körül, továbbá ez az érték az aktuális „kísérletsorozattól” függetlenül lényegében ugyanakkora. A relatív gyakoriságoknak ezt a szabályszerűségét **a nagy számok (*relatív gyakoriságokra vonatkozó*) tapasztalati törvényének** szokták hívni, melyet ***logikai úton bizonyítani nem lehet.***

Ennek alapján az eseményhez egy számértéket rendelhetünk, amely az esemény jellemzője: ha a vizsgált jelenségre vonatkozó hosszú „kísérletsorozatot” végzünk, akkor a kimeneteleknek körülbelül ennyied részében következik be az esemény. Ezt a számértéket az esemény valószínűségének nevezzük (15. megjegyzés).

3.1. Eseménytér

Ha egy jelenség megfigyelésének eredményét a körülmények nem határozzák meg egyértelműen, mint ahogy az legtöbbször lenni szokott, akkor kézenfekvő legalább elvileg megragadni **a megfigyelés összes lehetséges kimenetelét**. Ha ezeket mind beletesszük egy „absztrakt zsákba”, ezzel megadunk egy halmazt az eseményteret. Ha például a jelenség egy kör alakú céltáblára történő lövés és a találatok helyét figyeljük meg, akkor az eseménytér egy körlap, és a kimenetelek a körlap matematikai értelemben vett pontjai, tehát az (x,y) számpárok.

Ha egy 100 Ft-os és egy 200 Ft-os érmét egyszerre feldobva megfigyelhetjük, hogy az érméknek melyik oldala lesz felül, akkor négyféle kimenetel lehetséges: FF, FI, IF, II (16. példa). Ebben az esetben az eseménytér az alábbi négyelemű halmaz:

$$S_1 = \{\mathrm{FF}, \mathrm{FI}, \mathrm{IF}, \mathrm{II}\}\ . \qquad (4)$$

Ha ugyanezzel a jelenséggel kapcsolatban csak a dobott fejek számát figyeljük meg, akkor az eseménytér az alábbi háromelemű halmaz:

$$S_2 = \{2\mathrm{F}, 1\mathrm{F}, 0\mathrm{F}\}\ . \qquad (5)$$

3.2. Kapcsolatok események között

Elöljáróban fontos hangsúlyoznunk, hogy eseménynek tekintjük a biztos eseményt, ami mindig bekövetkezik, és a lehetetlen eseményt is, ami sohasem. A kockadobással kapcsolatban egy biztos esemény: „7-nél kisebbet dobunk”; egy lehetetlen pedig: „9-nél nagyobbat dobunk”. Megjegyezzük még, hogy a „hatost dobunk” esemény ellentettje a „nem hatost dobunk” esemény.

Korábban megmutattuk, hogy az események és a halmazok megfeleltethetők egymásnak, így az eseményeket legegyszerűbben halmazokkal szemléltethetjük.

Legyen egy dobozban 10 piros, 10 fehér és 10 zöld golyó. A jelenség álljon abból, hogy kiveszünk két golyót a dobozból. Színesnek nevezzük a piros és a zöld golyókat. (A fehér ebben a „kísérletben” a nem színes.) Tekintsük az alábbi két eseményt:

A ***esemény***: „az egyik piros, és a másik is színes”;
B ***esemény***: „az egyik zöld, és a másik is színes”.

Az *A* eseményt úgy is megfogalmazhatjuk, hogy „vagy két piros vagy egy piros és egy zöld golyót veszünk ki”, azaz a húzás eredménye: PP vagy PZ (sorrendre való tekintet nélkül, azaz PZ=ZP). Tehát az *A* halmazban ezt a két elemet találjuk:

$$A = \{\mathrm{PP}, \mathrm{PZ}\}\ . \qquad (6)$$

Hasonlóképpen a *B* esemény akkor következik be, ha a húzás eredménye: ZZ vagy ZP, (de ZP=PZ itt is igaz), így:

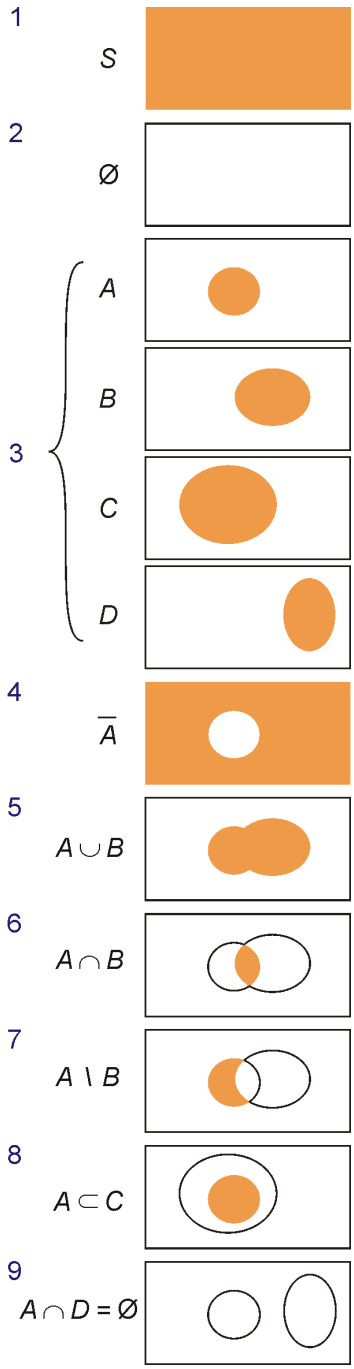
$$B = \{\mathrm{ZZ}, \mathrm{ZP}\}\ . \qquad (7)$$

A és *B* esemény egyesítése, uniója:

$$\mathrm{A} \cup \mathrm{B} = \{\mathrm{PP}, \mathrm{PZ{=}ZP}, \mathrm{ZZ}\}\ , \qquad (8)$$

azaz „mindkettő színes”; *A* és *B* **esemény közül legalább az egyik** bekövetkezik.

17. megjegyzés
Az egyesítés, illetve unió kifejezések helyett gyakran az **összeget** ($A + B$) ≡ (A vagy B); a közös rész, illetve metszet helyett pedig a **szorzatot** (AB) ≡ ($A * B$) ≡ (A és B) használjuk.



19. ábra
A halmazok diagramokkal történő szemléltetése lehetőséget ad az események és a velük kapcsolatos műveletek bemutatására.

A és B esemény közös része, metszete:

$$A \cap B = \{\text{PZ=ZP}\}\ , \qquad (9)$$

azaz „egyik piros, másik zöld”; A **esemény és** B **esemény is** bekövetkezik (17. megjegyzés).

A és B esemény különbsége:

$$A \setminus B = \{\text{PP}\}\ , \qquad (10)$$

azaz „mindkettő piros”. (Továbbá $B \setminus A = \{\text{ZZ}\}$, azaz „mindkettő zöld”.)

Azt mondjuk, hogy egy esemény maga után von egy másik eseményt, ha az eredeti eseménynek a bekövetkezése csak úgy képzelhető el, hogy a másik esemény is bekövetkezik. Például az „egyik piros, másik zöld” esemény maga után vonja a „mindkettő színes” eseményt.

Két eseményt egymást kizáró eseményeknek nevezünk, ha egyidejű bekövetkezésük lehetetlen. Ha több egymást kölcsönösen kizáró esemény közül egy mindig bekövetkezik, akkor azt mondjuk, hogy az események teljes eseményrendszert alkotnak, ami egyben azt is jelenti, hogy egyesítésük a biztos esemény. Az előbbi példa esetében az alábbi négy esemény teljes eseményrendszert alkot:

„mindkettő színes”	{PP, PZ=ZP, ZZ}
„egyik piros, másik fehér”	{PF=FP}
„egyik zöld, másik fehér”	{ZF=FZ}
„mindkettő fehér”	{FF}

A 18. táblázatban az eseményekkel és halmazokkal kapcsolatos fogalmak összevetése látható, amit a 19. ábrán diagramokkal is szemléltetünk.

1	biztos esemény	eseménytér halmaza	S
2	lehetetlen esemény	üres halmaz	$Ø$
3	esemény	eseménytér részhalmaza	A, B, C, D
4	esemény ellentettje	halmaz komplementere	$\overline{A}$
5	események egyesítése	halmazok uniója	$A \cup B$
6	események közös része	halmazok metszete	$A \cap B$
7	események különbsége	halmazok különbsége	$A \setminus B$
8	egyik esemény maga után vonja a másikat (ha A bekövetkezik, akkor C is)	egyik halmaz (A) része a másiknak	$A \subset C$
9	egymást kizáró események	diszjunkt halmazok	$A \cap D = Ø$

18. táblázat
Az eseményekkel és halmazokkal kapcsolatos fogalmak összevetése.

3.3. A relatív gyakoriság és a valószínűség alaptulajdonságai

Mivel a relatív gyakoriságot és a valószínűséget a nagy számok törvénye összekapcsolja, ezért az alábbi tulajdonságok mindkettőre érvényesek.

1. Egy esemény relatív gyakorisága nem lehet 0-nál kisebb vagy 1-nél nagyobb.
2. A biztos esemény relatív gyakorisága mindig 1, a lehetetlen eseményé pedig 0.
3. Egymást kizáró események egyesítésének relatív gyakorisága az egyes események külön-külön vett relatív gyakoriságainak az összege.

A 18. táblázat jelöléseit használva, továbbá, ha az A esemény valószínűségét (az angol probability első betűjét használva) $P(A)$-val jelöljük, akkor:

1. $0 \le P(A) \le 1$
2. a) $P(S) = 1$, b) $P(Ø) = 0$ (Kolmogorov axiómák)
3. $P(A \cup B) = P(A) + P(B)$, ha $A \cap B = Ø$.

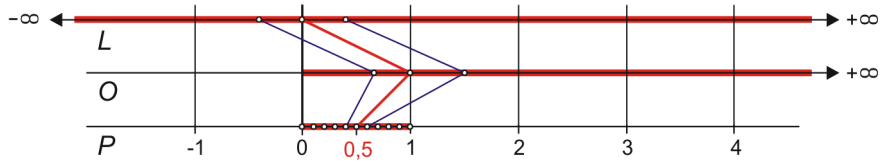
3.4. További skálák az esély mértékének megadására

Egy esemény bekövetkezésének esélyét a ***valószínűség*** a 0 és 1 közötti valós számok skáláján fejezi ki. Bár ez a leggyakrabban használt jellemző, vannak ugyanerre a célra alkalmas **egyéb mérőszámok** is.

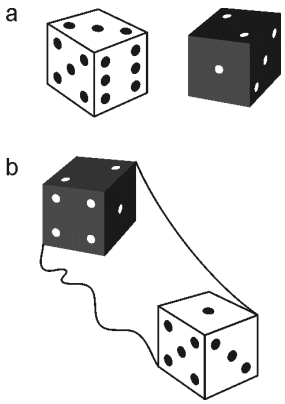
Az esélyértéket fogadásoknál szokták használni és azt fejezi ki, hogy a tét hányszorosa legyen a nyeremény. Egy esemény esélyértékét úgy fogalmazhatjuk meg, hogy **„hányszor akkora a valószínűsége annak, hogy az esemény bekövetkezik, mint annak, hogy nem következik be”**. Például, ha egy esemény bekövetkezésének valószínűsége 0,75, akkor az esélyérték 0,75/0,25 = 3 (három az egyhez). Látható, hogy 0,5-nél nagyobb valószínűség esélyértéke 1-nél nagyobb szám. Általánosan, ha (az angol „odds” első betűjét használva) ***O-val jelöljük az esélyértéket*** , akkor

$$O = \frac{P}{1-P}. \qquad (11)$$

Ez a transzformáció tehát a 0 és 1 közötti valós számokat a 0 és ∞ közötti valós számokra képezi le. Amennyiben az esélyértékek logaritmusát vesszük, akkor a skála a –∞ és +∞ közötti valós számok tartományára terjeszthető ki: $L = \ln O$, (angolul „logit”). Ezt a skálát azonban elég ritkán alkalmazzák (20. ábra).



20. ábra
Az esély további mérőszámai, skálái. *P*: valószínűség; *O*: esélyérték; *L*: az esélyérték logaritmusa.



21. ábra
Szabad (a) és összekötött (b) dobókockák.

a

	30	25	30	29	28	25
	24	27	31	27	24	27
	28	30	39	32	24	29
	28	28	22	26	27	33
	27	24	26	21	31	27
	30	25	32	30	29	25

b

	40	41	46	12	9	21
	51	38	37	13	22	15
	42	49	52	8	20	17
	8	10	15	36	52	44
	11	16	9	45	39	35
	10	17	8	43	41	28

22. ábra
1000 dobás eredménye (1000 hosszúságú kísérletsorozat abszolút gyakoriságai) szabad (a) és összekötött (b) dobókockák esetében.

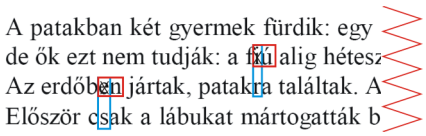
4.0. Összetett megfigyelés, függetlenség

Képzeljük el azt a jelenséget, amikor egy dobókockát és 6 darab 100 forintos pénzérmét egyszerre feldobunk. A jelenségben azt figyeljük meg, hogy hányast dobtunk a kockával, illetve, hogy hány érmén jött ki fej. A megfigyelés eredménye egy számpár, például (4,5). A jelenség lényegében változatlan marad, ha a kockát előbb dobjuk fel és az érméket csak utána. Ilyenkor a kockadobás eredményét előbb tudjuk meg, mint a fejek számát, ezért nevezzük őket „elsődleges”, illetve „másodlagos” megfigyelésnek. Ennél a megfigyelésnél akármi is az **elsődleges megfigyelés** eredménye, az a **másodlagos megfigyelés** eredményét nem befolyásolja. Ha azonban úgy járunk el, hogy csak annyi érmét dobunk fel, amennyi a kockadobás eredménye, akkor ez megszabja, hogy a hat lehetséges másodlagos jelenség közül melyiket hajtsuk végre.

Általánosan azt mondhatjuk, hogy az elsődleges és a végrehajtott másodlagos jelenség megfigyelésével összetett megfigyelést végzünk. A fenti példa alapján nem nehéz elképzelni, hogy többszörösen összetett jelenségek és megfigyelések is előfordulhatnak.

A következő példával először a függetlenség általános fogalmát szeretnénk tisztázni. Dobjunk fel két dobókockát (egy fehéret és egy feketét), és olvassuk le a dobott értékeket. Ennek a megfigyelésnek az eredményeként egy számpár adódik. A 21a ábrán ez épp a (3,2) számpár. Természetesen az is egy megfigyelés, ha az egyik kockával nem törődve csak a fehér (3), vagy csak a fekete (2) kocka értékét olvassuk le. Nevezzük ezeket rendre „fehér”, illetve „fekete” megfigyelésnek. Ebben az esetben eléggé kézenfekvő azt gondolni, hogy a fehér és a fekete megfigyeléseknek bizonyos értelemben semmi közük egymáshoz, tehát függetlenek.

Ha azonban a két kockát a 21b ábra szerint a testátlójuk mentén átfúrjuk és vékony cérnaszálakkal összekötjük, akkor a két kockának és ezáltal a fehér és a fekete megfigyeléseknek is lesz közük egymáshoz. Ez például abban mutatkozik meg, hogy egy hosszabb kísérletsorozat során megfigyelt abszolút gyakoriságok megváltoznak a szabad kockákon végzett megfigyelésekhez képest (22. ábra). Úgy tűnik, ha a fehér kockán az 1, 2, 3 számok valamelyike van felül, akkor a fekete kockán is inkább az 1, 2, 3 számok valamelyike kerül felülre. Hasonlóképpen, ha a fehér kockán a 4 az 5 vagy a 6 számok jönnek ki, akkor a fekete kockán is ezek



23. ábra
Egy magyar nyelven írt szövegből „rábökéssel” kiválasztott betű (×), illetve a mellette lévő (piros keret), és az alatta lévő (kék keret).

lesznek gyakrabban felül. Ez az eredmény azzal magyarázható, hogy az összekötések a fehér és a fekete kocka esetében is az 1, 2, 3, illetve a 4, 5, 6 számoknak megfelelő lapok közös csúcsainál vannak.

Egyszerűsítés végett a következő példában csak egy-egy esemény megfigyelésére koncentrálunk. A jelenség legyen az, hogy egy magyar nyelven írt szövegből „rábökéssel” kiválasztunk egy betűt, majd megfigyeljük a mellette (tőle jobbra), és az alatta lévőt is. Ha bármelyik helyen nem betű található, például szóköz vagy valamilyen írásjel, akkor a megfigyelés érvénytelen (23. ábra). Tekintsük az alábbi eseményeket:

A ***esemény***: „a **második** betű magánhangzó” (a 2. betű mgh)
B ***esemény***: „az **első** betű (amelyre ráböktünk) magánhangzó” (az 1. betű mgh)
$\overline{A}$ ***esemény***: „a második betű mássalhangzó” (a 2. betű msh)
$\overline{B}$ ***esemény***: „az első betű (amelyre ráböktünk) mássalhangzó” (az 1. betű msh).

Egy-egy hosszabb (100 hosszúságú) kísérletsorozat abszolút gyakoriságait a 24. táblázat mutatja, ahol sor- és az oszlopösszegeket is megadtuk.

a	B	$\overline{B}$	összesen
A	15	24	39
$\overline{A}$	25	36	61
összesen	40	60	100

b	B	$\overline{B}$	összesen
A	6	34	40
$\overline{A}$	36	24	60
összesen	42	58	100

24. táblázat
100 hosszúságú kísérletsorozat abszolút gyakoriságai a „rábökéssel” kiválasztott két egymás alatti (a) és két szomszédos (b) betű esetében.

Az oszlopösszegek a B esemény (az 1. betű mgh), illetve a $\overline{B}$ esemény (az 1. betű msh) bekövetkezésének az abszolút gyakoriságát mutatják. Számítsuk ki ezekre a szűkebb összességekre vonatkozó relatív gyakoriságokat, azaz a feltételes relatív gyakoriságokat (25. táblázat).

a		ha az 1. betű mgh	ha az 1. betű msh
a 2. betű	mgh	0,38	0,40
	msh	0,62	0,60

b		ha az 1. betű mgh	ha az 1. betű msh
a 2. betű	mgh	**0,14**	0,59
	msh	0,86	0,41

25. táblázat
100 hosszúságú kísérletsorozat feltételes relatív gyakoriságai a „rábökéssel” kiválasztott két egymás alatti (a) és két szomszédos (b) betű esetében, az 1. betű mgh, illetve az 1. betű msh szűkebb összességekre vonatkozóan.

A táblázat (a) részében azaz a két egymás alatti betű esetében az oszlopok lényegében megegyeznek egymással (0,38 ≈ 0,40, illetve 0,62 ≈ 0,60). Tehát **függetlenül** attól, hogy az 1. betű magánhangzó vagy mássalhangzó (tehát a B vagy a $\overline{B}$ esemény következik be), a 2. betű az eseteknek körülbelül a 0,4 részében magánhangzó (A esemény), a 0,6 részében mássalhangzó ($\overline{A}$ esemény).

Nem ez a helyzet a két szomszédos betű esetében, amit a táblázat (b) részében láthatunk. Magánhangzó után lényegesen ritkábban fordul elő magánhangzó, mint mássalhangzó után (0,14, 0,59-dal szemben), ezért azt mondhatjuk, hogy az említett két esemény, A és B **nem független** egymástól.

Az okok itt is érthetőek: a két egymás alatti betű esetében a függetlenség nem meglepő, hiszen a sortörések véletlenszerűek; a szomszédos betűk egymásutánját azonban a magyar nyelv szerkezete határozza meg, így a függetlenség nem is várható.

Egy B eseménytől függetlennek nevezünk egy A eseményt, ha az A esemény feltételes relatív gyakoriságai lényegében nem függenek attól, hogy B-re vonatkozóan vagy $\overline{B}$-re vonatkozóan számítottuk ki őket.

A és B független

a	B	$\overline{B}$	összesen
A	**0,15**	0,24	**0,39**
$\overline{A}$	0,25	0,36	0,61
összesen	**0,40**	0,60	1,00

A és B nem független

b	B	$\overline{B}$	összesen
A	**0,06**	0,34	0,40
$\overline{A}$	0,36	0,24	0,60
összesen	**0,42**	0,58	1,00

26. táblázat
100 hosszúságú kísérletsorozat relatív gyakorisági a „rábökéssel” kiválasztott két egymás alatti (a) és két szomszédos (b) betű esetében.

4.1. Feltételes valószínűség és függetlenség

Térjünk vissza az előző példához és használjuk a 24. táblázat abszolút gyakoriságait. A relatív gyakoriságokat könnyen kiszámíthatjuk, hiszen a kísérletsorozat $n = 100$ hosszúságú. Ezeket a 26. táblázatban foglaltuk össze. Először tekintsük a táblázat (b) részét (amikor az A és B esemény nem független). A B esemény bekövetkezésének relatív gyakorisága **0,42**, és **0,06** annak, amikor A esemény is bekövetkezik. Ennek a két számnak a hányadosa (0,06/0,42) ≈ **0,14**, ami megegyezik az a A esemény B eseményre vonatkozó feltételes relatív gyakoriságával (25b. táblázat első eleme). A nagy számok relatív gyakoriságokra vonatkozó tapasztalati törvénye szerint az említett gyakoriságok, illetve így, a belőlük képzett hányados is, a kísérletsorozat hosszúságának növekedtével, stabilitást mutatnak és a megfelelő valószínűségeket szolgáltatják:

$$\frac{k_B}{n} \approx P(B)\ , \qquad \frac{k_{A\cap B}}{n} \approx P(A\cap B)\ , \tag{12}$$

$$\frac{\frac{k_{A\cap B}}{n}}{\frac{k_B}{n}} = \frac{k_{A\cap B}}{k_B} \approx \frac{P(A\cap B)}{P(B)} = P(A\,|\,B)\ . \tag{13}$$

A $P(A\,|\,B)$ -vel jelölt valószínűséget az A esemény B eseményre vonatkozó feltételes valószínűségének nevezzük.

A relatív gyakoriságokra vonatkozó szorzási szabály mintájára a (13) összefüggés felhasználásával kapjuk a **valószínűségek szorzási szabályát**:

$$P(A\cap B) = P(A\,|\,B)P(B)\ . \tag{14}$$

Mivel $P(A\cap B) = P(B\cap A)$, ezért

$$P(A\,|\,B)P(B) = P(B\,|\,A)P(A) \tag{15}$$

is teljesül.

Ezek után a független események kritériumát ennek segítségével is megfogalmazhatjuk: **az *A* eseményt a *B* eseménytől függetlennek nevezzük,** ha

$$P(A\,|\,B) = P(A)\ , \tag{16}$$

ami ekvivalens azzal, hogy

$$P(A\cap B) = P(A)P(B)\ . \tag{17}$$

A (16) összefüggésnek megfelelő kapcsolat közvetlenül megfigyelhető a 25a. táblázatban található feltételes és a 26a. táblázatban található feltétel nélküli relatív gyakoriságok összehasonlításakor (0,38, illetve **0,39**). A 26a. táblázatban szereplő relatív gyakoriságok pedig nagyjából megegyeznek a (17) összefüggés alapján kiszámíthatókkal (például: **0,15** ≈ **0,39** · **0,40**) (27. megjegyzés).

27. megjegyzés
A függetlenség fogalmát a gyakorlatban kétféle értelemben használjuk. Az előző részben a szabad dobókockák esetében a „függetlenség” arra utal, hogy a két kockát akár külön-külön is feldobhattuk volna, és A az egyik (például a fehér), B pedig a másik (a fekete) megfigyeléssel kapcsolatos egy-egy esemény, tehát valóban semmi közük egymáshoz. Úgy gondoljuk, hogy ilyenkor már a kísérlet végrehajtásakor teljesül a függetlenség feltétele, így jogos ezen események (16) összefüggés szerinti matematikai függetlenségének a feltételezése is.

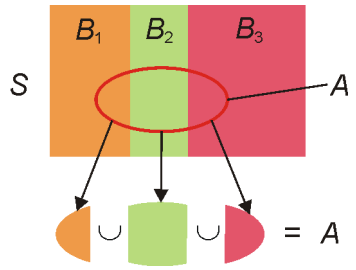
Máskor viszont éppen az a kérdés, hogy vajon két esemény független-e, és ezt – más lehetőségünk nem lévén – csak az események gyakorisága alapján kell eldöntenünk.

A feltételes valószínűség egy másik felhasználására nyújt lehetőséget a teljes valószínűség tétel. A tétel egyszerűsített változatának bemutatására használjuk a 28. ábrát. Ha az ábrán látható B_1, B_2, B_3 események teljes eseményrendszert alkotnak, azaz egymást kölcsönösen kizárják, de közülük egy mindig bekövetkezik, és A egy tetszőleges esemény, akkor

$$P(A) = P(A\,|\,B_1)P(B_1) + P(A\,|\,B_2)P(B_2) + P(A\,|\,B_3)P(B_3)\ . \tag{18}$$

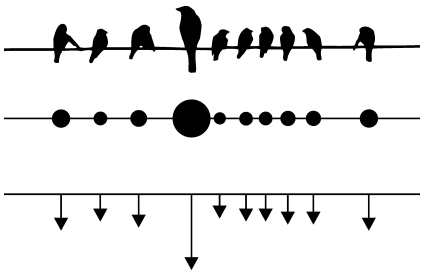
Ez ekvivalens azzal, hogy

$$P(A) = P(A\cap B_1) + P(A\cap B_2) + P(A\cap B_3)\ . \tag{19}$$

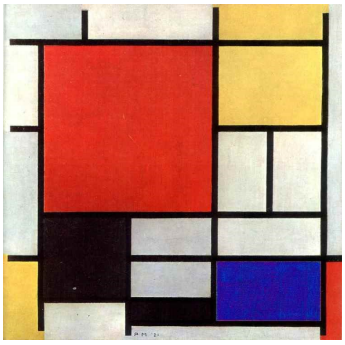


28. ábra
A teljes valószínűség tétel szemléltetése halmazok segítségével.

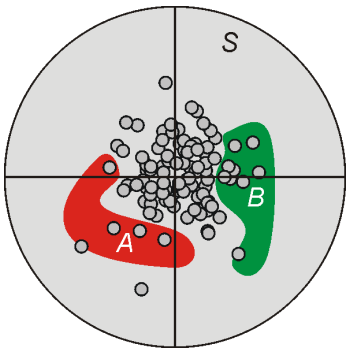
Az A eseménynek a B_1, B_2, B_3 egymást kölcsönösen kizáró eseményekkel vett közös része egymást kölcsönösen kizáró eseményeket eredményez. Ezek egyesítése épp az A eseményt adja meg, hiszen a B_1, B_2, B_3 események egyesítése a biztos esemény. A valószínűség 3. tulajdonsága szerint viszont egymást kizáró események egyesítésének valószínűsége az egyes események külön-külön vett valószínűségeinek az összege.



29. példa
Madarak „eloszlása” villanydróton, és az eloszlás kétféle szemléltetése: „tömegpontokkal” és erővektorokkal.



30. példa
Piet Mondrian: Kompozíció.



31. ábra
Céltábla mint alaphalmaz az eloszlás matematikai megfogalmazásához.

5.0. Az eloszlás szemléletes jelentése

Jutalomosztáskor a munkahelyi vezető a rendelkezésre álló keretösszeget szétosztja a beosztottak között. A matematika nyelvén ez azt jelenti, hogy a beosztottak halmazának elemeihez (minden egyes beosztotthoz) hozzárendelünk egy számot (pénzösszeget), amely persze 0 is lehet (azaz nem mindenki kap jutalmat). Ebédnél a ház asszonya szétosztja a tálnyi kelkáposzta főzeléket a családtagok között. A családtagok halmazának elemeihez itt is számokat rendelhetünk, például úgy, hogy ki hány merőkanállal kapott a főzelékből, vagy annak apján, hogy az egyes kelkáposzta adagoknak mennyi a grammokban mért tömege.

Különböző madarak különböző távolságokra ülnek a villanydróton. Eloszlásukat például úgy adhatjuk meg, hogy a drót mentén centiméterekben mért helyzetükhöz hozzárendeljük a newtonokban mért súlyukat (29. példa). A festőművész festékkel keni be a vásznat. Ennek eredményeként, mondjuk a piros festék úgy oszlik el a vásznon, ahogy azt a művész elképzelte. Ebben az esetben az eloszlás például a négyzetcentiméterekben kifejezett, adott területekhez rendelt grammokban mért festék mennyiségével adható meg (30. példa).

5.1. Az eloszlás matematikai megfogalmazása

Első lépésként egy halmazt kell megadnunk, amelyen az eloszlást vizsgáljuk. Legyen ez az S alaphalmaz egy kör alakú céltábla (31. ábra). Adjunk le száz lövést erre céltáblára légpuskával. Tegyük fel, hogy az ólomból készített légpuskalövedék mindegyike 1 g-os, és a céltáblát mindig eltaláljuk. Válasszuk ki S-nek egy tetszőleges A részhalmazát (piros terület), és vizsgáljuk meg, hogy ezen a területen belül hány gramm ólom található. A 31. ábra szerint ez 4,5 g. Hasonlóképpen kiválaszthatjuk S-nek egy tetszőleges másik B részhalmazát is (zöld terület), ahol – az előbbi eljárást követve – az ott található ólom 8,75 g. Ennek alapján az egyes részhalmazokhoz egy-egy számértéket rendelhetünk. Jelöljük $P(A)$-val az A halmazhoz $P(B)$-vel az B halmazhoz rendelt értéket. Esetünkben $P(A) = 4{,}5$, $P(B) = 8{,}75$. Mivel P függ attól, hogy melyik részhalmazt választjuk, ezért P egy függvény. Azt mondjuk, hogy ***P* az *S* alaphalmazon értelmezett halmazfüggvény**.

A 31. ábrából az is kitűnik, hogy az A és B diszjunkt halmazok egyesítéséhez a 13,25 érték rendelhető, mivel ennyi gramm ólom található összesen a két halmaz által kijelölt területen belül. Ez a matematika nyelvén azt jelenti, hogy

$$P(A \cup B) = P(A) + P(B), \qquad \text{ha } A \cap B = \emptyset \ , \tag{20}$$

azaz ***P* additív halmazfüggvény**.

Azt is megfigyelhetjük, hogy a (20) összefüggés alakilag azonos a valószínűség 3. tulajdonságával, ahol A és B egymást kizáró eseményeket jelölt. Ha a P additív halmazfüggvényről még azt is föltesszük, hogy **nem vesz fel negatív értékeket** (a valószínűség 1. tulajdonságának megfelelően), **és az üres halmazon felvett értéke 0** (a valószínűség 2b. tulajdonságának megfelelően), akkor a párhuzam már majdnem teljes. Az ilyen tulajdonságú halmazfüggvényt eloszlásnak nevezzük.

Amennyiben a céltáblára érkező lövedékek tömegét nem grammokban, hanem „hekto” grammokban mérnénk, akkor $P(S) = 1$ (hiszen 1 hg = 100 g és összesen 100 darab 1 g-os lövedéket lőttünk ki), ami a valószínűség 2a. tulajdonságával ekvivalens. Ilyenkor normált eloszlásról, vagy valószínűségeloszlásról beszélünk.

6.0. Valószínűségi változó

Egy jelenséggel kapcsolatban több dolgot is megfigyelhetünk. Ha ismét a kockadobás példáját vesszük, akkor

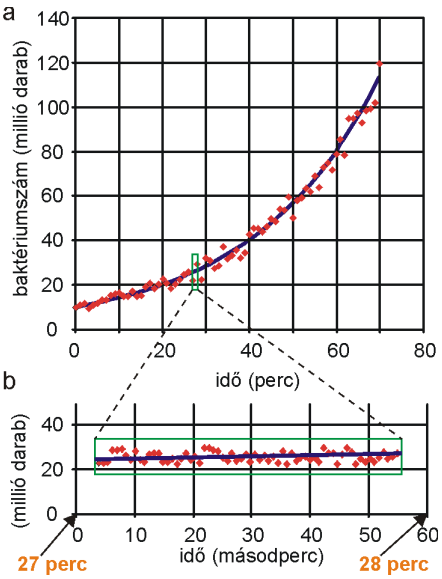
1. a ***dobott számon*** kívül megfigyelhetjük például azt is, hogy
2. a dobás kezdetétől a kocka megállásáig mennyi ***idő*** telik el;
3. a kocka megállása után az egyik adott éle mekkora ***szöget*** zár be az asztal élével; de azt is, hogy
4. az asztalon marad, vagy esetleg leesik a földre;
5. illetve azt, hogy melyik játékoshoz áll meg a legközelebb.

Az első három esetben ***kvantitatív***, azaz mennyiségi, tehát számokkal (és leg-

32. megjegyzés
Számokat a minőségekhez is rendelhetünk, de ezt valamilyen általunk kialakított szabály szerint kell tennünk. Például 1-et rendelünk ahhoz, ha a kocka az asztalon marad és 0-t, ha leesik onnan. Vagy a játékosokat is azonosíthatjuk számokkal, például a személyi számukkal. Ha a szabály egyértelmű, akkor ezeket a számokat is tekinthetjük egy valószínűségi változó felvett értékeinek.

Csak példaként említjük, hogy vannak olyan játék dobókockák, amelyek lapjain nem számok, hanem más fajta szimbólumok láthatók.

Az ilyen dobókocka esetében az a megfigyelés, hogy melyik lap van felül, nem szolgáltat valószínűségi változót, hiszen a megfigyelés eredményeként nem számokat kapunk. Ha azonban a szimbólumhoz számokat rendelünk, és megadjuk, hogy melyik szimbólumhoz melyik számot, akkor máris valószínűségi változóhoz jutunk. Bár ez az eljárás erőltetettnek tűnhet, a gyakorlatban sokszor van szükség ilyesmire (vö. 86. táblázat).



33. ábra
a) A baktérium kolónia szaporodása az 5. ábra szerint.
b) A baktérium kolónia szaporodása kb. a 28. percben. Ha igen gyorsan mérünk, mondjuk a duplázódási időnél (20 percnél) sokkal rövidebb idő (pl. 50 másodperc) alatt elég sok mérést (pl. 66-ot) el tudunk végezni, akkor a determinisztikus változás elhanyagolható.

többször mértékegységgel) megadható dolgot figyelünk meg. A másik kettőben a megfigyelés ***kvalitatív***, azaz minőségi, tehát számok automatikusan nem rendelődnek hozzájuk (32. megjegyzés). Ha egy jelenséggel kapcsolatban a megfigyelésünket **számokkal** jellemezzük, akkor azt mondhatjuk, hogy ezek a számok egy valószínűségi változó lehetséges értékei közül valók.

Egy **valószínűségi változót** azáltal adunk meg, hogy megmondjuk **mit és milyen körülmények között figyelünk meg, illetve hogyan mérünk.** Azt is lehet mondani, hogy egy valószínűségi változó lényegében egy **mérési, illetve megfigyelési eljárást jelent**. Ezek után ***nem azon kell meglepődnünk, ha egy változóról kiderül, hogy valószínűségi változó, hanem azon, hogyha nem az***. A matematikai modelljeinkben ugyanis absztrakcióval bármilyen változót definiálhatunk, de mihelyt ellenőrizni akarjuk azt, hogy egy modell mennyire felel meg tapasztalatainknak, megfigyeléseket, méréseket kell végeznünk, ahol a „véletlen”, azaz a számba nem vehető tényezők is szerepet kapnak.

6.1. A valószínűségi változó és a számszerű adatok kapcsolata

Amikor **megfigyeléseket**, méréseket végzünk, és ennek eredményeként ***számszerű adatokhoz*** jutunk, akkor azt is mondhatjuk, hogy ezeket az adatokat egy ***valószínűségi változó lehetséges értékei közül történő véletlen*** **kiválasztás** eredményezte. Másképpen mondva ilyenkor ***a megfigyelés ekvivalens egy véletlenszerű kiválasztással***. A kiválasztott számérték vagy benne van egy adott számhalmazban vagy nincs, és ennek megfelelően vagy bekövetkezett az adott esemény vagy nem. A véletlenszerű kiválasztás azt jelenti, hogy csupán a kiválasztás szempontjából a lehetséges értékek közül egyik sincs kitüntetve, tehát bármelyik kiválasztható, de ennek elemzésére még visszatérünk (13.1. rész).

Úgy is fogalmazhatunk, hogy a valószínűségi változó ***összes lehetséges értékét*** beletesszük egy „absztrakt zsákba”, ezzel ***egy halmazt adunk meg, amely ekvivalens az eseménytérrel***. ***Többszöri megfigyelés***, illetve mérés során ennek a halmaznak valamilyen ***részhalmazát*** kapjuk meg.

Ezek után térjünk vissza a baktérium kolónia szaporodásának problematikájára. Korábban már megállapítottuk, hogy a „változások” determinisztikus és statisztikus része általában együtt van jelen. Most mégis tegyük fel, hogy valamilyen trükkel szét tudjuk választani őket, és figyelmünket fordítsuk a statisztikus részre (33b. ábra). Ebben az esetben az n-szeri megfigyelés, illetve mérés egy n hosszúságú „kísérletsorozatnak” felel meg, ugyanis a „kísérleteket” ***lényegében azonos körülmények között végeztük***. Az eredmény pedig egy n-elemű adatrendszer.

A számszerű adatok esetén meg kell különböztetnünk a diszkrét és a folytonos változót, ami sokszor nem is olyan egyszerű feladat. A baktérium kolóniánál maradva például megszámlálhatjuk az egyedeket, ami a ***darabszám*** definíciója szerint csak egész szám lehet, tehát **diszkrét** változó. Ugyanakkor sok milliós egyedszám esetén a megszámlálás kivitelezhetetlen, helyette például a baktériumokat tartalmazó oldat fényszórásából következtetünk az egyedszámra. Ilyenkor egy **folytonos** fizikai mennyiség, nevezetesen a ***fényintenzitás*** segítségével fejezhető ki a jellemző adat. Megjegyezzük azonban, hogy a fényintenzitást sem tudjuk tetszőleges pontossággal mérni, hiszen ahhoz végtelen tizedes törtekre lenne szükségünk. Így a valódi különbség majd ott húzódik a kétféle változó között, hogy az a matematikai modell, amelyet használni akarunk, elméletileg folytonos-e, vagy diszkrét.

Vannak azért elég egyértelmű esetek, ilyen például a kockadobás is. Ha valódi kockát, azaz hexaédert használunk, akkor a dobás eredményét megadó valószínűségi változó csupán hat diszkrét értéket vehet föl. Az eseménytér az $S = \{1, 2, 3, 4, 5, 6\}$ halmaz. A halmaz elemeinek kiválasztása úgy is megvalósítható, hogy a dobások helyett az 1, 2, 3, 4, 5, 6 számokat egyesével felírjuk mondjuk egy-egy golyóra, amelyek egyéb szempontból megkülönböztethetetlenek. Ezután a golyókat beletesszük egy dobozba, és a dobozból csukott szemmel úgy húzunk, hogy minden húzás után a leolvasott értéket feljegyezzük, majd a kihúzott golyót visszatesszük. n dobás, vagy n húzás a fentiek szerint egy n-elemű adatrendszert eredményez. Amennyiben az adatok kiválasztásának sorrendje nem érdekel bennünket, akkor a legjobb jellemzésükre az adatrendszer eloszlásfüggvényét használhatjuk, amelyet definíció szerint a következő összefüggéssel adhatunk meg:

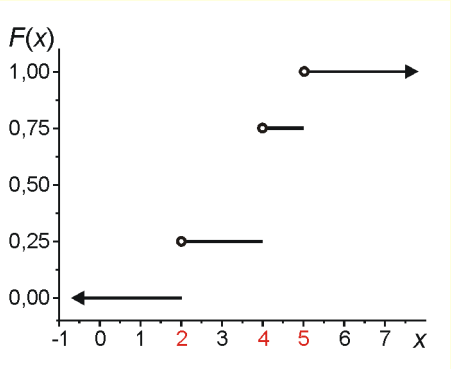
$$F(x) = \frac{\text{az } x \text{ - nél kisebb adatok darabszáma}}{n}. \qquad (21)$$

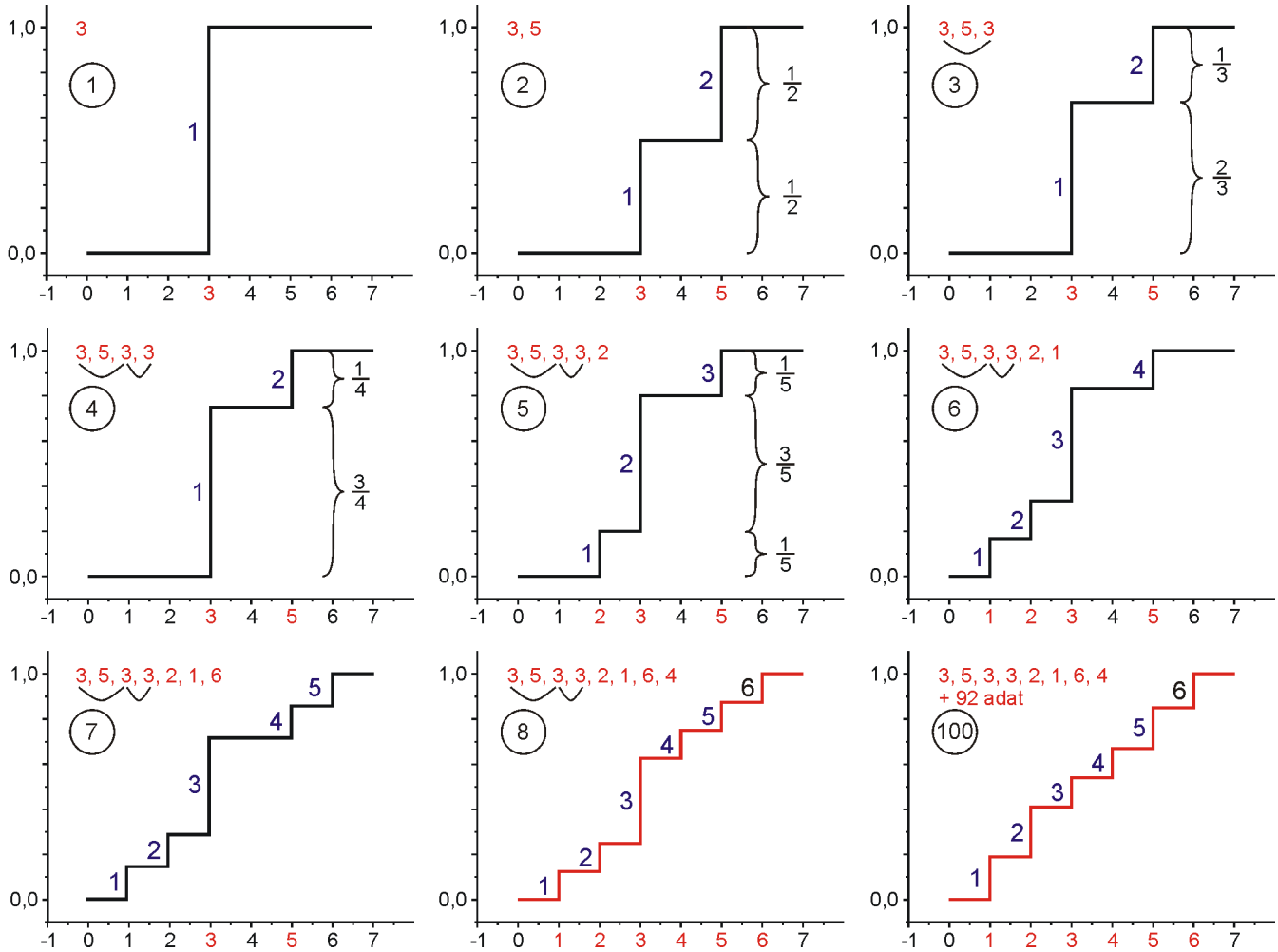
Látható, hogy $F(x)$ értékei 0 és 1 között változhatnak, ugyanis, a legkisebb adatnál kisebb és a legnagyobb adatnál nagyobb adatok darabszáma 0, emiatt $F(x) = 0$ egészen addig, amíg a nagyobb x-ek felé haladva a legkisebb adaton túl nem jutunk és $F(x) = 1$ mihelyt túljutottunk a legnagyobb, tehát az utolsó, n-edik adaton is. Az is nyilvánvaló, hogy $F(x)$ lépcsőfüggvény lesz, mert a darabszám diszkrét változó és azokban az intervallumokban, ahol az adatok darabszáma nem változik, $F(x)$ állandó értéket vesz föl, ahol viszont változik, ott a függvénynek mindig ugrása lesz. Amennyiben azonos értékű adataink nincsenek, akkor minden ugrás éppen $1/n$ nagyságú, ha vannak, akkor az ugrások $1/n$-nek egészszámú többszörösei is lehetnek.

Mintafeladat

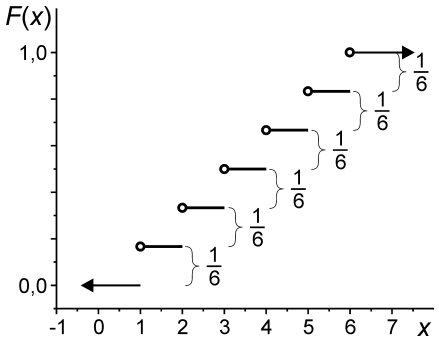
Ábrázoljuk grafikusan a [2, 4, 4, 5] adatrendszer eloszlásfüggvényét!

Megoldás: Mivel 2 a legkisebb adat, ezért 2-ig $F(x) = 0$. A nagyobb x-ek felé haladva, 2-t elhagyva $F(x) = 0{,}25$, hiszen mondjuk 2,3-nél csak 1 darab kisebb adatunk van és $n = 4$. Ez az érték nem változik egészen 4-ig. 4-et elhagyva $F(x) = 0{,}75$, hiszen mondjuk 4,1-nél 3 darab kisebb adatunk van (2, 4, 4). 5 felett $F(x) = 1$, hiszen mind a négy adaton túljutottunk. Az ábra az $F(x)$ eloszlásfüggvényt szemlélteti. A karikák azt jelentik, hogy ott a függvény még az alsó értéket veszi föl, azaz: $F(2) = 0$, $F(4) = 0{,}25$, $F(5) = 0{,}75$. Kicsit továbbmenve (2+, 4+, 5+, azaz 2-nél, 4-nél, 5-nél kicsit nagyobb értékek esetében) azonban $F(2+) = 0{,}25$, $F(4+) = 0{,}75$, $F(5+) = 1$.





34. ábra
Egy kockadobás sorozat eredményének megadása az adatrendszerek eloszlásfüggvényeivel. A karikában lévő szám mutatja azt, hogy hányadik dobásról van szó. A fölötte lévő piros számok az addigi dobások eredménye. A lépcsőfüggvények függőleges vonalai természetesen nem részei a függvényeknek, csak a jobb szemléltetés kedvéért rajzoltuk be őket. A lépcsőfokokat szintén megszámoztuk (kék számok).



35. ábra
Az ideális kockára, azaz a kocka modelljére vonatkozó dobás eredményének mint valószínűségi változónak az eloszlásfüggvénye.

36. ábra
Az ideális kockával történő dobások eredményéhez hozzárendelt diszkrét valószínűségek szemléltetése.

A kockadobás példájánál maradva használjuk ezt a szemléltetést arra, hogy az egymás utáni dobások eredményét megadjuk.

A 34. ábrán megfigyelhetjük, hogy lépcsőmagasságok a relatív gyakoriságokkal egyenlők. Pl. az 5. dobás után $F(2+) - F(2) = 1/5$; $F(3+) - F(3) = 3/5$ és $F(5+) - F(5) = 1/5$. Az is látható, hogy 8. dobásnál már mind a 6 szám legalább egyszer szerepelt, így ettől kezdve az adatrendszerek eloszlásfüggvénye 6 lépcsőfokból áll (piros görbék). A további dobások (pl. 100 dobás) után a relatív gyakoriságok a ***nagy számok tapasztalati törvényének megfelelően*** stabilizálódást mutatnak, így a lépcsőmagasságok különbözősége csökken.

Határozzuk meg ezután az ***ideális kockára, azaz a kocka modelljére vonatkozó*** dobások eredményének mint ***valószínűségi változónak*** az eloszlását. Az ideális kocka különlegessége éppen az, hogy mind a 6 eredmény egyformán valószínű. Ezért a hozzárendelhető matematikai modell például egy olyan eloszlásfüggvénynyel adható meg, amelyben a lépcsőmagasságok – a 34. ábra utolsó függvényén is látható kicsit különböző relatív gyakoriságok helyett – az **egyenlő valószínűségek** (35. ábra).

Az adatrendszer eloszlásfüggvényéhez képest (vö. (21)) a valószínűségi változó eloszlásfüggvényét kicsit másként definiáljuk. Jelöljük ξ-vel (görög kszi bnetűvel) a valószínűségi változót, ekkor

$$F(x) = P(\xi < x) = \sum_{x_j < x} P(\xi = x_j) \qquad (22)$$

annak a valószínűsége, hogy a ξ valószínűségi változó az adott x-nél kisebb értéket vesz föl. Így, ha például $x = 3{,}2$, akkor $F(x) = 0{,}5$, mivel annak a valószínűsége, hogy 3,2-nél kisebbet dobjunk, azaz 1-et, 2-t vagy 3-at, éppen 0,5. A valószínűségi változó eloszlása jellemezhető a lépcsőmagasságokkal, azaz a valószínűségekkel is:

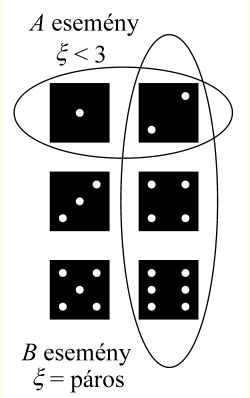
$$p_j = P(\xi = x_j)\,. \qquad (23)$$

Esetünkben x_j az 1, 2, 3, 4, 5, 6 értékeket veheti föl, ahol $p_j = 1/6$. Ezt szemlélteti a 36. ábra, ahol a számegyenes minden egyes x_j pontjába p_j nagyságú függőleges vonalat, „pálcikát” húztunk. (Vö. A madarak „eloszlása” villanydróton, 29. ábra.)

Egy valószínűségi változóról az eloszlása mindent elárul. Ennél többet nem is lehet róla tudni. Segítségével viszont összetett események valószínűségeit is meg tudjuk határozni. A kockadobás esetében például arra is alkalmas, hogy egy igazi kocka „jóságát” teszteljük azáltal, hogy a valószínűségi változó eloszlásfüggvényét összehasonlítjuk egy dobássorozat során gyűjtött adatrendszer eloszlásfüggvényével. Erre a problémára is visszatérünk (vö. 15.8. rész).

Mintafeladat

Független-e az a két esemény, ha egy (ideális) kockával egyszer dobunk, hogy az eredmény 3-nál kisebb (A esemény), illetve, hogy páros (B esemény)?



Megoldás: A feltett kérdést kétféle módon is eldönthetjük:

1. Annak a valószínűsége, hogy az A esemény és B esemény is bekövetkezik, ugyanakkora-e, mint a két esemény külön-külön vett valószínűségeinek szorzata? $P(A\cap B) = P(A)P(B)$?
2. Az A esemény B eseményre vonatkozó feltételes valószínűsége megegyezik-e A esemény feltétel nélküli valószínűségével? $P(A|B) = P(A)$?

Annak a valószínűsége, hogy a dobás eredménye 3-nál kisebb (A esemény), leolvasható a 35. ábrán látható eloszlásfüggvényről: $P(A) = P(\xi < 3) = F(3) = 2/6 = 1/3$. A páros (2, 4, 6) eredmények valószínűsége külön-külön (1/6), a 36. ábráról olvasható le, és az egymást kizáró események additív tulajdonsága miatt összeadódnak, tehát a B esemény valószínűsége: $P(B) = P(\xi = \text{páros}) = 3/6 = 1/2$.

1. szerint: $P(A)P(B) = (1/3)(1/2) = 1/6$, másrészről a két halmaz metszete, $(A\cap B)$ csak a 2-es dobás, aminek valószínűsége, $P(A\cap B)$ a 36. ábra alapján valóban 1/6.

2. szerint: mivel a (13) definíciós egyenlet szerint $P(A|B) = P(A\cap B)/P(B)$, ezért $P(A|B) = (1/6)/(1/2) = 1/3$. Másrészről a 2, 4, 6 számok (3 darab) közül csak (1 darab) a 2 kisebb 3-nál, ami szintén a $P(A|B) = 1/3$ eredményre vezet.

Látható, hogy minkét megfogalmazásra pozitív választ kaptunk, tehát a fenti két esemény (A, B) független egymástól.

A	B	C	D
85	85,2	85,15	85,153
85	85,0	85,02	85,021
85	85,0	84,96	84,965
85	84,9	84,95	84,946
85	84,7	84,73	84,727
85	85,2	85,22	85,216
85	84,9	84,94	84,935
85	85,4	85,44	85,441
85	85,0	84,99	84,987
85	85,0	85,02	85,016
85	85,3	85,28	85,282
85	85,1	85,09	85,091
85	84,8	84,80	84,804
85	85,0	85,04	85,037
85	84,7	84,72	84,723
85	85,1	85,09	85,087
85	85,3	85,29	85,293
85	85,1	85,12	85,122
85	85,0	84,98	84,976
85	85,1	85,12	85,123
85	85,0	85,00	85,005
85	84,9	84,91	84,909
85	85,1	85,07	85,074
85	85,0	84,99	84,990
85	84,9	84,94	84,936

mind egyenlő

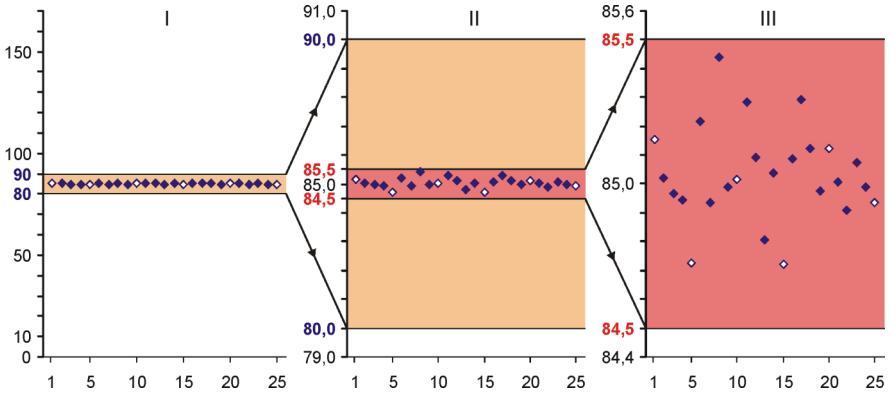
mind különböző

37. ábra
A szórt fény intenzitás mint folytonos változó mW/m^2 egységekben, négy különböző érzékenységű mérőállásban mérve.
A oszlop:0, B oszlop:1, C oszlop:2, D oszlop:3 tizedes jegyre kerekítve.

7.0. Folytonos valószínűségi változó jellemzése, eloszlásfüggvénye

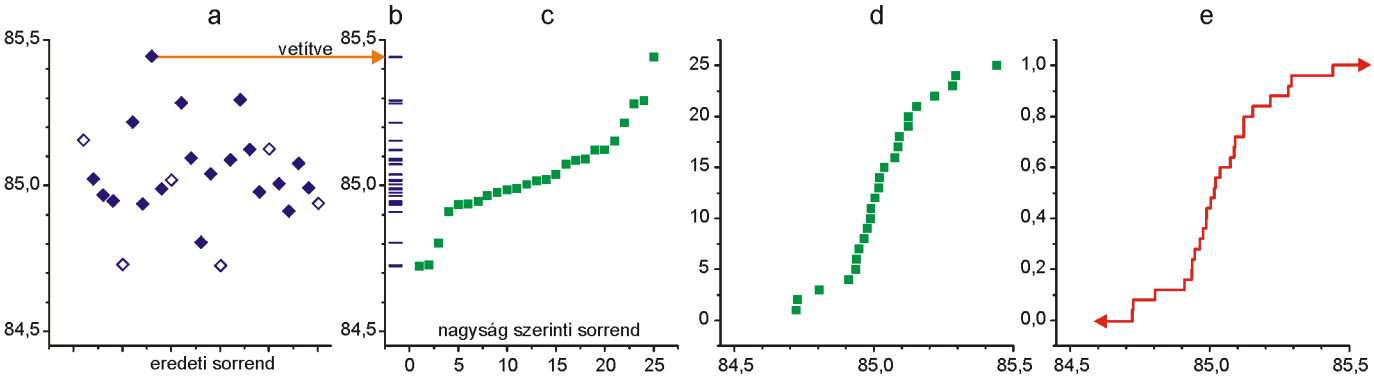
Tekintsük ismét a példaként már felhozott baktérium kolóniát. **Elvileg** a kolónia egyedszáma a 28. perc elteltével, a 29. perc kezdetén egy jól **meghatározott érték**. A baj csak az, hogy nem tudjuk megmondani, hogy mennyi. A kérdés megválaszolhatósága érdekében tegyük fel, hogy sikerült a „változások" determinisztikus és statisztikus részét egyértelműen szétválasztanunk például úgy, hogy mesterségesen leállítottuk a kolónia szaporodását. Így alkalmunk nyílik kizárólag a statisztikus rész tanulmányozására. Tegyük fel továbbá, hogy a egyedszám mint **diszkrét** változó helyett, a vele arányos szórt fény intenzitást mint **folytonos** változót mérjük, mondjuk mW/m^2 egységekben. A 37. ábra A, B, C, D oszlopaiban egy képzeletbeli mérési sorozatot mutatunk be egyre érzékenyebb mérőállásban mérve. A folytonos megjelölés azt jelentené, hogy a 25-elemű adatrendszer egyes értékei a valóságban végtelen tizedes törtek, és csak a mérőberendezés érzékenységén múlik, hogy hány tizedes jegyre tudjuk meghatározni őket. A véges tizedes törtek viszont nem változhatnak folytonosan, hiszen az utolsó tizedes jegybeli különbség mindig ugrást, azaz diszkrét változást eredményez. A dolog lényegét tekintve csak addig növeltük a tizedes jegyek számát, amíg mind a 25 érték különbözik egymástól.

Az A oszlopban minden érték egyenlő, ami a változatlanságra utal. A valóságban ez csak azt jelenti, hogy minden érték biztosan 84,5 és 85,5 között van, tehát a legkisebb kimutatható változás 1, de ezek az adatok ennél kevésbé térnek el egymástól. A B oszlopban a legkisebb változás 0,1, a C oszlopban 0,01, a D oszlopban pedig 0,001.



38. ábra
Az eredeti szórt fény intenzitások számértéke három különböző skálán (I: 0–170, II: 79–91, III: 84,4–85,6) ábrázolva. A jobb összehasonlíthatóság kedvéért fehér közepű szimbólumokkal jelöltük a vízszintes tengelyen olvasható sorszámoknak megfelelő értékeket.

Az egyre érzékenyebb mérőállás grafikusan a skála széthúzásával is szemléltethető. A 38. ábra az eredeti, kvázi folytonos 25-elemű adatrendszert (a 37. ábra D oszlopának megfelelően) ily módon mutatja be három különböző skálán (I, II, III). A 38.I ábrát megfigyelve első ránézésre azt láthatjuk, amit várunk, azaz „nincs változás", és az egymás utáni mérések látszólag nem különböznek egymástól, hi-

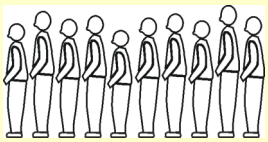


39. ábra
A szórt fény intenzitás statisztikus „változásának" többféle szemléltetése. a: eredeti időbeli sorrend szerint, b: egy helyre vetítve c: nagyság szerinti sorrendben, d: a tengelyek felcserélése után, e: az adatrendszer eloszlásfüggvényével. (A 34. ábrához hasonlóan a függőleges vonaldarabok itt sem részei a függvénynek.)

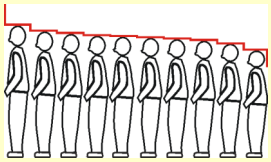
41. megjegyzés
Egy adatrendszer eloszlásfüggvényét jól szemléltethetjük az iskolai tornasorral.

Feldolgozandó adataink legyenek egy tanulócsoport tíz fiú tagjának testmagasságértékei (például cm-ekben mérve). Az alábbi ábrasor azt mutatja be, hogy hogyan származtatható az adatrendszer eloszlásfüggvénye.

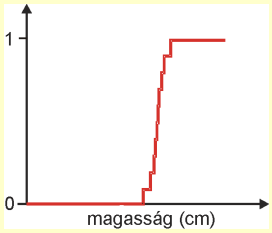
1. „eredeti” sorrend, ami valójában nem érdekel bennünket



2. nagyság szerinti sorrend (tornasor)

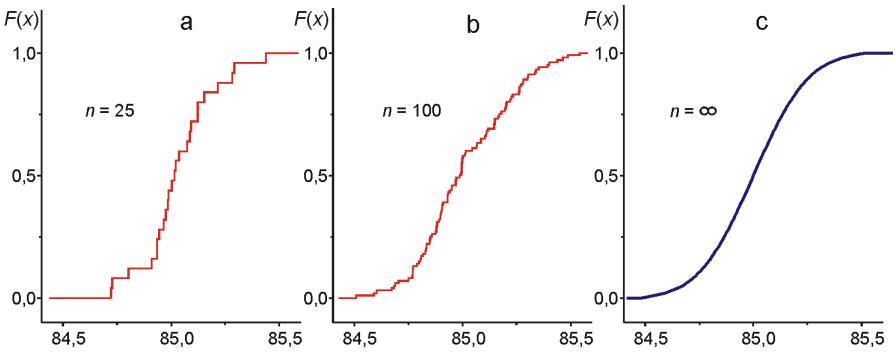


3. a fejtetőket burkoló lépcsős görbe -90°-os elforgatása és átskálázása



szen a kolónia szaporodását mesterségesen leállítottuk. Ha azonban fokozatosan széthúzzuk a skálát (38.II, III ábrák), az egyes mért értékek egyre jobban elkülönülnek. A kérdés az, hogy hogyan jellemezhetjük ezt a most már kizárólag statisztikusnak mondható „változást”.

Mivel úgy tűnik, hogy az adatok és azok eredeti időbeli sorrendje között nincs kapcsolat, ezért erről meg is feledkezhetünk. Így akár egy helyre vetítve vagy nagyság szerinti sorrendben is ábrázolhatjuk őket (39bc. ábra). Amennyiben a nagyság szerinti ábrázolás után a tengelyeket felcseréljük (39d. ábra) és az 1-től 25 -ig terjedő skálát (25-tel való osztással) 0 és 1 közé transzformáljuk, akkor az adatrendszer $F(x)$ eloszlásfüggvényéhez jutunk (39e. és 40a. ábra) (41. megjegyzés).



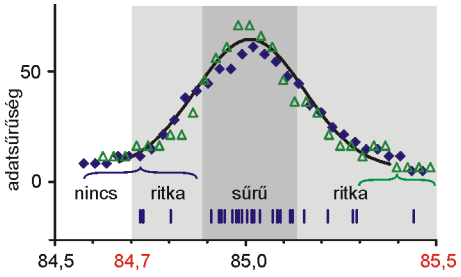
40. ábra
mW/m^2 egységekben mért szórt fény intenzitás mint folytonos **valószínűségi változó** (c) és ennek lehetséges értékeiből véletlenszerűen kiválasztott **adatrendszerek** (a,b) eloszlásfüggvényeinek szemléltetése.

Ez a lépcsőfüggvény 25 lépcsőből áll, hiszen mind a 25 adat különböző és a lépcsőmagasságok értéke 1/25. Hasonló a helyzet akkor is, ha sokkal több adatot gyűjtünk össze. n = 100 adat esetén 100 lépcsőfok jellemzi az eloszlásfüggvényt, amennyiben feltesszük, hogy mérőeszközünk a legkisebb különbségeket is érzékeli (40b. ábra). Folytonos változó esetén ugyanis 0 a valószínűsége annak, hogy két mért érték tökéletes egyezést mutasson. Ebben az esetben a lépcsőmagasságok értéke 1/100.

Az adatok számának további növelése ($n = \infty$) a lépcsők eltűnéséhez vezet, és gyakorlatilag egy folytonos függvényt eredményez. Ez a függvény tekinthető az adott körülmények közötti szórt fény intenzitás mint **folytonos valószínűségi változó eloszlásfüggvényének** (40c. ábra).

7.1. Adatsűrűség, gyakorisági eloszlás, relatív gyakorisági eloszlás, hisztogram, gyakoriságsűrűség, relatív gyakoriságsűrűség

Bár ***az F(x) eloszlásfüggvény minden esetben egyértelműen jellemzi az adatrendszert, illetve a valószínűségi változót***, grafikonjuk első ránézésre nem sokat mond, nem eléggé szemléletes. Adatrendszer esetén annyi kiolvasható belőle, hogy mekkora egy adott x értéknél kisebb adatok relatív darabszáma (vö. (21)), illetve relatív gyakorisága. Valószínűségi változó esetén pedig az, hogy mekkora a valószínűsége annak, hogy a ξ valószínűségi változó x-nél kisebb értéket vesz föl (vö. (22)).



42. ábra
Az adatok sűrűségének jellemzése.
0,2 hosszúságú (△) és 0,3 hosszúságú (◆) intervallumra kiszámított adatsűrűségek. A ritkábban, illetve sűrűbben elhelyezkedő adatokat egészen jól szemlélteti a folytonos görbe is.

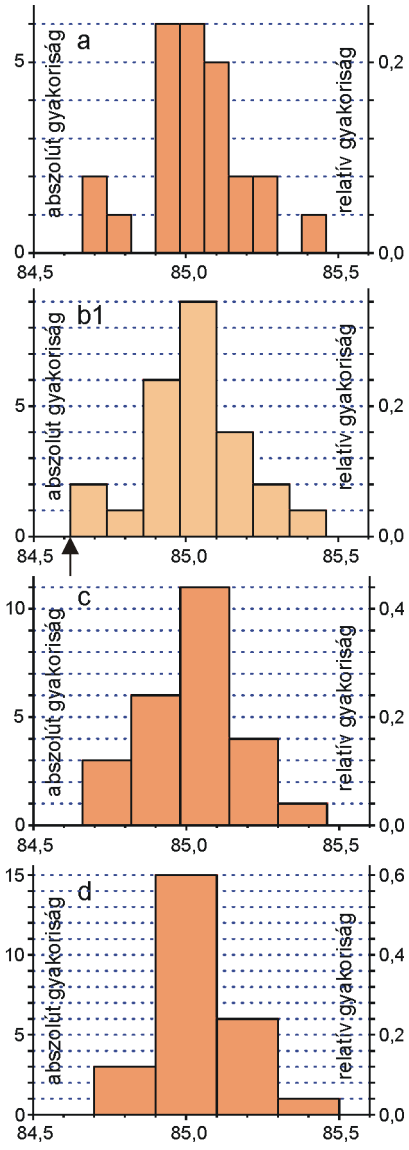
Tekintsük a 39b. ábrán látható, egy helyre vetített adatokat megjelenítő vonalkákat, de a könnyebb áttekinthetőség kedvéért forgassuk el -90°-kal (42. ábra). Megfigyelhetjük, hogy 84,7-nél kisebb adat nem szerepel a sorban. Ha a nagyobb értékek felé haladunk, először csak ritkán fordul elő egy-egy adat, 84,9-ig mindössze 3 darab. Ezt követően az adatok besűrűsödnek, de 85,2 környékén már újra elég ritkán követik egymást, és 85,5-nél nagyobb adat megint nem szerepel. Az ábra alapján azt gondolhatjuk, hogy az ***adatok sűrűsége*** lehet az a jellemző, amely szemléletes tartalommal bír, és amelyet a vonalkák fölé rajzolt zöld illetve kék szimbólumok valamint a fekete görbe juttat kifejezésre. Ez utóbbi eredetére később térünk vissza (vö. 7.2. rész).

Esetünkben az adatsűrűséget például úgy határozhatjuk meg, hogy választunk egy intervallumot (a 42. ábrán például a zöld vagy a kék vízszintes kapcsos zárójelnek megfelelően), amely elegendően hosszú ahhoz, hogy az adatrendszer több

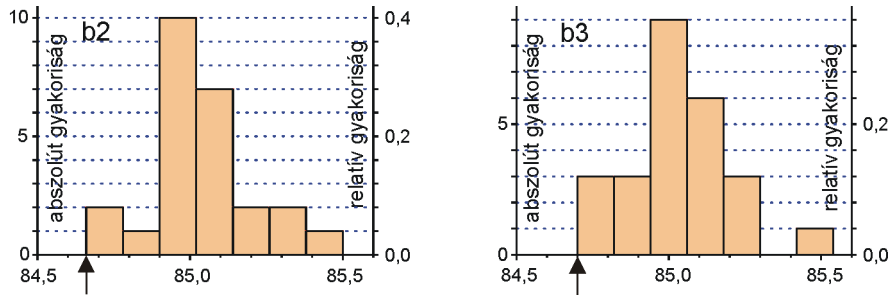
elemét is tartalmazhatja, de azért az összes adatnál mindig lényegesen kevesebbet. Ennek közepét beállítjuk egy adott x értékre, és megnézzük, hogy hány darab adat található az intervallumon belül, majd ezt a számot elosztjuk az intervallum hosszával. Így bármelyik x értékhez hozzárendelhetünk egy adatsűrűség értéket. Ha x-szel végighaladunk a 84,5 és 85,5 közötti tartományon, akkor az adatsűrűség kezdetben növekszik, majd csökken, de hogy pontosan hogyan, az láthatóan a választott intervallum hosszától is függ, de nem olyan nagyon (42. ábra).

Az adatsűrűség jellemzésének egy másik módja a gyakorisági eloszlás, illetve a relatív gyakorisági eloszlás megadása. Ennek érdekében osszuk fel a számegyenes adatainkat is tartalmazó részét intervallumokra, és számoljuk meg, hogy az így kapott osztályokban hány adat található, ezzel meghatározhatjuk az egyes osztályokban a gyakoriságokat, illetve a relatív gyakoriságokat. Megjegyezzük, hogy az osztályokat nem kell feltétlenül ugyanakkorára választanunk, de célszerű.

Használjuk ismét a 37. ábra D oszlopának adatait. Természetesen, mivel az osztályhatárok megválasztása önkényes, ugyanazokból az adatokból többféle gyakorisági eloszlás is készíthető, melyeket a jobb áttekinthetőség kedvéért oszlop diagramokkal szemléltethetünk. Nem azonos osztályszélesség esetén is célszerű a következő ábrázolásmód: **minden osztály fölé olyan téglalapot** („oszlopot”) **rajzolunk, melynek területe arányos az osztályba eső adatok gyakoriságával**. Az így kapott ábrát hisztogramnak nevezzük. Egy hisztogramon belül az ***azonos osztályszélesség*** választás épp azért előnyös, mert ilyen esetben a téglalap területe és a magassága arányos egymással. A téglalapok magassága, azaz a függőleges tengely skáláján feltüntetett értékek ilyenkor vagy az abszolút gyakoriságok, vagy a relatív gyakoriságok lehetnek (43. és 44. ábra).



43. ábra
A 37. ábra D oszlopának adatai alapján készített
a) 0,08; b1) 0,12; c) 0,16; d) 0,2
osztályszélességű hisztogramok 84,5-es kezdőértékkel és egy hisztogramon belül azonos osztályszélességgel.



44. ábra
A 43b1. ábrán látható hisztogramnak megfelelő további hisztogramok ugyancsak 0,12-es osztályszélességgel, de különböző kezdőérték esetén. Az első nem üres osztály kezdőértéke (nyíllal jelölve):
b1) 84,5+0,12; b2) 84,5+0,16; b3) 84,5+0,2.

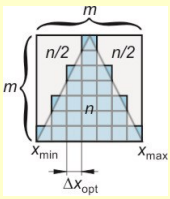
Van olyan eset, amikor az adatok már eleve csak valamilyen csoportosításban állnak rendelkezésünkre. A 45. táblázatban egy üzem eFt-ben (ezer forint) megadott fizetési kimutatása olvasható, ahol az összeghatárok különböző mértékben változnak. Ez a táblázat az adott osztályokhoz tartozó gyakoriságokat tartalmazza, amelyekből a relatív gyakoriságok is meghatározhatók. Bár ezeket a mennyiségeket is ábrázolhatjuk egy diagramon (46a. ábra), hisztogramhoz úgy juthatunk, ha a megadott abszolút gyakoriságokat, illetve relatív gyakoriságokat a hozzájuk tartozó osztályszélességgel elosztjuk és az így kapott gyakoriságsűrűségeket vagy relatív gyakoriságsűrűségeket ábrázoljuk (46b. ábra).

osztályhatárok (eFt)	abszolút gyakoriság	relatív gyakoriság (kerekítve)	gyakoriság-sűrűség	relatív gyakoriság-sűrűség
120 ≤ 130	124	0,14	12,4	0,014
130 ≤ 140	195	0,22	19,5	0,022
140 ≤ 160	293	0,33	14,65	0,0165
160 ≤ 200	195	0,22	4,875	0,0055
200 ≤ 300	80	0,09	0,8	0,0009
összesen	887	1		

45. táblázat
Egy üzem eFt-ben megadott fizetési kimutatása.

47. megjegyzés
A hisztogram akkor „esztétikus”, ha nem hézagos, de azért van szerkezete, azaz nincs minden adat egy-két osztályba bezsúfolva. Itt a sokféle lehetőség közül csak egyet mutatunk be.

Ha egy négyzet alakú diagramban akarjuk az „optimális” hisztogramot megrajzolni, akkor az intervallumok száma körülbelül megegyezik az egy intervallumba eső elemek maximális számával, mindkettőt jelöljük *m*-mel.



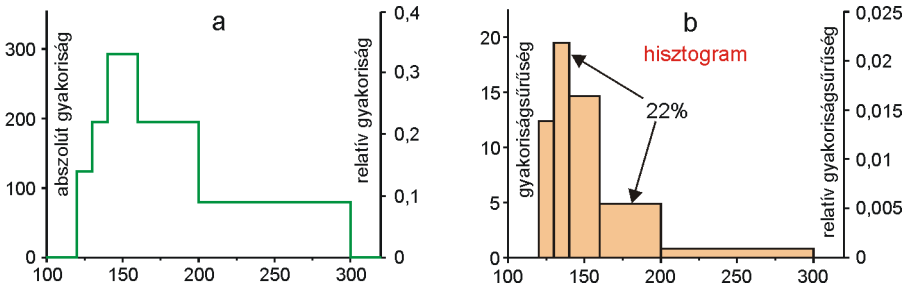
Ekkor egy elem egy négyzet alakú területet foglal el (a hosszúkás téglalap helyett). Az ábra alapján az intervallumok optimális száma:

$$m = \sqrt{2n}$$.

Az intervallumok optimális méretét (Δx_{opt}) megkaphatjuk, ha a legnagyobb (x_{max}) és a legkisebb (x_{min}) adat különbségét elosztjuk az optimális intervallumszámmal:

$$\Delta x_{opt} = \frac{x_{max} - x_{min}}{m}$$,

majd ezt általában kerekítjük.

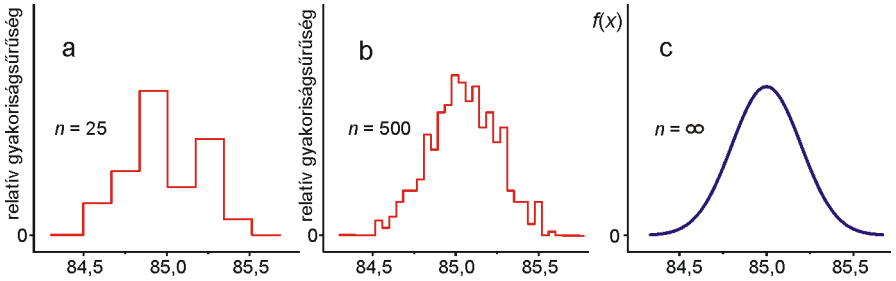


46. ábra
A 45. táblázatban szereplő adatok szemléltetése: a) abszolút gyakoriságok és relatív gyakoriságok; b) hisztogram (47. megjegyzés).

7.2. Sűrűségfüggvény

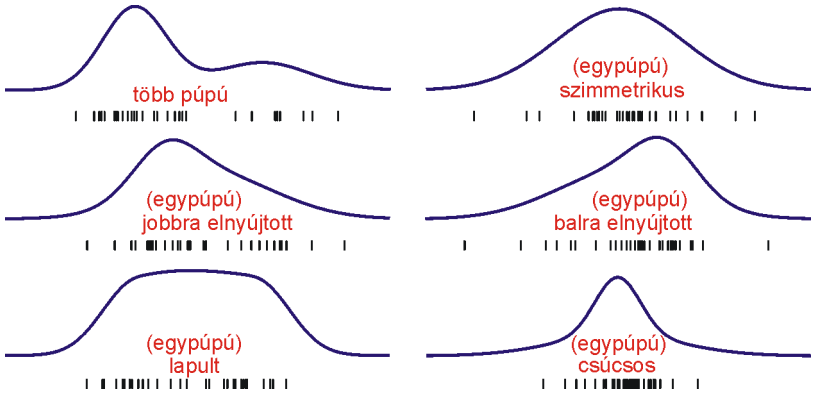
Emlékeztetőül megjegyezzük, hogy a szóban forgó folytonos valószínűségi változó a baktériumkolónia egyedszámával arányos szórt fény intenzitás mW/m^2 egységekben mérve. A 43. és 44. ábra azt mutatja, hogy míg az ***eloszlásfüggvény egyértelműen*** jellemzi az adatrendszert, ***ugyanazokból az adatokból igen különböző hisztogramok*** készíthetők, de néhány közös vonás azért ezekre is jellemző. Megfigyelhetjük például, hogy az összes ábrázolt hisztogram „kipúposodik” a közepe tájékán nagyjából ugyanannál az értéknél, de ezen túlmenően még a „szélességük” is hasonló mértékű. Ha a függőleges tengelyen a ***relatív gyakoriság-sűrűséget*** ábrázoljuk, továbbá az adatok számát növeljük, az osztályok szélességét pedig csökkentjük (és ezt akármeddig folytathatnánk), akkor hisztogramjaink durva lépcsős burkoló görbéi egyre jobban kisimulnának és egyetlen folytonos, sima görbébe mennének át (48. ábra). Ez a függvény tekinthető az adott folytonos valószínűségi változó sűrűségfüggvényének, $f(x)$-nek.

49. megjegyzés
Ha a valószínűségi változót sokszor megfigyeljük, akkor ott helyezkednek el sűrűbben az adatok, ahol a sűrűségfüggvény éréke nagyobb. A sűrűségfüggvény az eloszlásfüggvényhez hasonlóan elméletileg -∞ és +∞ között értelmezhető, de míg $F(-\infty) = 0$ és $F(+\infty) = 1$; addig $f(-\infty) = f(+\infty) = 0$. Mivel $f(x)$ a relatív gyakoriságsűrűséget méri, a görbe alatti összterület szükségszerűen 1.



48. ábra
Az adatok számának növelésével és az osztályszélesség csökkentésével a hisztogramot burkoló görbe a sűrűségfüggvényt közelíti.

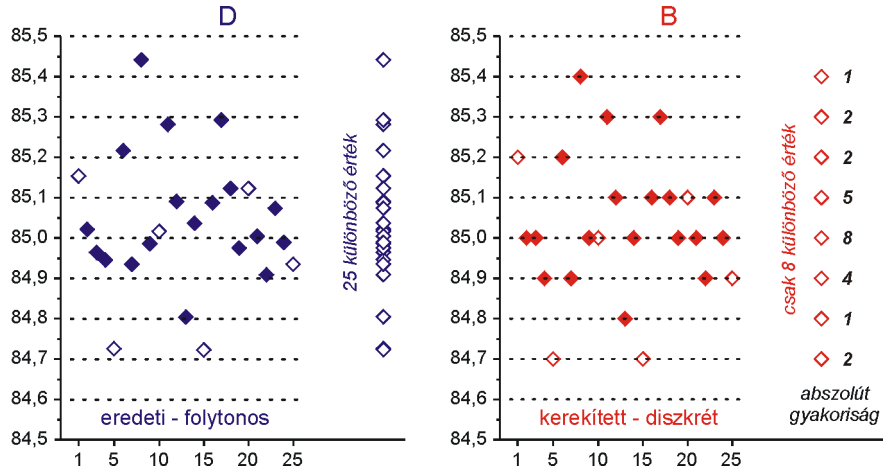
A ***sűrűségfüggvény*** valójában ugyanazt fejezi ki, mint az ***eloszlásfüggvény***, de másként, annál kicsit szemléletesebben (49. megjegyzés). **Alakja** többféle is lehet, a legfontosabb típusokat kvalitatívan megfogalmazva a 50. ábra mutatja.



50. ábra
A sűrűségfüggvény jellegzetes típusai és a hozzájuk tartozó adatok szemléltetése.

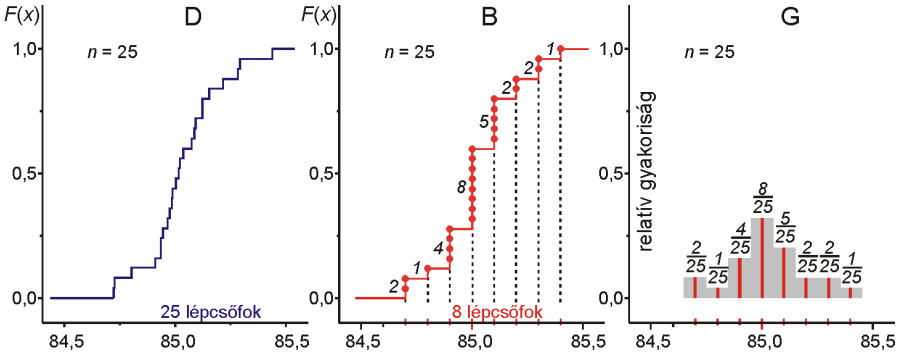
8.0. Diszkrét valószínűségi változó jellemzése, eloszlásfüggvénye

A folytonos, illetve diszkrét valószínűségi változó közötti leglényegesebb különbség megértéséhez a 37. ábra D és B oszlopának összevetése vezethet el bennünket. A valószínűségi változó továbbra is a baktériumkolónia egyedszámával arányos szórt fény intenzitás mW/m^2 egységekben mérve, de míg a D oszlopban elvileg csak végtelen tizedestörtekkel megadható értékek szerepelnek (folytonos eset), a B-ben az egy tizedesjegyre kerekített, ***diszkrét értékeket*** adtuk meg (51. ábra).

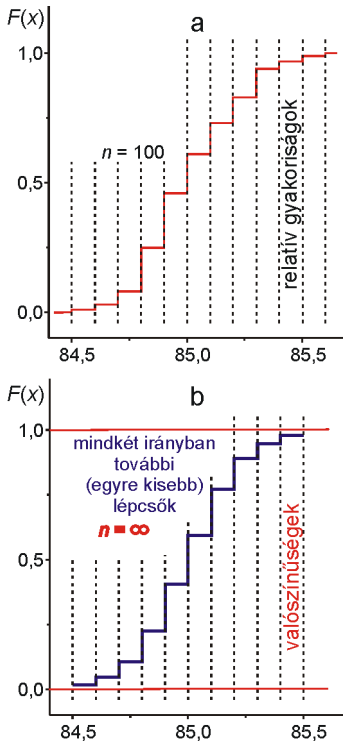


51. ábra
A 37. ábra D és B oszlopának grafikus összevetése. A folytonos esetben mind a 25 érték különböző, tehát a 39.ab ábra szerinti vetítés eredményeként mind a 25 érték látható is marad. A diszkrét esetben csak 8 különböző érték marad, tehát a 25 érték között vannak azonosak, így az adatrendszer jellemzéséhez a gyakoriságok ismerete is szükséges.

A 25-elemű diszkrét adatrendszer eloszlásfüggvényét a (21) összefüggés alapján ugyanúgy állíthatjuk elő, mint a folytonosét, természetesen a gyakoriságok figyelembevételével (52. ábra).



52. ábra
Az 51. ábrán bemutatott adatok eloszlásfüggvényei (D, B), valamint a B rész relatív gyakoriságai (piros oszlopok), illetve a „megfelelő” hisztogram (szürke terület) (G).



53. ábra
Az egy tizedesjegyre kerekített szórt fény intenzitás mint diszkrét valószínűségi változó, 100-elemű adatrendszerének (a) és a valószínűségi változónak (b) az eloszlásfüggvénye.

A folytonos és a diszkrét eset közötti különbség szembetűnő. Míg az ábra D részében mind a 25 lépcsőfok látható, addig a B-ben csak 8. A diszkrét esetben az eloszlás az adott értékekhez tartozó ***relatív gyakoriságokkal*** is ***egyértelműen*** jellemezhető, ugyanúgy, ahogy magával az eloszlásfüggvénnyel. Ez a folytonos esetben nem így volt (lásd a különböző hisztogramokat). Amennyiben a hisztogram készítéséhez szükséges osztályokat a kerekítési szabályoknak megfelelően választjuk meg, a két reprezentáció nagyfokú hasonlóságot mutat (52. ábra G rész).

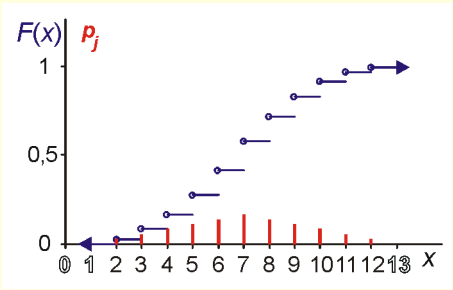
Hogyha a valószínűségi változó lehetséges értékeit tartalmazó halmazból egyre több elemet választunk ki véletlenszerűen, akkor a diszkrét esetben (a folytonos esettel ellentétben; lásd 40. ábra) az eloszlásfüggvény jellege nem fog megváltozni. Így **a diszkrét valószínűségi változó eloszlásfüggvénye ugyanúgy lépcsős marad, mint a neki megfelelő adatrendszeré** (53. ábra). Különbség csak abban fog mutatkozni, hogy a relatív gyakoriságok szerepét a valószínűségek veszik át.

Mintafeladat

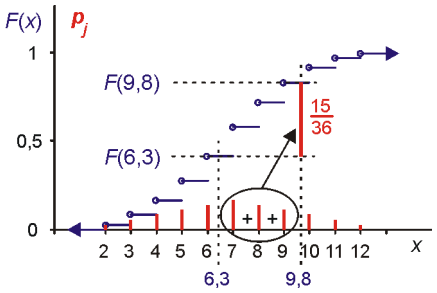
Válasszuk képzeletben a „kockadobás két ideális kockával” ***jelenséget***. Legyen a valószínűségi változó $\xi = i + k$, ahol i illetve k a dobás eredményei, azaz az 1, 2, 3, 4, 5, 6 számok valamelyike. Eszerint ξ lehetséges értékei x_j = 2-től 12-ig bármelyik természetes szám. Adjuk meg ξ valószínűségi változó jellemzését a lehetséges értékek valószínűségével illetve eloszlásfüggvénnyel.

Megoldás: Felhasználjuk a kockák függetlenségéről tanultakat és az ideális kockára vonatkozó korábbi ismereteinket (36. ábra). Ezek szerint a legkisebb értékű dobás az $x_j = 2$, ami csak akkor jön ki, ha mindkét kockán 1-et kapunk. Hasonlóképpen a legnagyobb értékű dobás az $x_j = 12$, ami csak akkor jön ki, ha mindkét kockán 6-ot kapunk. A (17) szorzási szabály alapján mindkét értékhez az 1/36 valószínűség rendelhető. Az $x_j = 3$ és az $x_j = 11$ kétféleképpen is kijöhet (1 + 2 = 2 + 1 = 3; 5 + 6 = 6 + 5 = 11), így a valószínűségek kiszámításához az összegzési szabályt is figyelembe kell vennünk: $P(\xi = 3) = P(\xi = 11) = 2/36$. Ezt a módszert folytatva a p_j-kre ($p_j = P(\xi = x_j)$) az alábbi táblázatban adtuk meg az eredményeket:

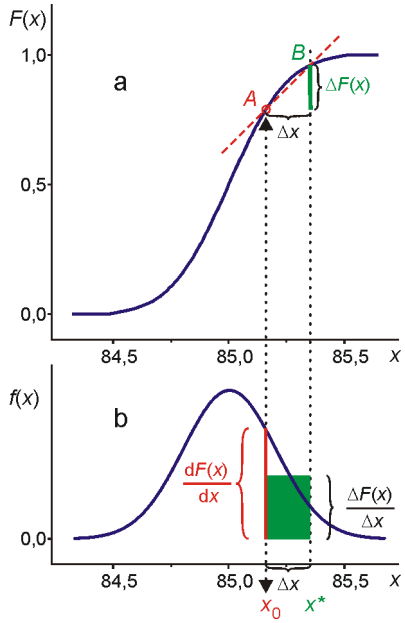
x_j	2	3	4	5	6	7	8	9	10	11	12
p_j	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



A valószínűségek ismeretében az eloszlásfüggvény a (22) definíciós egyenlet alapján állítható elő (ábrázolás a függőleges szakaszok nélkül):



54. ábra
A valószínűségek és az eloszlásfüggvény kapcsolata, ***diszkrét*** valószínűségi változó esetén a fenti mintafeladat alapján.



55. ábra
Az $f(x)$ valószínűségsűrűségek származtatása a ***folytonos*** valószínűségi változó eloszlásfüggvényéből, $F(x)$-ből. A (b) részben a függőleges tengelyt úgy kell beosztani, hogy a zöld ***terület*** éppen $\Delta F(x)$ legyen, azaz egyezzen meg az (a) részben a zöld oszlop ***magasságával***.

9.0. A valószínűségi változók további tulajdonságai

Az 52.G ábrán az oszlopok magassága a relatív gyakoriságokat mutatja, melyeket egymás után összeadva megkapjuk az eloszlásfüggvény megfelelő értékeit. Hasonló módon a fenti mintafeladat ábráján megfigyelhető, hogy az $F(x)$ eloszlásfüggvény lépcsőinek magassága éppen akkora, mint az adott x_j értékhez tartozó p_j valószínűség. Látható, hogy a kétféle reprezentáció teljes mértékben megfelel egymásnak. Egy lényeges különbség azért van közöttük: míg a p_j valószínűségek csak a megfelelő x_j értékekhez vannak hozzárendelve, az eloszlásfüggvény minden x-re értelmezve van. Így feltehető az a kérdés is, hogy mit jelent az $F(9,8) - F(6,3)$ különbség (54. ábra). Úgy is fogalmazhatunk, hogy mi annak a valószínűsége, hogy a két kockával való dobás eredménye a (6,3; 9,8) intervallumba esik. Mivel ennek a kikötésnek a 7-es, 8-as és a 9-es dobás bármelyike megfelel, de más dobás nem, ezért a fenti táblázatból kiolvasva a megfelelő értékeket, illetve az 54. ábra alapján a 6/36 + 5/36 + 4/36 = 15/36-ot kapjuk eredményül.

Diszkrét valószínűségi változó esetén a valószínűségek és az eloszlásfüggvény közötti kapcsolatot általánosan a következőképpen írhatjuk fel:

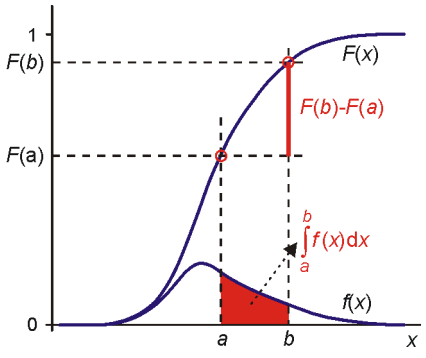
$$F(b) - F(a) = P(a < \xi < b) = \sum_{a<x_j<b} P(\xi = x_j), \qquad (24)$$

azaz egy adott intervallumban az eloszlásfüggvény függvényértékeinek ***különbsége*** adja meg a megfelelő valószínűséget, és fordítva, a valószínűségek ***összege*** adja meg az eloszlásfüggvény megfelelő függvényértékeinek különbségét.

Folytonos valószínűségi változó esetén a valószínűségeket a sűrűségfüggvénnyel is kifejezhetjük. Vizsgáljuk meg, hogy mi a kapcsolat egy folytonos valószínűségi változó $F(x)$ eloszlásfüggvénye (40.c ábra) és $f(x)$ sűrűségfüggvénye (48.c ábra) között. A sűrűségfüggvényt a ***relatív gyakoriságsűrűségek*** segítségével vezettük be, melyeket úgy határoztunk meg, hogy az adott intervallumon kapott relatív gyakoriságot elosztottuk az intervallum hosszával. Ennek mintájára az $F(x)$ eloszlásfüggvény ***valószínűségeiből*** is meghatározhatók a ***valószínűségsűrűségek*** (55. ábra).

Először válasszuk ki a kívánt intervallumot ($\Delta x = x^* - x_0$) majd az eloszlásfüggvény megfelelő függvényértékeinek különbségét ($\Delta F(x)$-et) osszuk el az intervallum hosszával (Δx-szel). A kapott eredmény az adott feltételek melletti valószínűségsűrűség, melynek számértéke nem más, mint az ábrán látható AB egyenes meredeksége.

Az így kapott valószínűségsűrűség akkor lesz független az intervallum hosszától (nevezetesen attól, hogy az x_0-hoz hogyan választjuk meg x^*-ot), ha Δx-szel 0-hoz tartunk. Ekkor az AB egyenes az $F(x)$ eloszlásfüggvény A pontbeli érintőjébe



56. ábra
A ***folytonos*** valószínűségi változó eloszlásfüggvénye és sűrűségfüggvénye közötti kapcsolat szemléltetése.

megy át. Mivel x_0-at bárhol megválaszthatjuk, ezért minden x-re igaz, hogy:

$$f(x)=\frac{\mathrm{d}F(x)}{\mathrm{d}x}, \tag{25}$$

azaz a sűrűségfüggvény az eloszlásfüggvény ***deriváltja***.

Folytonos valószínűségi változóra a (24) összefüggés alakja is módosul:

$$F(b)-F(a)=P(a<\xi<b)=\int_a^b f(x)\mathrm{d}x, \tag{26}$$

ahol az összegzés helyére a megfelelő ***integrál*** kerül (56. ábra).

Ezek szerint annak a valószínűsége, hogy az adott $F(x)$ eloszlásfüggvénnyel, vagy $f(x)$ sűrűségfüggvénnyel jellemzett folytonos valószínűségi változó lehetséges értékei közül egyet véletlenszerűen kiválasztva a kapott érték éppen az adott (a,b) intervallumba essen, megegyezik:

1. az $F(x)$ eloszlásfüggvény b illetve a helyen vett függvényértékeinek különbségével, valamint
2. az $f(x)$ sűrűségfüggvény (a,b) intervallumra eső görbe alatti területével.

Amennyiben $b=\infty$ és $a=-\infty$, az csak annyit jelent, hogy a kiválasztott érték ***valahol van***, ami egy biztos esemény, ezért ennek valószínűsége 1.

10.0. Adatrendszer és valószínűségi változó számszerű jellemzői

A sűrűségfüggvény bevezetésénél láthattuk, hogy annak ***kvalitatív*** jellegzetességeire, alakjára csak akkor következtethetünk, ha a valószínűségi változót elég sokszor megfigyeljük, azaz elég sok adatot gyűjtünk (50. ábra). Ezek a jellegzetességek például az egy púp vagy több púp, a szimmetria, vagy a púpok helye és szélessége. A továbbiakban olyan ***kvantitatív*** jellemzőket keresünk, amelyek ***kevés adat esetén is*** lehetőséget nyújtanak arra, hogy ezeket az ismérveket valahogy bemutassuk. (A továbbiakban az ***A***, illetve ***V*** jelölés arra utal, hogy a jellemző az ***adatokra***, vagy a valószínűségi ***változóra*** vonatkozik.)

c_1	= 5	d_1	= 1
c_2	= 2	d_2	= 4
c_3	= 2	d_3	= 6
c_4	= 8	d_4	= 6
c_5	= 1	d_5	= 2
c_6	= 9	d_6	= 1
c_7	= 5	d_7	= 2
c_8	= 5	d_8	= 9
c_9	= 3	d_9	= 6
c_{10}	= 2	d_{10}	= 5
c_{11}	= 2		

57. ábra
Számszerű adatok két adatrendszere (c, d).

A számolások egyszerűbb elvégezhetősége végett tekintsük az 57. ábrán feltüntetett két adatrendszert (c, d). Az első (c) $n=11$, a második (d) $n=10$ adatot tartalmaz. Először adjuk meg azokat a számszerű jellemzőket, amelyek azt mérik, hogy hol van az adatrendszer, illetve az eloszlás ***„közepe”***. A korábbi tanulmányokból már tudhatjuk, hogy középből többféle is van, ezért itt is több definíció adható.

1***A***. átlag, vagy számtani közép:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i. \tag{27}$$

A két konkrét esetben:

$$\bar{c}=\frac{5+2+2+8+1+9+5+5+3+2+2}{11}=4, \tag{28}$$

$$\bar{d}=\frac{1+4+6+6+2+1+2+9+6+5}{10}=4{,}2. \tag{29}$$

Nyilvánvalóan ugyanerre az eredményre jutunk, ha az adatokat először csoportosítjuk,

$c_5=\mathbf{1},\ c_2=c_3=c_{10}=c_{11}=\mathbf{2,2,2,2},\ c_9=\mathbf{3},\ c_1=c_7=c_8=\mathbf{5,5,5},\ c_4=\mathbf{8},\ c_6=\mathbf{9};$

$d_1=d_6=\mathbf{1,1},\ d_5=d_7=\mathbf{2,2},\ d_2=\mathbf{4},\ d_{10}=\mathbf{5},\ d_3=d_4=d_9=\mathbf{6,6,6},\ d_8=\mathbf{9},$

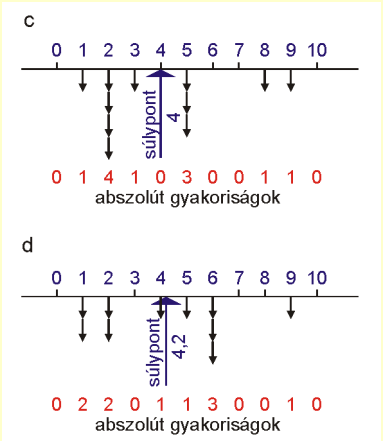
majd az összegeket az ***abszolút gyakoriságok*** figyelembevételével írjuk fel (58. megjegyzés):

$$\bar{c}=\frac{1\cdot\mathbf{1}+4\cdot\mathbf{2}+1\cdot\mathbf{3}+3\cdot\mathbf{5}+1\cdot\mathbf{8}+1\cdot\mathbf{9}}{11}=4, \tag{30}$$

$$\bar{d}=\frac{2\cdot\mathbf{1}+2\cdot\mathbf{2}+1\cdot\mathbf{4}+1\cdot\mathbf{5}+3\cdot\mathbf{6}+1\cdot\mathbf{9}}{10}=4{,}2. \tag{31}$$

58. megjegyzés
Az átlagnak a (30), (31) kifejezésekkel történő kiszámítása olyan, mint a fizikában a súlypont helyének meghatározása.

Akasszunk egy súlytalan rúdra, pontosabban az adatoknak megfelelő helyekre az abszolút gyakoriságokkal arányos súlyokat (vö. 29. ábra), és keressük meg, hogy hol van a súlypont.

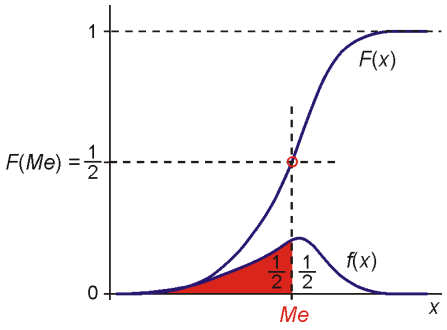


59. megjegyzés
Az előző mintafeladatban („kockadobás két ideális kockával”) szereplő $\xi = i + k$ valószínűségi változó várhatóértékét a (33) formula segítségével egyszerűen meghatározhatjuk.

A megfelelő x_j és p_j párok ismertek:

x_j	p_j	$x_j\,p_j$
2	1/36	2/36
3	2/36	6/36
4	3/36	12/36
5	4/36	20/36
6	5/36	30/36
7	6/36	42/36
8	5/36	40/36
9	4/36	36/36
10	3/36	30/36
11	2/36	22/36
12	1/36	12/36
	Σ	252/36 = 7

Így ξ várhatóértéke, $E(\xi) = 7$.



60. ábra
A medián szemléletes jelentése (folytonos valószínűségi változóra bemutatva): az az **osztóérték**, amely a sűrűségfüggvényt két egyforma területű részre osztja.

Általánosan:

$$\bar{x} = \frac{\sum_{j=1}^{m} k_j x_j}{\sum_{j=1}^{m} k_j} = \frac{1}{n}\sum_{j=1}^{m} k_j x_j = \sum_{j=1}^{m} \frac{k_j}{n} x_j\,, \qquad (32)$$

ahol a k_j együtthatók az abszolút gyakoriságokat jelölik és $\sum_{j=1}^{m} k_j = n$.

A (32) összefüggésben azt is megfigyelhetjük, hogy a különböző x_j-k együtthatói végső soron a k_j/n ***relatív gyakoriságok*** lesznek.

Ennek mintájára és felhasználva azt, hogy a nagy számok törvénye (lásd 7. oldal) összekapcsolja a relatív gyakoriságokat a valószínűségekkel, megadhatjuk egy ξ valószínűségi változóra, illetve annak eloszlására vonatkozó hasonló jellemzőt is.

1***V***. A ξ várhatóértékét ***diszkrét*** esetben az

$$E(\xi) = \sum_{j=1}^{m} p_j x_j \qquad (33)$$

formula adja meg, ahol az x_j-k jelölik ξ lehetséges értékeit, és a p_j-k az ezekhez tartozó valószínűségeket (59. megjegyzés).

A ***folytonos*** esetben

$$E(\xi) = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x\,, \qquad (34)$$

ahol x jelöli ξ lehetséges értékeit, és $f(x)$ a ξ sűrűségfüggvényét. (Formálisan az $f(x)\mathrm{d}x$ szorzat felel meg a valószínűségeknek, és az egyszerű összegzés integrállá módosul.)

2***A***.***V***. Egy másik lehetséges „közép” a medián (itt *Me*-vel jelöljük), amelyet – mind az adatrendszerre, mind a valószínűségi változóra vonatkozóan – a legegyszerűbben az eloszlásfüggvény segítségével lehet megadni (60. ábra). A medián az az érték (vagy értékek), amelyre:

$$F(Me) = \frac{1}{2}\,. \qquad (35)$$

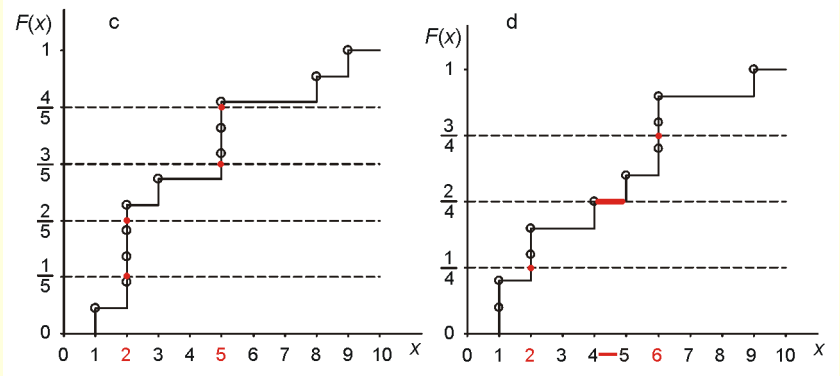
3***A***.***V***. A további osztóértékeket általánosan kvantiliseknek nevezzük. Ezek lehetnek pl. harmadoló, negyedelő, ötödölő, tizedelő vagy akár századoló értékek, amelyek név szerint rendre a tercilisek, kvartilisek, kvintilisek, decilisek, percentilisek. Ezek szerint pl. az alsó ($Q1$), a középső ($Q2 = Me$), illetve a felső ($Q3$) kvartilis definíciója a következő:

$$F(Q1) = \frac{1}{4}\,, \qquad F(Q2) = F(Me) = \frac{1}{2}\,, \qquad F(Q3) = \frac{3}{4}\,. \qquad (36)$$

Mintafeladat

Határozzuk meg az 57. ábrán bemutatott c adatrendszer kvintiliseit (K) és a d adatrendszer kvartiliseit (Q)!

Megoldás: Készítsük el a két adatrendszer eloszlásfüggvényét (lépcsőfüggvény ábrázolásban), majd olvassuk le a megfelelő értékeket.



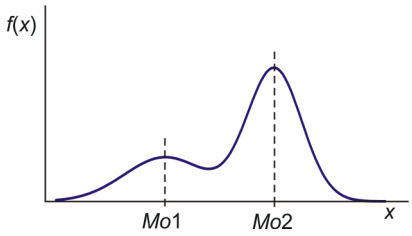
c eset:

$K1 = K2 = 2$; $K3 = K4 = 5$.

d eset:

$Q1 = 2$; $4 < Q2 < 5$; $Q3 = 6$.

Érdekesség, hogy a c esetben a két-két kvintilis megegyezik, továbbá, hogy a d esetben a középső kvartilis egy nyílt intervallum bármelyik eleme lehet, amely éppen két egyenlő részre osztja az adatokat, tehát 5-5 adatot választ el.



61. ábra
Egy kétpúpú sűrűségfüggvény móduszai.

4***A***. Az 50. ábrán, ahol a sűrűségfüggvények jellegzetes típusait mutatjuk be, megkülönböztetünk egypúpú és több púpú függvényeket. Az egypúpúakat tekintve azt mondhatjuk, hogy a púp helyének a környékén a valószínűségi változó megfigyelt értékei (azaz a megfelelő adatok) igen gyakran fordulnak elő. Ezt alapul véve egy újabb „közép” megadására nyílik lehetőség.

Ha az adatrendszerben vannak azonos értékű elemek, akkor azt, amelyikből a ***legtöbb*** van (leggyakoribb érték) az adatrendszer móduszának nevezzük (itt *Mo*-val jelöljük).

4***V***. A módusz ***diszkrét*** esetben a valószínűségi változó lehetséges értékei közül a legvalószínűbb:

$$P(Mo) = \max . \tag{37}$$

A ***folytonos*** esetben a sűrűségfüggvény lokális maximum helyével adható meg:

$$f(Mo) = \max , \tag{38}$$

ami azt jelenti, hogy ennek a helynek (*Mo*) a környékén a legvalószínűbb a valószínűségi változó megfigyelt értékeinek az előfordulása. Amennyiben lokális maximumból több is van, akkor multimodális (több púpú) sűrűségfüggvényről beszélünk (61. ábra).

A különböző „közepek” után nézzünk egy más fajta jellemzőt, ami azt adja meg, hogy ***mennyire szóródnak*** az adatok, illetve, hogy ***milyen széles*** az eloszlás.

5***A***. Az adatrendszer legnagyobb és legkisebb elemének az eltérése az adatrendszer terjedelme:

$$T = x_{\max} - x_{\min} , \tag{39}$$

6***A.V***. További, más fajta „terjedelem” is megadható az ***osztóértékek segítségével***. Ilyen például az interkvartilis terjedelem, ami a felső és az alsó kvartilis különbsége:

$$IQT = Q3 - Q1 . \tag{40}$$

7***A***. Az adatrendszer átlagától vett ***négyzetes eltérések átlagát*** az adatrendszer szórásnégyzetének (vagy varianciájának) nevezzük (62a. és 63a. megjegyzések):

$$s_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 ; \tag{41}$$

ennek négyzetgyöke pedig az adatrendszer szórása:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} . \tag{42}$$

7***V***. A (41) összefüggés mintájára, felhasználva a várhatóérték (33, 34) definícióit, a ξ valószínűségi változóra vonatkozóan is megadhatunk egy hasonló jellem-

62. megjegyzés
a) Mivel az adatok átlagos eltérése az átlagtól mindig nulla:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}n\bar{x} = 0 ,$$

ez a mennyiség nem alkalmas az adatok szóródásának jellemzésére;

b) hasonlóképpen a várhatóértéktől való eltérések várhatóértéke:

$$E(\xi - E(\xi)) = 0 .$$

Mintafeladat

Mutassuk meg az átlag minimum tulajdonságát, nevezetesen azt, hogy amennyiben a szórásnégyzet (41) képletében az $\bar{x}$ helyére bármilyen más értéket írunk, akkor a kifejezés értéke növekszik.

Megoldás: Tekintsük x^*-ot változónak és írjuk $\bar{x}$ helyére, majd keressük x^*-nak azt az értékét, amely a kifejezést minimalizálja.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - x^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i{}^2 - 2x_i x^* + x^{*2}) = \frac{1}{n}\sum_{i=1}^{n}x_i{}^2 \underbrace{\left(-\frac{1}{n}\sum_{i=1}^{n}2x_i\right)}_{B} x^* + \underbrace{\left(\frac{1}{n}n\right)}_{A}x^{*2}$$

Felhasználva az $Ax^{*2} + Bx^* + C = 0$ alakú másodfokú egyenletek megoldó képletét: $x^*_{1,2} = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$;

illetve azt, hogy a gyökök a szélsőértékre (esetünkben, mivel $A > 0$, ezért minimumra) nézve szimmetrikusak, ezért: $x^*_{\min} = \frac{-B}{2A}$. A bekarikázott együtthatókat leolvasva: $A = 1$, B-re pedig éppen a az átlag –2-szerese adódik, így $x^*_{\min} = \bar{x}$.

63. megjegyzés
a) Az adatrendszer szórásnégyzetének kiszámításához a (41) összefüggésen kívül egy egyszerűbben használhatót is megadhatunk, ha felhasználjuk az előbbi mintafeladat eredményeit.

A végső kifejezésben szereplő első tag nem más, mint az adatok négyzetének átlaga:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \overline{x^2} \ .$$

A másik két tag pedig $x^* = \overline{x}$ helyettesítéssel, az összevonás elvégzése után nem más, mint az átlag négyzetének –1-szerese, így:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2 \ .$$

b) Általánosan a

$$\frac{1}{n}\sum_{i=1}^{n} x_i^k$$

kifejezést az adatrendszer *k*-adik momentumának, a

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$$

kifejezést pedig az adatrendszer *k*-adik centrális momentumának nevezik.

c) Ezeket a mennyiségeket – a várhatóérték definíciójának felhasználásával – valószínűségi változókra is meg lehet adni. Ezek szerint a várhatóérték az első momentum, a variancia a második centrális momentum, ami a fentiek alapján a következőképpen is felírható:

$$Var(\xi) = E(\xi^2) - E^2(\xi) \ .$$

A harmadik momentummal az eloszlás ferdesége, a negyedik momentummal az eloszlás csúcsossága (lapultsága) jellemezhető.

65. megjegyzés
a) Az $\binom{n}{x}$ (*n* alatt az *x*) szimbólum egy szám,

amely azt adja meg, hogy hányféleképpen lehet *n* különböző elemből *x* különbözőt kiválasztani úgy, hogy a sorrendjükre nem vagyunk tekintettel; értékeit *n* = 6-ig bezárólag az alábbi táblázat mutatja:

		x						
		0	1	2	3	4	5	6
n	1	1	1					
	2	1	2	1				
	3	1	3	3	1			
	4	1	4	6	4	1		
	5	1	5	10	10	5	1	
	6	1	6	15	20	15	6	1

b) A faktoriális definíciójának ismeretében:

$$\binom{n}{x} = \frac{n!}{(n-x)!\,x!} \ ,$$

ahol pl. 5 faktoriális = 5! = 1·2·3·4·5 = 120.

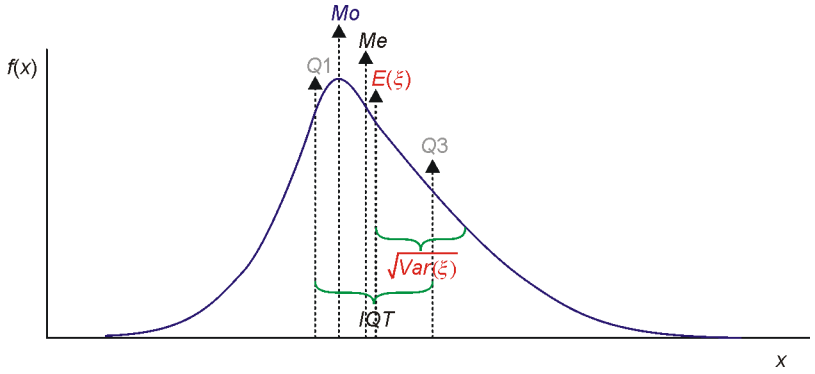
zőt, a varianciát (vagy szórásnégyzetet) (62b. 63c. megjegyzések):

$$Var(\xi) = E\left[(\xi - E(\xi))^2\right] . \qquad (43)$$

Még további jellemzők is megadhatók, olyanok, amelyek az 50. ábrán lévő sűrűségfüggvényeken is megfigyelhető alaki sajátságokat közvetlenül tükrözik: ***ferdeség***, ***csúcsosság*** (***lapultság***); de ezekre itt nem térünk ki.

10.1. A számszerű jellemzők áttekintése és összehasonlítása

Mindezeket legszemléletesebben a sűrűségfüggvény segítségével tudjuk bemutatni. A 64. ábrán feltüntettük egy valószínűségi változó összes eddig bevezetett számszerű jellemzőjét. (Mivel a terjedelem csak adatrendszerre van definiálva, ezért nem szerepel az ábrán.)



64. ábra
Egy valószínűségi változó legfontosabb számszerű jellemzői a sűrűségfüggvényén bemutatva.

Bár szimmetriára, illetve alaki sajátságokra vonatkozó közvetlen jellemzőket nem adtunk meg, a meglévők felhasználásával ilyenekre is tudunk következtetni. Egypúpú szimmetrikus sűrűségfüggvény esetén például:

$$E(\xi) = Me = Mo \ . \qquad (44)$$

Jobbra elnyújtott esetben (lásd 64. ábra) $Mo < Me < E(\xi)$, balra elnyújtott esetben a fordított egyenlőtlenség igaz. (A lapultságra, csúcsosságra a különböző interkvantilis terjedelmek összehasonlításából következtethetünk.)

11.0. Nevezetes eloszlások, matematikai modellek

Diszkrét eloszlások

Az egyenletes eloszlást az ***ideális kocka*** esetében már megbeszéltük (36. ábra), amit itt csak a számszerű jellemzőkkel egészítünk ki. A (33) és (43) összefüggések felhasználásával:

$$E(\xi) = \frac{1}{6}(1+2+3+4+5+6) = 3{,}5 \ ; \qquad Var(\xi) \approx 2{,}92 \ ; \qquad (45)$$

és a 35. illetve 36. ábrákról leolvasva:

$$3 < Me < 4 \ ; \qquad Mo = x_i \ \text{ minden } i\text{-re.} \qquad (46)$$

Binomiális eloszlás (Bernoulli-eloszlás, vagy ismételt alternatívák eloszlása). Maradjunk az ***ideális kocka*** eseténél, de most azt a kérdést tesszük föl, hogy mi a valószínűsége annak, hogy mondjuk 5 megismételt (egymástól független) dobás során egyszer sem jön ki 6-os. Általánosabban fogalmazva: annak a valószínűségét keressük, hogy egy jelenség *n*-szeri megismétlődésekor (*n*-szer megfigyelve) éppen *x*-szer következik be a *p* valószínűségű *A* esemény. Az (*n*,*p*) paraméterű eloszlást az alábbi formula adja meg:

$$P(\xi = x) = \binom{n}{x} p^x (1-p)^{n-x} \ , \qquad (47)$$

ahol a jobb oldal első tényezője a megfelelő binomiális együttható (65. megjegyzés).

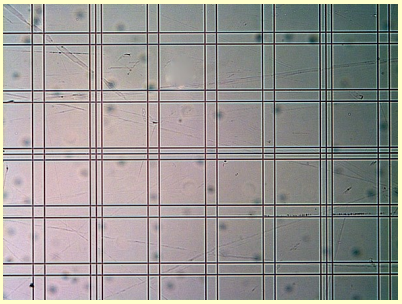
Jakob Bernoulli (1654-1705), svájci matematikus.

67. megjegyzés
a) A várhatóérték (48) formulája könnyen érthetővé válik, ha arra gondolunk, hogy kb. 6 dobásonként várunk egy hatost.

b) Azt is megfigyelhetjük, hogy kis valószínűségű események esetén $(1-p) \approx 1$, ami azt jelenti, hogy a variancia (49) kifejezése gyakorlatilag megegyezik (48)-cal, azaz a variancia majdnem ugyanakkora, mint a várhatóérték.

Siméon Denis Poisson (1781-1840), francia matematikus és fizikus; nevének érdekessége, hogy a poisson halat jelent, melyek eloszlása például egy halastóban éppen Poisson-eloszlással írható le.

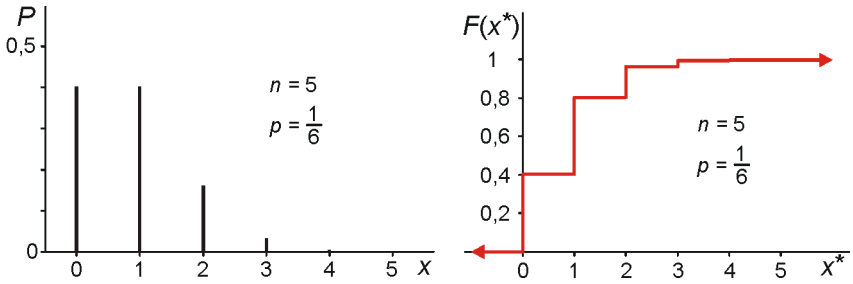
69. megjegyzés
Számlálókamrás eljárások
A számlálókamrák különlegesen kiképzett mikroszkópi tárgylemezek, amelyeken ismert területű beosztás található. A fedőlemez feltételével meghatározott magasságú réteg keletkezik és emiatt a beosztásban elhelyezkedő folyadék térfogata jól meghatározott.
Ez a módszer használható például vörösvértest koncentráció meghatározásra is (Bürker-kamra).



Ha $n = 5$, $p = 1/6$, akkor $P(\xi = 0) = 1 \cdot (1/6)^0 \cdot (5/6)^5 = (3125/7776) \approx 0{,}40188$. A többi x-re $P(\xi = x)$ az alábbi táblázatban foglaltak szerint változik:

x	0	1	2	3	4	5
P	0,40188	0,40188	0,16075	0,03215	0,00321	0,00013

Szemléltetésül vagy a valószínűségek „pálcika”, vagy az eloszlásfüggvény „lépcsős” reprezentációja használható.



66. ábra
Az (5, 1/6) paraméterekkel jellemzett binomiális eloszlás szemléltetése kétféle módon (míg x diszkrét, x^* folytonos változó).

Az eloszlás számszerű jellemzői (67. megjegyzés):

$$E(\xi) = np; \quad \text{esetünkben} = \frac{5}{6} \approx 0{,}833; \tag{48}$$

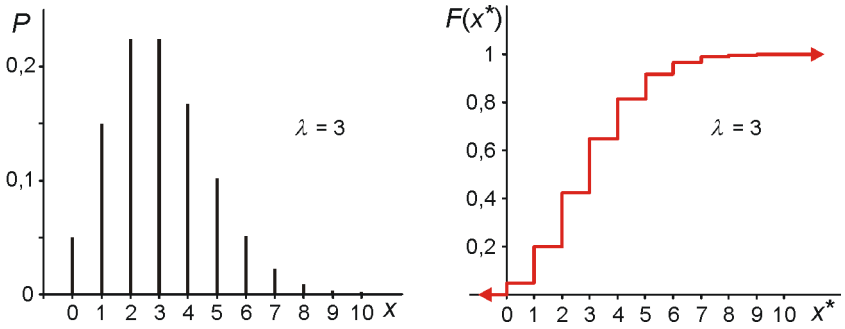
$$Var(\xi) = np(1-p); \quad \text{esetünkben} = \frac{5}{6} \cdot \frac{5}{6} \approx 0{,}694; \tag{49}$$

a medián és a móduszok az ábrákról leolvashatók:

$$Me = 1; \qquad Mo1 = 0; \qquad Mo2 = 1. \tag{50}$$

Bár a Poisson-eloszlás egy független, önálló matematikai modell, legegyszerűbben úgy juthatunk hozzá, ha megkeressük a binomiális eloszlásnak azt a határesetét, amikor $n \to \infty$, $p \to 0$, miközben $np = \lambda$ véges érték marad. Ennek eredménye a λ paraméterű Poisson-eloszlás:

$$P(\xi = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad (x = 0, 1, 2, \ldots) \text{ (65b. megjegyzés).} \tag{51}$$



68. ábra
A $\lambda = 3$ paraméterű Poisson-eloszlás szemléltetése kétféle módon (x diszkrét, x^* folytonos változó).

Itt az egyetlen paraméter λ a várhatóértékkel és a varianciával is megegyezik (vö. 67b. megjegyzés):

$$E(\xi) = Var(\xi) = \lambda; \quad \text{esetünkben} = 3; \tag{52}$$

a medián és a móduszok itt is leolvashatók az ábrákról:

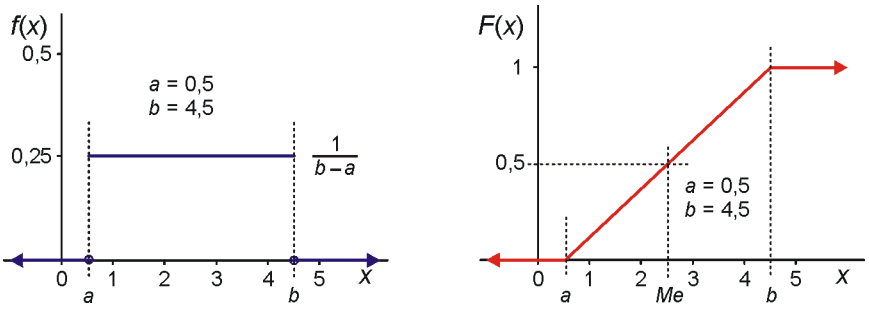
$$Me = 3; \qquad Mo1 = 2; \qquad Mo2 = 3. \tag{53}$$

Példaként felhozhatjuk a radioaktív preparátumban adott idő alatt elbomló atomok számát, mint valószínűségi változót, ami jól modellezhető Poisson-eloszlással. Hasonlóképpen jó példának számít az adott térfogatban lévő részecskék száma, mint valószínűségi változó (69. megjegyzés).

Folytonos eloszlások

Az egyenletes eloszlást a diszkrét esetben már megbeszéltük (15.oldal), illetve kiegészítettük (25. oldal). Nézzük meg milyen különbségek származnak abból, ha a valószínűségi változó folytonos. Milyen az (a,b) paraméterű egyenletes eloszlás?

Szemléltetésül a sűrűségfüggvényt és az eloszlás függvényt használjuk.



70. ábra
Az (a = 0,5; b = 4,5) paraméterű egyenletes eloszlás szemléltetése kétféle módon.

$f(x) = 0$ az (a,b) intervallumon kívül, azon belül pedig nullától különböző állandó érték, éppen akkora, hogy a görbe alatti terület 1 legyen. Ennek megfelelően $F(x)$ lineáris függvény az (a,b) intervallumban, továbbá $F(x) = 0$, ha $x \le a$ *és* $F(x) = 1$, ha $x > b$.

Az eloszlás számszerű jellemzői:

$$E(\xi) = \frac{a+b}{2}; \quad \text{esetünkben} = 2{,}5\,; \tag{54}$$

$$Var(\xi) = \frac{(b-a)^2}{12}; \quad \text{esetünkben} = \frac{4}{3} \approx 1{,}33\,; \tag{55}$$

a medián és a móduszok az ábrákról leolvashatók:

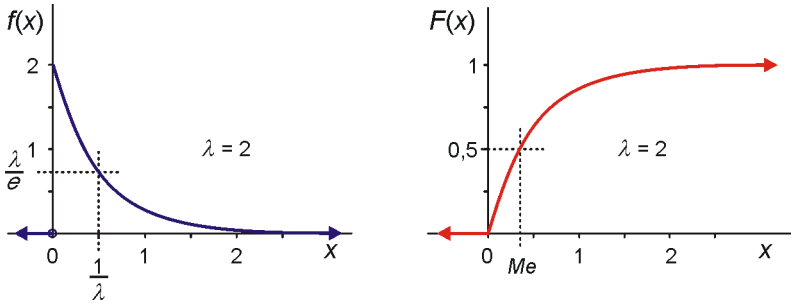
$$Me = 2{,}5\,; \qquad 0{,}5 < Mo < 4{,}5\,. \tag{56}$$

Egy üres teremben a sűrűség vagy a hőmérséklet eloszlása (az ablakoktól és a fűtőtestektől „távol") jól modellezhető folytonos egyenletes eloszlással.

A (diszkrét) Poisson-eloszláshoz hasonlóan az exponenciális eloszlásnak is csak egy paramétere van, amit ráadásul szintén λ-val szokás jelölni, de ennek jelentése – mint látni fogjuk – egészen más. Az exponenciális eloszlás sűrűségfüggvénye, illetve eloszlásfüggvénye:

$$f(x) = \lambda e^{-\lambda x}\,; \qquad F(x) = 1 - e^{-\lambda x}\,. \tag{57}$$

Ezeket szemléltetjük az alábbi ábrán.



71. ábra
A $\lambda = 2$ paraméterű exponenciális eloszlás szemléltetése kétféle módon.

Az eloszlás számszerű jellemzői (72. megjegyzés):

$$E(\xi) = \sqrt{Var(\xi)} = \frac{1}{\lambda}; \quad \text{esetünkben} = 0{,}5\,; \tag{58}$$

$$Me = \frac{\ln 2}{\lambda}; \quad \text{esetünkben} \approx 0{,}35\,; \qquad Mo = 0\,. \tag{59}$$

72. megjegyzés
Láthatjuk, hogy az exponenciális eloszlás sűrűségfüggvényének (57/1) kezdeti értéke éppen λ, hiszen $x = 0$ esetben

$$f(0) = \lambda e^{-\lambda \cdot 0} = \lambda\,,$$

ami egyben a maximális érték is, tehát $Mo = 0$ (vö. 71/1. ábra).

Amennyiben x helyére $1/\lambda$-át írunk, akkor a kitevő éppen -1-gyel egyenlő, tehát $f(x) = \lambda/e$.

Me meghatározásához a szokásos $F(Me) = 0{,}5$ egyenlet megoldása vezet. Az eloszlásfüggvény (57/2) kifejezését átrendezve kapjuk az

$$e^{-\lambda \cdot Me} = 0{,}5\,, \text{ illetve az } e^{\lambda \cdot Me} = 2$$

egyenletet, majd a két oldal természetes alapú logaritmusát véve Me egyszerű osztással megkapható.

73. megjegyzés
A gerjesztett állapotban lévő atomok számának az időtől való függését kétféle módon is felírhatjuk:

$$N = N_0 e^{-\frac{t}{\tau}} = N_0 2^{-\frac{t}{T}} ,$$

ahol τ az átlagos élettartamot, T pedig a felezési időt jelöli. A kettő közötti kapcsolat: $T = \ln 2\,\tau$, vagy $1/\tau = \lambda$ helyettesítéssel $T = \ln 2/\lambda$ (vö. (59/1)). Ezután, ha kifejezzük az alapállapotba visszatérő elektronok arányát $1-(N/N_0)$-t, akkor az exponenciális eloszlásfüggvényhez jutunk.

Mindezek alapján azt mondhatjuk, hogy a $t = \tau$ a várhatóértéknek, a $t = T$ a mediánnak, a $t = 0$ pedig a módusznak felel meg.:

Carl Friedrich Gauss (1777-1855), német matematikus és fizikus.

Radioaktív bomlás során az egyes atomok élettartama, vagy gerjesztés után a gerjesztett állapotot elhagyó elektronok várakozási ideje jól modellezhető exponenciális eloszlással (73. megjegyzés).

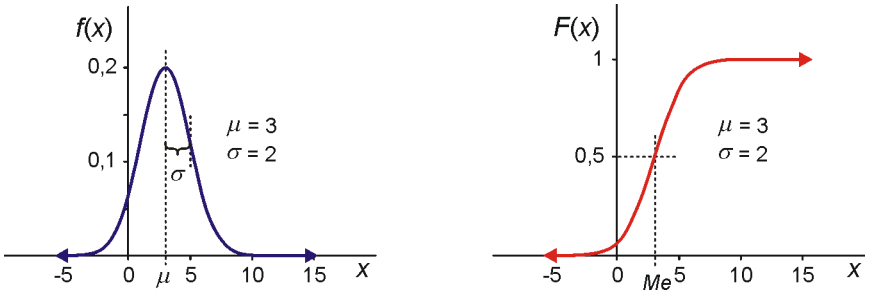
A leggyakrabban használt eloszlás a normális eloszlás (vagy Gauss-eloszlás). Ezt az eloszlást két paraméter jellemzi (μ és σ), sűrűségfüggvénye pedig a következő:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} . \qquad (60)$$

Ez a matematikai kifejezés az eddigieknél bonyolultabbnak tűnik, de valójában nem más, mint az

$$f(x) = e^{-x^2} \qquad (61)$$

függvény, néhány állandóval kiegészítve. (Az eloszlásfüggvény már lényegesen bonyolultabb – a (26) összefüggés szerint a (60) sűrűségfüggvény integrálja – így azt itt nem is írtuk fel, de grafikonon bemutatjuk.) A sűrűségfüggvényt és az eloszlásfüggvényt a 74. ábra szemlélteti.



74. ábra
A $\mu = 3$, $\sigma = 2$ paraméterű normális eloszlás szemléltetése kétféle módon.

Az eloszlás számszerű jellemzői:

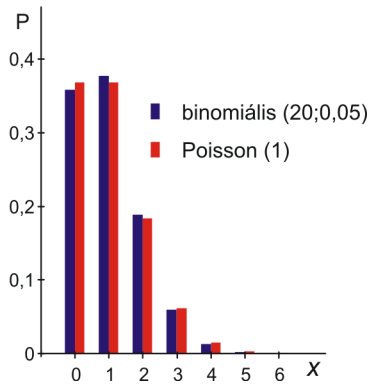
$$E(\xi) = \mu = Me = Mo; \quad \text{esetünkben} = 3 ; \qquad (62)$$

$$\sqrt{Var(\xi)} = \sigma ; \quad \text{esetünkben} = 2 ; \qquad (63)$$

Igen sok esetben használhatunk modellként normális eloszlást, aminek okára még visszatérünk (vö. 11.1. rész és 77. megjegyzés). Csak egy példát említve, egy sejtkultúrában a sejtek karakterisztikus lineáris mérete (például a leghosszabb átmérő), mint valószínűségi változó ilyen.

11.1. A normális eloszlás kitüntetett szerepe

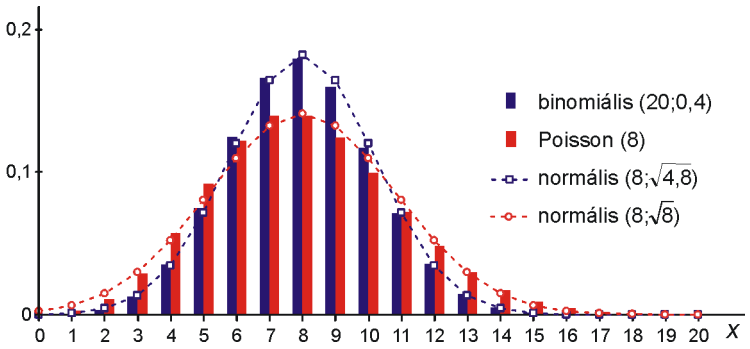
Mivel a Poisson-eloszlást a binomiális eloszlás határeseteként vezettük be, nem meglepő, ha kis p értékek esetén, $np = \lambda$ helyettesítéssel ez a két eloszlás nagyon hasonló. Anélkül, hogy a probléma matematika hátterének részleteibe betekintenénk, csupán a 75. ábra alapján megfigyelhetjük, hogy például $p = 0{,}05$ esetén a hasonlóság szembetűnő.



75. ábra
Az $n = 20$, $p = 0{,}05$ paraméterű binomiális és az $np = \lambda = 1$ paraméterű Poisson-eloszlás hasonlóságának szemléltetése.

Ha azonban p növekszik, a binomiális eloszlás csak $p = 0{,}5$-ig szélesedik, utána keskenyedik, ugyanis $Var(\xi) = np(1–p)$ akkor maximális, ha $p = (1–p)$, (mivel azonos kerületű téglalapok közül a négyzet a maximális területű). Ezzel szemben λ növekedtével a Poisson-eloszlás egyre szélesebb lesz, hiszen $Var(\xi) = \lambda$. Ennek az a következménye, hogy nagyobb p értékek esetén a hasonlóság is megszűnik. A 76. ábrán bemutatott binomiális eloszlás abban különbözik a 75. ábrán bemutatottól, hogy p-t 0,05-ról 0,4-re növeltük. Így $Var(\xi)_{\text{binomiális}} = 4{,}8$, míg az ennek „megfelelő” $Var(\xi)_{\text{Poisson}} = 8$, ami az ábrán is megfigyelhető.

Belátható, illetve esetünkben megmutatható, hogy mindkét eloszlás **jól közelíthető** egy-egy **normális eloszlással**, melyek várhatóértéke azonos ($\mu = np = \lambda$), de szélességük, pontosabban a szórásuk különböző ($\sigma_{\text{binomiális}} = \sqrt{np(1-p)}$, $\sigma_{\text{Poisson}} = \sqrt{\lambda}$) (76. ábra). Így a folytonos normális eloszlás sűrűségfüggvényéből diszkrét eloszlások valószínűségeit becsülhetjük meg, mivel a diszkrét esetek egységnyi lépésközei miatt a terület és a magasság számértéke megegyezik.



76. ábra
Az $n = 20$, $p = 0{,}4$ paraméterű binomiális és az $np = \lambda = 8$ paraméterű Poisson-eloszlás, valamint a nekik megfelelő normális eloszlások szemléltetése.

77. megjegyzés
Ezzel magyarázható többek között az a tapasztalat is, hogy a természetben előforduló változók jelentős része normális eloszlással jól modellezhető.

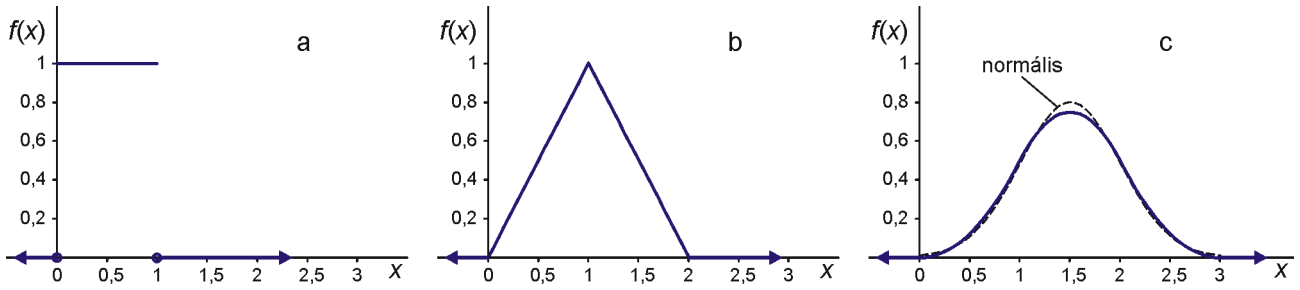
Vegyük például a testmagasságot, mint valószínűségi változót, amelynek értékét rengeteg véletlenszerű hatás – azaz független valószínűségi változó – összegeződése határozza meg (például a táplálkozás sok összetevője, a genetikai tényezők stb.), akkor a centrális határeloszlás tétel értelmében azt mondhatjuk, hogy ennek a valószínűségi változónak jó közelítéssel normális eloszlásúnak kell lennie, ami tapasztalatainkkal is egybevág.

A normális eloszlás kiemelkedő jelentőségére a valószínűségszámítás egyik nevezetes tétele, a centrális határeloszlás tétel mutat rá. Matematikailag, de nem a szokásos precizitással ez a következőt jelenti: **ha sok független valószínűségi változót összegzünk, akkor elég általános feltételek teljesülése esetén az összeg normális eloszlású valószínűségi változó lesz** (77. megjegyzés). Bár ezt a tételt sem kívánjuk bizonyítani, azt azért megmutatjuk, hogy a tételben szereplő összegző hatás eredménye viszonylag milyen hamar érvényre jut (78. ábra).

Elevenítsük fel a „kockadobás két ideális kockával” feladatot (21. oldal), amelyből az derült ki, hogy míg az egy kockával való dobás eredménye mint valószínűségi változó egyenletes eloszlású (36. ábra), addig a két kocka esetén kapott összeg 2–7-ig egyenletesen növekvő valószínűségű, majd 7–12-ig szimmetrikusan, egyenletesen csökkenő. Így az egyes értékekhez tartozó valószínűségeket ábrázolva, azok háromszögszerűen változnak.

Hasonló eredményt kapunk a folytonos esetben is. Tekintsük a (0,1) intervallumban folytonos egyenletes eloszlású valószínűségi változót, amelynek sűrűségfüggvénye a 78a. ábrán látható. Ha két ilyen változót összeadunk, és ezek függetlenek (lásd: kockadobás két kockával), akkor az összeg, mint új folytonos valószínűségi változó sűrűségfüggvénye – a diszkrét esethez hasonlóan – egyenlőszárú háromszöget formáz (78b. ábra).

Három ilyen változó összegének sűrűségfüggvénye már olyan (parabolaívekből összerakott) „haranggörbét” mutat, amely szemre nagyon hasonlít a normális eloszlás sűrűségfüggvényéhez (78c. ábra, Mintafeladat).

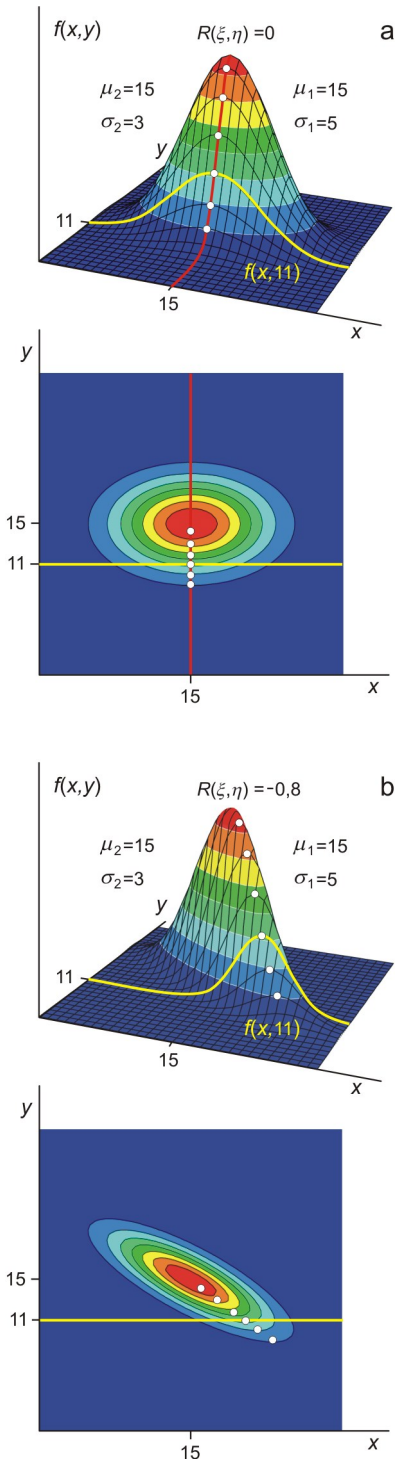


78. ábra
A centrális határeloszlás tételben szereplő összegző hatás érvényre jutásának szemléltetése. a: a (0,1) intervallumban folytonos egyenletes eloszlású valószínűségi változó sűrűségfüggvénye; b, c: 2 illetve 3 ilyen független változó összegének sűrűségfüggvénye. A c ábrán a normális eloszlás sűrűségfüggvényét is feltüntettük szaggatott vonallal.

Mintafeladat

Mutassuk meg, hogy a „kockadobás három ideális kockával” probléma valószínűségei haranggörbeszerűen változnak!

Megoldás: Az könnyen belátható, hogy a „kockadobás két ideális kockával” problémához hasonlóan ez az eloszlás is szimmetrikus lesz, hiszen minden dobáseredmény-hármasnak megvan a párja: például (111;666) vagy (112;665) stb. 1+1+1=3-at csak 1-féleképpen lehet dobni, 4-et 3-féleképpen (4=2+1+1 és a 2-es mind a 3 kockán kijöhet), 5-öt 6-féleképpen (5=1+2+2=3+1+1 és az 1-es, illetve a 3-as is mind a 3 kockán kijöhet). Azt már nehezebb belátni, hogy 9-et 25-féleképpen, 10-et pedig 27-féleképpen lehet dobni, de a különbség oka az, hogy míg a 9=3+3+3 dobás csak 1-féleképpen, addig a 10=3+3+4 dobás 3-féleképpen jöhet ki. Láthatjuk, hogy ebben a sorozatban 1, 3, 6,…, 25, 27 a növekmények (3–1=2; 6–3=3; …, 27–25=2) – a haranggörbe alakjának (felszálló ágának) megfelelően – növekvő majd csökkenő tendenciát mutatnak.



79. ábra
Két folytonos valószínűségi változó (ξ és η) együttes eloszlásának szemléltetése az együttes sűrűségfüggvénnyel (térben, illetve az xy síkra eső merőleges vetülettel ábrázolva).

a) ξ és η független változók, mert például a különböző $\boldsymbol{y}$ helyeken az x tengellyel párhuzamos metszetgörbék paraméterei megegyeznek a vetületeloszlás paramétereivel (μ_1 és σ_1) és csak a görbék alatti terület változik. Megfigyelhető, hogy a piros görbe mentén $x_{max} = \mu_1 = 15$, állandó érték.

b) ξ és η nem független változók, mert itt például az előbbi x_{max} helyek eltolódnak, ha $\boldsymbol{y}$ változik.

12.0. Két valószínűségi változó együttes eloszlása, feltételes eloszlása, függetlensége

Vannak esetek, amikor két (vagy több) adat együttesen szükséges valaminek a jellemzéséhez. Például az, hogy valaki kövér vagy sem, nem dönthető el csupán a testsúlya alapján, mert tudjuk, hogy a testmagasság is számít. Ilyenkor a megfelelő valószínűségi változókat is együttesen kell vizsgálnunk. Nem elegendő például külön-külön ismernünk az eloszlásukat, mert így nem kapunk felvilágosítást a közöttük lévő esetleges kapcsolatról (mondjuk például arról, hogy a nehezebb emberek általában magasabbak is). (A következőkben az egyszerűbb szemléltetés kedvéért normális eloszlású változókat használunk, de a kapott eredmények más eloszlású változókra is igazak).

Legyen ξ és η (éta) akármilyen valószínűségi változó, a (22) összefüggés mintájára értelmezhető az együttes eloszlásfüggvényük a következőképpen:

$$F(x,y) = P(\xi < x, \eta < y) \ . \qquad (64)$$

Folytonos esetben ebből deriválással megkaphatjuk az együttes sűrűségfüggvényüket $f(x,y)$-t is (79. ábra).

Az együttes eloszlás egyértelműen meghatározza a külön-külön vett valószínűségeloszlásokat, a peremeloszlásokat vagy vetületeloszlásokat. Ha $F_1(x)$ a ξ, $F_2(y)$ pedig az η eloszlásfüggvénye, akkor:

$$\begin{aligned} F_1(x) &= F(x,\infty) \\ F_2(y) &= F(\infty,y) \end{aligned} \ , \qquad (65)$$

azaz $F(x,y)$-nak az y illetve a x tengellyel párhuzamos „végtelenben” vett metszetei. Ha az együttes eloszlás folytonos, akkor a (25) összefüggés alapján ξ és η sűrűségfüggvényei, $f_1(x)$ és $f_2(y)$ is megadhatók. Ha ismerjük ezeket a függvényeket is, akkor bevezethetők a ξ valószínűségi változó $\eta = \boldsymbol{y}$ feltétel melletti, valamint az η valószínűségi változó $\xi = \boldsymbol{x}$ feltétel melletti feltételes sűrűségfüggvényei:

$$f_1(x \mid \boldsymbol{y}) = \frac{f(x,\boldsymbol{y})}{f_2(\boldsymbol{y})} \ , \qquad f_2(y \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x},y)}{f_1(\boldsymbol{x})} \ . \qquad (66)$$

Ezek a függvények az $f(x,y)$ együttes sűrűségfüggvénynek a feltételben megadott helyeken vett – az x, illetve az y tengellyel párhuzamos – metszetei (lásd például a sárga görbét, $f(x,11)$-t a 79. ábrán) elosztva a megfelelő normálási faktorral.

Mint az ábrán is látható, az $f(x,\boldsymbol{y})$ típusú függvények már „sűrűségfüggvényszerűek”, de nem normáltak, ami közvetlenül látszik abból, hogy a görbék alatti terület eltérő. Emiatt szükséges a megfelelő vetületeloszlás sűrűségfüggvényének az $\boldsymbol{y}$ helyen felvett értékét, $f_2(\boldsymbol{y})$-t felhasználni a normáláshoz.

Amennyiben

$$f_1(x \mid \boldsymbol{y}) = f_1(x) \ , \qquad (67)$$

azaz a feltételes és a feltétel nélküli sűrűségfüggvény megegyezik egymással, ami (66/1) alapján ekvivalens azzal, hogy

$$f_1(x) f_2(y) = f(x,y) \ , \qquad (68)$$

akkor a (16) és (17) összefüggések mintájára azt mondhatjuk, hogy ξ és η független valószínűségi változók (79a. ábra). (Az ábrán a függetlenség például abban mutatkozik meg, hogy a vetületi képen az ellipszis tengelyei párhuzamosak a koordináta tengelyekkel.)

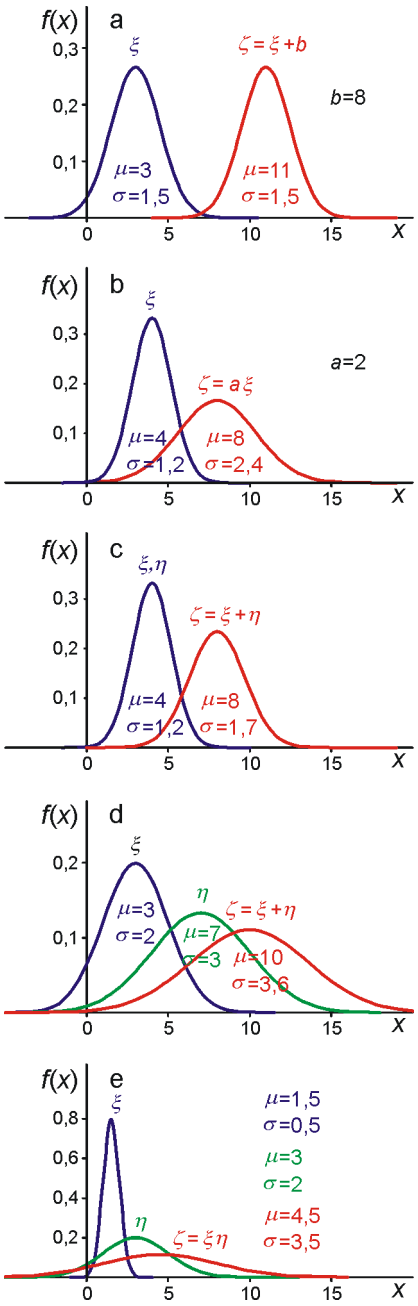
Általánosan úgy fogalmazhatunk, hogy ξ és η valószínűségi változók akkor függetlenek, ha tetszőleges (x,y)-ra teljesül, hogy a $\{\xi < x\}$, $\{\eta < y\}$ események függetlenek, azaz

$$P(\xi < x, \eta < y) = P(\xi < x) P(\eta < y) \ , \qquad (69)$$

vagy

$$F(x,y) = F_1(x) F_2(y) \ . \qquad (70)$$

A valószínűségi változók függetlenségével kapcsolatban ugyanazt mondhatjuk el, mint amit az események függetlenségével kapcsolatban mondtunk (27. megjegyzés). Egyes esetekben éppen azt kell majd megállapítanunk, hogy két változó független-e vagy sem, amire még visszatérünk (vö. 15.8. rész).



80. ábra
A normális eloszlású valószínűségi változók legegyszerűbb transzformációinak szemléltetése.

12.1. A valószínűségi változók transzformációi

Mivel a valószínűségi változók igen sokfélék, célszerű egyszerű matematikai műveletekkel olyan átalakításokat (transzformációkat) végezni rajtuk, amelyek után a transzformált új valószínűségi változó olyan tulajdonságú, hogy a korábban kidolgozott módszerek alkalmazhatók rájuk. A legegyszerűbb transzformációk között említhetjük egy állandó hozzáadását, vagy egy állandóval való szorzást, de két változó összeadása vagy szorzása is ide tartozik. Jelöljük ξ-vel, illetve η-val a transzformáció előtti változókat és legyen ζ (dzéta) a transzformált változó. Erről belátható, hogy szintén valószínűségi változó, így a kérdés az, hogy milyen lesz ζ eloszlása? A precíz bizonyítást ismét mellőzve, csupán józan eszünket használva sok esetben ki tudjuk találni az eredményt. (Követve előbbi módszerünket, az egyszerűbb szemléltetés kedvéért itt is normális eloszlású változókat használunk, de továbbra is érvényes, hogy a kapott eredmények más eloszlású változókra is igazak (80. ábra).)

Állandó hozzáadása**,** eltolási transzformáció: $\zeta = \xi + b$, ahol b állandó. Ha például ξ, (3;1,5) paraméterű normális eloszlású változó, és 8-cal pozitív irányba eltoljuk akkor ζ szintén normális eloszlású, de (11;1,5) paraméterű lesz, hisz az eltolás az eloszlás szélességét nem befolyásolja (80a. ábra).

Állandóval való szorzás**,** nyújtási (zsugorítási) transzformáció: $\zeta = a\xi$, ahol a állandó. a = 2 esetén minden érték megkétszereződik, így az eloszlás közepe, (például a várhatóértéke) és a szélessége is (például a szórása) (80b. ábra).

Ha az előbbi két transzformációt egymás után alkalmazzuk, méghozzá úgy, hogy $b = -\mu$, $a = 1/\sigma$, akkor standardizálást hajtunk végre. Általánosan:

$$\zeta = \frac{\xi - E(\xi)}{\sqrt{Var(\xi)}} . \qquad (71)$$

Amennyiben ξ normális eloszlású változó (μ,σ) paraméterekkel, akkor a (71) transzormációval a (0;1) paraméterekkel jellemzett standard normális eloszlású ζ változóhoz jutunk.

Két azonos eloszlású valószínűségi változó összege: $\zeta = \xi + \eta$, de $\xi \neq \eta$. Ha a két változó egyenlő is lenne, akkor visszakapnánk a 2-vel való szorzás eredményét. Itt azonban a véletlen kiválasztás miatt ξ lehetséges kisebb értéke mellé nagyobb valószínűséggel kerül η lehetséges nagyobb értéke és fordítva, ami ζ várhatóértékét nem befolyásolja, de az eloszlása keskenyebb lesz, mint a 2-vel való szorzás esetén. Ha azt is feltesszük, hogy ξ és η ***függetlenek***, akkor kicsit részletesebb számolással megmutatható, hogy 2-szeres helyett éppen $\sqrt{2}$ -szörös lesz az eloszlás szélessége (80c. ábra).

Két valószínűségi változó összege, (általános eset): $\zeta = \xi + \eta$. Mint az ábrán is látható (80d. ábra) a várhatóértékek összeadódnak (3 +7 = 10), a ***függetlenség*** fennállása esetén pedig a varianciák (szórásnégyzetek) is ($2^2 + 3^2 = 13 \approx 3{,}6^2$).

Két valószínűségi változó szorzata: $\zeta = \xi\eta$. ***Független*** valószínűségi változók várhatóértékekei összeszorzódnak (80e. ábra), a varianciákra pedig az alábbi, bonyolultabb (77) összefüggés érvényes.

A várhatóérték és a variancia néhány fontos tulajdonsága összefoglalva:

$$E(\xi + b) = E(\xi) + b \ , \qquad Var(\xi + b) = Var(\xi) \ , \qquad (72)$$

$$E(a\xi) = aE(\xi) \ , \qquad Var(a\xi) = a^2 Var(\xi) \ , \qquad (73)$$

$$E(\xi + \eta) = E(\xi) + E(\eta) \ . \qquad (74)$$

Ha ξ és η ***független*** valószínűségi változók, akkor

$$Var(\xi + \eta) = Var(\xi) + Var(\eta) \ , \qquad (75)$$

$$E(\xi\eta) = E(\xi)E(\eta) \ , \qquad (76)$$

$$Var(\xi\eta) = E^2(\xi)Var(\eta) + E^2(\eta)Var(\xi) + Var(\xi)Var(\eta) \ . \qquad (77)$$

Ha n darab ***független*** ξ_i változó varianciája mind ugyanakkora, azaz például $Var(\xi_i) = \sigma^2$ minden i-re, akkor (75) szerint

$$Var(\xi_1 + \xi_2 + \ldots + \xi_n) = n\sigma^2 \ . \qquad (78)$$

12.2. Két valószínűségi változó függősége, korreláció, regresszió

Láthattuk, hogy két valószínűségi változó függetlenségét a (70) egyenlőség egyértelműen leírja, így azt gondolhatjuk, hogy annak mintájára az

$$F(x,y) \neq F_1(x)F_2(y) \qquad (79)$$

összefüggés majd épp ilyen jól jellemzi a függőséget. Bár a statisztikus kapcsolat tényét valóban megadhatjuk így is, a (79) összefüggés nem ad számot a függőség, „erősségéről". Annak érdekében, hogy ilyen jellemzőhöz jussunk, először vizsgáljuk meg azokat a lépéseket, amelyek a (75) összefüggéshez vezetnek, azaz hogyan kaptuk meg két független valószínűségi változó összegének varianciáját (Mintafeladat).

Mintafeladat

Igazoljuk, hogy amennyiben ξ és η független valószínűségi változók, akkor $Var(\xi+\eta) = Var(\xi)+Var(\eta)$!

Megoldás: Először írjuk fel a varianciát a (42) definíció alapján, majd alkalmazzuk az összeg várhatóértékére vonatkozó (74) összefüggést és a kapott tagokat csoportosítsuk az alábbiak szerint:

$$Var(\xi+\eta) = E\left[\left((\xi+\eta)-E(\xi+\eta)\right)^2\right] = E\left[\left((\xi-E(\xi))+(\eta-E(\eta))\right)^2\right] = ,$$

ezután végezzük el a négyzetre emelést és ismét használjuk a (74) és (73/1) összefüggéseket:

$$= E\left[(\xi-E(\xi))^2+(\eta-E(\eta))^2+2(\xi-E(\xi))(\eta-E(\eta))\right] = E\left[(\xi-E(\xi))^2\right]+E\left[(\eta-E(\eta))^2\right]+2E\left[(\xi-E(\xi))(\eta-E(\eta))\right]$$

A kapott eredmény első két tagja definíció szerint $Var(\xi)$ és $Var(\eta)$, a pirossal áthúzott tag pedig a függetlenség fennállása esetén 0, hiszen ilyenkor a (76) összefüggés szerint a szorzat várhatóértéke egyenlő a várhatóértékek szorzatával, amelyek pedig külön-külön is 0-át adnak eredményül (vö. 62b. megjegyzés).

A fentiek alapján felmerül a gondolat, hogy amennyiben ξ és η függetlensége esetén az ***áthúzott tag*** 0-át ad eredményül, akkor két változó függőségének mértékét éppen egy ilyen mennyiséggel mérhetjük. Azt is megfigyelhetjük, hogy ez a tag a varianciára emlékeztet, sőt amennyiben $\eta = \xi$, akkor a 2-es faktortól eltekintve egyenlő is vele. Így bevezethetjük a kovarianciát, amellyel két változó statisztikus kapcsolatának „erőssége" jellemezhető:

$$Cov(\xi,\eta) = E\left[(\xi-E(\xi))(\eta-E(\eta))\right]. \qquad (80)$$

Ez a mennyiség elvileg $-\infty$ és ∞ között változhat, így a különböző feltételek esetén kapott kovarianciák összehasonlítása körülményes. Célszerű ezért egy hasonló, standardizált mennyiség bevezetése. Belátható, hogy a kovariancia abszolút értéke nem lehet nagyobb, mint a varianciák szorzatának négyzetgyöke, így a kovarianciát ezzel a számmal osztva, mindig –1 és 1 közé eső számértéket kapunk, amit korrelációs együtthatónak nevezünk:

$$R(\xi,\eta) = \frac{Cov(\xi,\eta)}{\sqrt{Var(\xi)Var(\eta)}}. \qquad (81)$$

Azt már megbeszéltük, hogy amennyiben ξ és η független valószínűségi változó, akkor $Cov(\xi,\eta) = 0$ és így $R(\xi,\eta) = 0$ is teljesül. Ha viszont $R(\xi,\eta) = 0$, akkor csak azt mondhatjuk, hogy ξ és η korrelálatlanok, ami **nem jelent** még feltétlenül **függetlenséget**. Ez kétségtelenül a korrelációs együttható egyik „hibája". A másik „hiba" az, hogy $R(\xi,\eta) = 1$ akkor és csak akkor teljesül, ha ξ és η között lineáris kapcsolat van, márpedig bármilyen függvénykapcsolat elegendő lenne ahhoz, hogy a legnagyobb mértékű függőségről beszéljünk (81., 82. megjegyzés).

Ha $R(\xi,\eta) > 0$, akkor azt mondhatjuk, hogy ξ és η között pozitív korreláció áll fenn. Ilyenkor abból, hogy $\xi > E(\xi)$, általában arra lehet következtetni, hogy $\eta > E(\eta)$. Negatív korreláció esetén viszont, azaz amikor $R(\xi,\eta) < 0$, ha $\xi > E(\xi)$, akkor általában $\eta < E(\eta)$. A megfelelő analógia felhasználásával, a (41) összefüggés segítségével a kovariancia, és annak alapján a korrelációs együttható adatrendszerre vonatkozóan is meghatározható (vö. (138); 175. megjegyzés):

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} = \frac{s_{xy}}{s_x s_y}. \qquad (82)$$

81. megjegyzés
Bizonyítható, hogy amennyiben ξ és η valószínűségi változók együttes eloszlása **normális** eloszlás, akkor a korrelációs együttható a két változó függőségének elméleti szempontból kifogástalan mérőszáma. (Nincs vele „hiba".) Ekkor ugyanis ξ és η között más függvénykapcsolat a lineárison kívül nem lehetséges, továbbá ilyenkor, ha $R(\xi,\eta) = 0$ akkor ξ és η független.

82. megjegyzés
A ξ és η valószínűségi változók együttes eloszlására úgy is tekinthetünk, mint egy kétdimenziós valószínűségi vektorváltozó komponenseinek együttes eloszlására, vagyis a vektorváltozó eloszlására.

Látva a korrelációs együttható „hibáit”, felmerül a kérdés, hogy van-e más mód arra, hogy két változó közötti statisztikus kapcsolat mértékét jellemezzük? **A statisztikus kapcsolat** a függvénykapcsolatnál általában „gyengébb”. **Az egyik változó a másikat nem határozza meg egyértelműen, de befolyással lehet rá**, azaz bizonyos feltételeket róhat ki rá. Folytonos esetben a feltételes sűrűségfüggvény éppen ilyen szituációk leírására alkalmas.

Ha például ismerjük az η valószínűségi változó $\xi = \boldsymbol{x}$ feltétel melletti sűrűségfüggvényét, akkor a (34) definíciós összefüggés alapján η feltételes várhatóértéke is meghatározható:

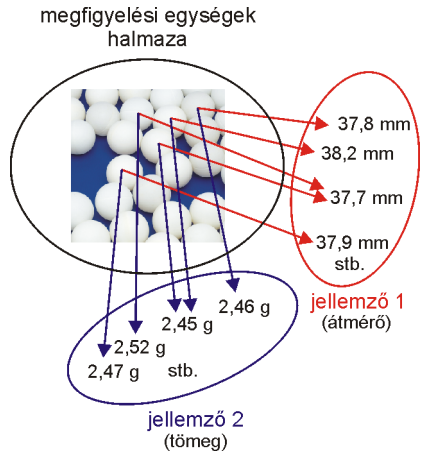
$$E(\eta \mid \xi = \boldsymbol{x}) = \int_{-\infty}^{\infty} y f_2(y \mid \boldsymbol{x}) \mathrm{d}y \,. \qquad (83)$$

Mivel ez a várhatóérték tetszőleges x-re megadható, ezért $E(\eta \mid \xi = x) = m(x)$ az x változó függvénye. Ezt a függvényt az η valószínűségi változó ξ-re vonatkozó regressziójának, a függvény grafikonját pedig regressziós görbének nevezzük. Ez a görbe jó jellemzője a két vizsgált változó statisztikus kapcsolatának. Ha például $m(x)$ monoton növekvő függvény, akkor ξ nagyobb értékeihez η-nak is általában nagyobb értékei tartoznak. Sok esetben a cél éppen az, hogy a regressziós görbét, illetve annak paramétereit meghatározzuk avégett, hogy η konkrét értékét közelítőleg megkapjuk ξ konkrét értékének ismeretében.

13.0. A statisztika alapfogalmai; alapsokaság, minta, változó

Mint azt a 2.0. részben már említettük, a valószínűségszámítási alapok nélkülözhetetlenek az induktív statisztikában, ezért az előző részekben elsősorban a statisztikához is szükséges legfontosabb valószínűségszámítási ismereteket foglaltuk össze. Mindezek után rátérhetünk olyan alapfogalmak bevezetésére is, amelyek kifejezetten a statisztikához kapcsolhatók.

A statisztikai elemzések mindig valamilyen alapsokaságra (sokaságra, populációra) vonatkoznak, arra irányulnak. Az alapsokaságot legszemléletesebben függvényként definiálhatjuk (83. ábra). A függvény a legegyszerűbb esetben két halmaz elemeinek valamilyen megfeleltetését jelenti.



83. ábra
Az alapsokaság mint függvény. A megfigyelési egységek itt pingponglabdák, melyekhez kétféle jellemzőt is hozzárendeltünk.

Esetünkben az egyik halmaz elemei a megfigyelési egységek, ezeken végezzük a megfigyeléseket, vizsgálatokat, méréseket. Ez a halmaz felel meg a függvény értelmezési tartományának, ennek a halmaznak az általános eleme a független változó. Megfigyelési egység lehet például egy személy, egy vérminta, egy pingponglabda stb.

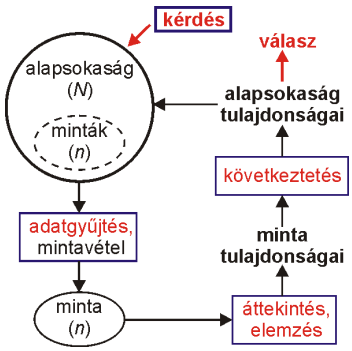
A másik halmaz elemei az adott szempontok, kísérleti feltételek vagy mérési utasítások szerint a megfigyelési egységekhez rendelhető adatok, azaz mennyiségi illetve minőségi jellemzők. Az egyszerűség kedvéért először feltesszük, hogy minden megfigyelési egységhez csak egy jellemző adat tartozik. Ezek halmaza felel meg a függvény „értékkészletének”, ennek a halmaznak az általános eleme a függő változó (84. megjegyzés). Ha a megfigyelési egységeket nem változtatjuk meg, de más jellemző adat iránt érdeklődünk, akkor természetesen egy másik alapsokaságról van szó (83. ábra).

Egy statisztikai elemzés akkor a legmegbízhatóbb, ha a vizsgálat az alapsokaság minden elemére kiterjed. Ilyen teljes körű adatfelvétel történik például népszámláláskor. Van azonban nagyon sok olyan eset, amikor ezt ***nem lehet megoldani***, mert mondjuk végtelen elemű a sokaság, vagy ***nem akarjuk megoldani***, mert például igen költséges lenne, vagy, mert nincsen értelme (például töréstesztek). Ilyenkor az alapsokaság helyett annak csak bizonyos, lehetőleg kevés számú elemét vizsgáljuk meg. A megvizsgálásra szánt elemek összessége a minta. A mintát reprezentatívnak nevezzük, ha az **híven tükrözi a sokaság minden fontos jellemzőjét**. Éppen emiatt a legtöbb esetben a mintaelemek kiválasztása, azaz a mintavétel a statisztikai elemzést megelőző kulcsfontosságú lépés.

A kiválasztás tulajdonképpen megfigyelést jelent (vö. 6.1. rész), számszerű adatok esetén azt figyeljük meg, hogy egy valószínűségi változó éppen milyen értékeket vesz föl. Így azt is mondhatjuk, hogy a mintavétel céltudatos adatgyűjtés. **A cél pedig az, hogy a minta megismert tulajdonságaiból a sokaság ismeretlen tulajdonságaira vonatkozóan vonjunk le következtetéseket** (85. ábra).

84. megjegyzés
Itt rögtön ki kell térnünk egy furcsaságra, mert van olyan eset, amikor a két halmaz egybeesik (identikus leképezés). Vegyük példaként a pingponglabda esetét. Ha a labdák méretének egyformaságát akarjuk ellenőrizni, akkor az összes legyártott labda az átmérő adatával együtt alkotja az alapsokaságot. Ennek elképzelése különösebb nehézséget nem okoz, a véges sokaság elemei direkt módon megfigyelhetők, szemünk előtt vannak, fizikailag is léteznek (83. ábra).

Ha azonban arra vagyunk kíváncsiak, hogy egy kiszemelt labda mennyire gömbölyű, akkor egyetlen labda összes, lehetséges átmérő adatát tekintjük, amelyek ilyenkor egybeesnek a sokaság elemeit meghatározó megfigyelési egységekkel. A sokaság tehát csak absztrakció, egy konkrét vizsgálat során a végtelen halmaz bármelyik eleme realizálódhat, de mindegyikre biztosan nem kerül sor.



85. ábra
A statisztikai cél eléréséhez vezető főbb lépések.

salmonellosis	1
scarlatina	2
egyéb bakteriális eredetű betegségek	3
hepatitis infectiosa	4
mononucleosis infectiosa	5
lyssa	6
egyéb vírusos eredetű betegségek	7
egyéb fertőző betegségek	8

86. táblázat
A korábban már szereplő fertőző betegségek számokká alakításának egy lehetséges módja.

87. megjegyzés
Orvosi szempont például az is, hogy egy vizsgálat során a betegeket nem tesszük ki felesleges kockázatoknak. Ebből következően mondjuk egy klinikai hatóanyag vizsgálata esetén számos kizáró ok merülhet fel, ami miatt az adott személy nem szerepelhet a mintában. Ilyen helyzet állhat elő, ha tudjuk, hogy a beadandó szer bizonyos esetekben allergiás tüneteket okozhat. Emiatt a konkrét személyek kiválasztásához használt adatlapok több olyan kérdést is tartalmazhatnak, amelyeknek látszólag semmi közük a vizsgálathoz, de éppen a kizáró okok keresésére alkalmasak.

88. megjegyzés
A szisztematikus mintavételi módszert leggyakrabban akkor szokták használni, ha a megfigyelési egységek külső beavatkozás nélkül, spontán módon „választód-nak ki". Ilyen eset valósul meg például, amikor a betegek elmennek az orvosi rendelőbe. Tegyük fel, hogy minden nap az első beteget választjuk a mintába. Ezzel a választással azonban feltehető, hogy a munkába sietők aktív dolgozók nagyobb arányban kerülnek a mintába, mint például a nyugdíjasok, akik jobban ráérnek. Így szisztematikus mintavétel esetén a szokásos statisztikai következtetéseket fenntartással kell kezelnünk. Általánosságban azt mondhatjuk, hogy mindig előfordulhat olyan eset, hogy a mintavételi szabályunk kapcsolatban van valamilyen más változóval, és ekkor könnyen lehet, hogy a minta már nem reprezentálja kellőképpen a sokaságot.

A változó fogalma már korábban szerepelt, mint egy halmaz általános eleme (vö. 2.2. rész). A ***statisztikában*** a változó valójában a **sokaság definíciójában szereplő függő változóval egyezik meg**.

Problémát csak az jelenthet, hogy ez minőségi jellemző is lehet, tehát nem feltétlenül szám, ami viszont szükséges lenne ahhoz, hogy valószínűségi változóként a valószínűségszámítás módszereit alkalmazni tudjuk. Ilyenkor azt a megoldást választhatjuk, hogy valamilyen általunk kialakított szabály szerint a minőségi jellemzőkhöz is számokat rendelünk (vö. 32. megjegyzés). Például a 8. táblázatban, illetve a 9. ábrán is szereplő betegségek esetében a megfigyelési egységek az egyes személyek, a ***változó*** „a fertőző betegség típusa", amelynek lehetséges „értékeit" mondjuk a 86. táblázatban feltüntetett megfeleltetések szerint adhatjuk meg. Természetesen azzal tisztában kell lennünk, hogy egy ilyen átalakítás után a kapott számoknak más a jelentésük, mint mondjuk egy pingponglabda átmérőnek, mert például az összeadás műveletének itt nem sok értelme van: scarlatina (2) + hepatitis infectiosa (4) ≠ lyssa (6).

Matematikai szempontból a megfigyelési egységek már lényegtelenek. Ha a mintavétel megtörtént, akkor a továbbiakban a statisztikai elemzésekhez nincs is szükség rájuk, csak a hozzájuk rendelt adatok, pontosabban azok mérőszámai fontosak. Természetesen a végső eredmények közlésekor a **mértékegységekről sem felejtkezhetünk meg**. Végső soron tehát az alapsokaság egy N elemű halmaz (adatrendszer), ahol N végtelen is lehet és ennek n elemű részhalmaza a minta (vö. 85. ábra).

13.1. Mintavételi módszerek

A fő kérdés az, hogy hogyan válasszuk ki a mintaelemeket ahhoz, hogy az elemzés alapján a sokaságra érvényes következtetéseket vonhassunk le. A legegyszerűbb a véletlen mintavétel, amikor az alapsokaság mindegyik eleme ugyanakkora eséllyel kerül a mintába. Bár a kiválasztás után a megfigyelési egységekhez tartozó számértékek adottak, a minta elemei valószínűségi változóknak is tekinthetők, hiszen a véletlenszerűségből következően más értékeket is kaphattunk volna.

A mintavétel lehet visszatevéses vagy visszatevés nélküli. Visszatevéses mintavételről akkor beszélünk, ha a mintába már kiválasztott elem a következő kiválasztás előtt „visszakerül" a sokaságba, annak érdekében, hogy az is újraválasztható legyen; más szavakkal egy megfigyelési egység többször is kiválasztható a sokaságból. Ilyenkor a sokaság minden mintaelem kiválasztása előtt ugyanolyan összetételű marad, tehát a mintába kerülő értékek, mint valószínűségi változók függetlenek egymástól. Az ilyen mintát független mintának nevezzük.

A függetlenség nyilvánvalóan nem teljesül a visszatevés nélküli mintavételre. Megjegyezzük azonban, hogy amennyiben a minta elemszáma kicsi az alapsokaság elemszámához képest, azaz $n << N$, akkor a sokaság összetétele visszatevés nélküli mintavétel esetén ugyan megváltozik, de csak csekély mértékben, ezért ilyenkor a kiválasztott mintaelemek lényegében függetlennek tekinthetők. Végtelen sokaság esetén a mintaelemek függetlensége automatikusan teljesül.

Ha az alapsokaságot egyik jellemzője szerint csoportosítjuk, például nemek, korcsoportok, vagy mondjuk orvosi szempontok (87. megjegyzés) szerint rétegekre bontjuk, és ezekből egymástól függetlenül veszünk egyszerű véletlen mintákat, akkor rétegzett mintavételről beszélünk. Ahhoz, hogy az ily módon választott mintából az egész sokaságra érvényes következtetéseket vonhassunk le, tudnunk kell, hogy az egyes rétegek az alapsokaság hány százalékát teszik ki, és az ezekből vett részminták elemszámát ennek megfelelően kell meghatároznunk. Az így nyert minta sok esetben reprezentatívabb lehet, mint amit egy közönséges véletlen mintavétel esetén kapnánk.

Véletlen mintavételre nem mindig van módunk. A legfőbb akadály általában az, hogy a sokaságbeli megfigyelési egységekről nincs teljes áttekintésünk. Ilyenkor nem az összes mintaelemet választjuk véletlenszerűen, hanem csak az elsőt, a többit pedig ebből kiindulva, valamilyen szabály szerint; például minden tizedik elemet választjuk ki, vagy háromperecenként mintavételezünk. Ezt a módszert szisztematikus mintavételi eljárásnak nevezzük (88. megjegyzés). Még sok egyéb mintavételi módszer létezik, de ezek kiértékelése további ismereteket igényelne, így megmaradunk az egyszerű véletlen mintavételnél, természetesen, adott esetben a kizáró okok figyelembevételével (vö. 87. megjegyzés).

13.2. A minta és az alapsokaság hasonlósága, a statisztika alaptétele

Matematikai szempontból a minta független – az alapsokaság eloszlásával megegyező – azonos eloszlású valószínűségi változók véges sorozata, de a mintaelemek tényleges kiválasztása után:

$$\xi_1 = x_1, \xi_2 = x_2, \ldots, \xi_n = x_n , \qquad (84)$$

az x_1, x_2, …, x_n értékhalmazt is mintának nevezzük. A minta mint adatrendszer eloszlásfüggvénye (vö. 6.1. rész), $F_n(x)$ az empirikus vagy tapasztalati eloszlásfüggvény, megkülönböztetésképpen az alapsokaság mint adatrendszer eloszlásfüggvénye, $F_N(x) \equiv F(x)$ pedig az elméleti eloszlásfüggvény. Abból a feltételből kiindulva, hogy az összes lehetséges (azonos elemszámú) minta egyformán valószínű, a mintából meghatározott bármely érték is olyan valószínűségi változó lesz, amelynek eloszlása a valószínűségszámítás segítségével megadható. Ennek megfelelően beszélünk tapasztalati illetve elméleti várhatóértékről, varianciáról stb. vagy általánosságban bármilyen egyéb számszerű jellemzőről.

A tapasztalati várhatóérték megegyezik az n elemű minta átlagával (vö. (27)). Ehhez hasonlóan az elméleti várhatóérték megegyezik az N elemű sokaság átlagával, amennyiben N véges. Itt az a fő probléma, hogy a sokaság minden elemét általában még akkor sem ismerjük ha N véges, nem beszélve a végtelen elemű sokaságokról. Éppen emiatt van szükségünk a becslésekre. Ehhez pedig, ha tehetjük, fel fogjuk használni az ismertetett matematikai modelleket, nevezetes eloszlásokat (vö. 11.0. rész).

Visszatérve az eredeti alapkérdésünkhöz, nevezetesen a hasonlóság és különbözőség problematikájához (vö. 1.0.), oldjuk meg az alábbi mintafeladatot.

Mintafeladat

Igazoljuk, hogy adott feltételek mellett a minta és az alapsokaság eloszlásfüggvénye nagyfokú hasonlóságot mutat!

Megoldás: Jelöljük ξ-vel az adott alapsokaságra vonatkozó változót, amelynek eloszlását az $F(x)$ elméleti eloszlásfüggvény adja meg. $F(x)$ definíció szerint minden x-re éppen a $\{\xi < x\}$ esemény valószínűsége, azaz $p(\xi < x) = F(x)$ (vö. (22)). Tetszőlegesen választott, de rögzített $\boldsymbol{x}$ esetén tekintsük a $\{\xi < \boldsymbol{x}\}$, $\{\xi \geq \boldsymbol{x}\}$ alternatívát. Egy n elemű mintához tartozó tapasztalati eloszlásfüggvény $\boldsymbol{x}$ helyen felvett értéke, $F_n(\boldsymbol{x})$ azt adja meg, hogy a mintában az $\boldsymbol{x}$-nél kisebb elemek hányad része található, így az $nF_n(\boldsymbol{x})$ szorzat éppen az $\boldsymbol{x}$-nél kisebb elemek darabszámával egyenlő (vö. (21)). Ez az eredmény úgy is interpretálható, hogy n-szeri ismétlés esetén, $nF_n(\boldsymbol{x})$-szer következik be a $\{\xi < \boldsymbol{x}\}$ esemény. A binomiális eloszlás éppen az ilyen ismételt alternatívák matematikai modellezésére való. Ezt felhasználva azt mondhatjuk, hogy az $nF_n(\boldsymbol{x})$ valószínűségi változó eloszlása az $(n, p = F(\boldsymbol{x}))$ paraméterű binomiális eloszlással adható meg. Ennek várhatóértéke:

$$E(nF_n(\boldsymbol{x})) = np = nF(\boldsymbol{x}) \text{ (vö. (48)),}$$

varianciája pedig:

$$Var(nF_n(\boldsymbol{x})) = np(1-p) \text{ (vö. (49)).}$$

A (73) összefüggés felhasználásával, majd n-nel illetve n^2-tel minkét oldalt elosztva kapjuk:

$$E(F_n(\boldsymbol{x})) = F(\boldsymbol{x}) \quad , \quad Var(F_n(\boldsymbol{x})) = \frac{p(1-p)}{n} \leq \frac{1}{4n} \quad .$$

Az utolsó egyenlőtlenség azon alapul, hogy $p(1–p)$ akkor maximális, ha $p = (1–p) = 0{,}5$ (vö. 11.1. rész).

Mivel x-et tetszőlegesen megválaszthatjuk, ezért a kapott két egyenlőség értelmében minden x-re igaz, hogy $F_n(x)$ $F(x)$ körül ingadozik, mégpedig n növekedtével (átlagosan) egyre kisebb mértékben.

A mintafeladat alapján megállapíthatjuk, hogy egy adott alapsokaság és a belőle vett – elég nagy elemszámú – minta esetén, **a tapasztalati eloszlásfüggvény nagyon jól megközelíti az elméleti eloszlásfüggvényt**. A matematika szigorú szabályai szerint ennél erősebb állítás is megfogalmazható, de lényegében ez a statisztika alaptételének szemléletes tartalma. Ezek szerint azt várjuk, hogy **minél gyakoribb egy értékcsoport előfordulása a mintában, annál valószínűbb a megjelenése az alapsokaságban is**. Ezen a tételen alapszik mindazon eljárások jogosultsága, amelyek során valamely valószínűségi változó eloszlására mintavétel útján következtetünk.

13.3. A leíró és az induktív statisztika kapcsolata

A statisztikai módszerek végső célja a következtetés. Ennek eléréséhez elengedhetetlenül szükséges a leíró statisztika alkalmazása, melynek első lépése az adatgyűjtés, további eszközei a különféle táblázatok, diagramok (vö. 2.1. rész), illetve az adatrendszerből meghatározható számszerű jellemzők (vö. 10.0. rész). Tipikus a leíró statisztika használata akkor, amikor teljes körű adatfelvétel történik, például népszámlálási vagy választási adatok, bejelentési kötelezettséggel járó fertőző betegségek stb. esetén. Ilyenkor a „valódi" induktív statisztika alkalmazására nem is kerül sor. A leíró statisztika viszont azért fontos, mert nélküle a hatalmas mennyiségű adat áttekinthetetlen és így használhatatlan lenne. Elképzelhetjük, hogy mire mennénk például a népszámlálási adatokkal, ha azokat feldolgozatlanul, úgymond ömlesztve tenné közzé a Központ Statisztikai Hivatal.

Hasonló a helyzet akkor is, amikor ugyan nincs módunk arra, hogy a sokaság minden elemét megvizsgáljuk, de viszonylag nagy elemszámú minta áll rendelkezésünkre. A kis elemszámú minták esetén pedig a számszerű jellemzők meghatározásán túl azért előnyös a leíró statisztika módszereinek – különösen a grafikus megjelenítésnek – az alkalmazása, mert észrevehetjük a minta olyan tulajdonságait, amelyek döntően befolyásolhatják az elemzésükhöz legmegfelelőbb induktív statisztikai módszerek kiválasztását.

Összefoglalva azt mondhatjuk, hogy a leíró és induktív statisztika a legtöbb esetben nem választható szét. A folyamat az adatgyűjtéssel kezdődik, ezt követi az adatok lényegre törő áttekintése, majd elemzése annak érdekében, hogy a megfigyelésekből általános érvényű következtetéseket vonhassunk le (vö. 85. ábra).

A statisztikai következtetések sémája kicsit hasonlít a logikai következtetések sémájához. A logikában a szillogizmus a következtetés egyik fajtája, amelyben bizonyos dolgok megállapításából **szükségszerűen következik** valami más. Az általánosból következtetünk az egyedire, azaz a következtetésünk deduktív (89. megjegyzés).

A statisztikai következtetés lényege éppen az, hogy itt az egyediből szeretnénk az általánosra következtetni, azaz a következtetésünk induktív (90. megjegyzés). Innen származik az induktív statisztika elnevezés is.

Azt mondhatjuk tehát, hogy a statisztikai következtetés annyiban tér el a logikaitól (ami persze nem jelentéktelen különbség), hogy míg a **logikai következtést teljes** (100%-os) **bizonyossággal** állíthatjuk**,** addig a **statisztikait** csak **adott,** (100%-nál mindig kisebb) **bizonyossággal**. Ebből következik tehát, hogy a statisztikai következtetéseknél **tévedhetünk**.

A megbízhatóság, illetve a tévedési valószínűség pontos jelentését nem könnyű megérteni. A legegyszerűbb, ha úgy gondolunk rá, hogy ha sokszor alkalmazzuk a szóban forgó módszert, akkor várhatóan az esetek hány százalékában kapunk helyes, illetve téves eredményt. Ha például valamit 95% bizonyossággal állítunk, akkor az azt jelenti, hogy átlagosan minden 100 eset közül 5 esetben tévedünk; a bizonytalanság 5%-os (91. megjegyzés). Állításaink biztonságát tehát számszerűen is kifejezhetjük.

A következtetés újrafogalmazásával a bizonytalanságot általában tetszés szerint csökkenthetjük, ez azonban többnyire azzal a veszteséggel jár, hogy állításunk, azaz a következtetés egyre semmitmondóbbá válik. Erre is lássunk egy példát (92. megjegyzés)! Ha tehát következtetésünk bizonytalanságát ily módon csökkentjük, akkor annak általában az az ára, hogy a következtetés értéke, használhatósága is csökken. Következtetéseinket tehát e két ellentétes tendencia figyelembevételével kell levonnunk.

Az induktív statisztika két legjellemzőbb feladata a becslés és a hipotézisvizsgálat. A becslés a „Mennyi? Mekkora? Hány százalék? stb." kérdésekre vár választ, mégpedig egy vagy néhány számot. A hipotézisvizsgálatban ezzel szemben „Igen/Nem" választ várunk az „Igaz-e? Fennáll-e? Van-e összefüggés ... ? Van-e hatása ... ? **Van-e különbség ... ?** (vö. 1.0. rész) stb." típusú kérdésekre.

89. megjegyzés
A klasszikus példa szerint: ***minden görög halandó, Szókratész görög, tehát Szókratész halandó.***

Ha matematikai szempontból vizsgáljuk ezt a példát, akkor azt mondhatjuk, hogy az első állítás egy halmaz minden elemére vonatkozik (minden görög), így nyilvánvaló, hogy amennyiben ebből a halmazból választok ki egy elemet (Szókratész), akkor az állítás erre is igaz.

90. megjegyzés
Ha megfordítjuk a 89. megjegyzésben szereplő állítások sorrendjét: Szókratész halandó, Szókratész görög, tehát minden görög halandó, akkor egy halmaz egyetlen elméről állítok valamit, de az egyáltalán nem biztos, hogy ez az állítás a halmaz minden elemére igaz, így a végső állítás is bizonytalan.

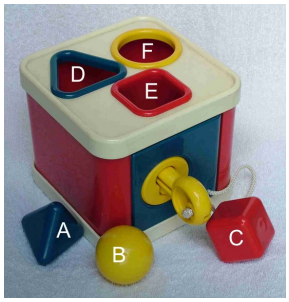
A helyzeten úgy javíthatunk, ha a kérdéses halmaznak több elemét is megvizsgáljuk és mondjuk kiderül, hogy Platón és Arisztotelész is halandó (és természetesen görögök). Ebből azonban még továbbra sem következik a végső állítás igazsága, ugyanis ***logikailag*** elképzelhető, hogy egy nem halandó görögöt is megfigyelhetünk …

91. megjegyzés
Egy ilyen állítás értékére talán jobban rávilágíthatunk a következő példával. Annak a valószínűsége, hogy egy dobókockával 6-ost dobjunk 1/6, kb. 17%.

Ha szabályos dobókocka helyett egy ikozaédert (20 lapú szabályos testet) használnánk, aminek minden lapja meg van számozva, akkor annak a valószínűsége, hogy 20-ast dobjunk 1/20 = 0,05, azaz 5%. Átlagosan tehát 20 dobásból 19-szer nem 20-as jön ki, **de a maradék egy kijöhet az első dobásra is**. „Csak" ennyit jelent a 95%-os „bizonyosság".

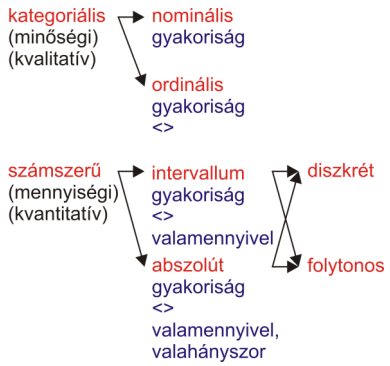
92. megjegyzés
Egy bankrablást követően az eseményekről beszámoló ***első*** rendőri ***jelentés*** így írt: „A rendőrség az eddigi adatok birtokában azt feltételezi, hogy a tettesek saját gépkocsijukon hagyták el a helyszínt."

Később a rendőrségi szóvivő a szemtanúk meghallgatása után már így nyilatkozott: „A szemtanúk látták, amint az elkövetők gépkocsiba szállnak; a rendőrség megállapította, hogy a tettesek vagy saját gépkocsijukon hagyták el a helyszínt, vagy taxival távoztak, vagy lopott, esetleg bérelt autót használtak".



változó / eset	alak	szín	méret (cm)	dimenzió
B	gömb	**sárga**	4,2	3
A	gúla	**kék**	4,6	3
C	kocka	**piros**	3,8	3
F	kör	**sárga**	4,5	2
D	háromszög	**kék**	4,9	2
E	négyzet	**piros**	4,1	2

93. ábra
A fenti gyermekkocka játékhoz kapcsolódó adatok elrendezése táblázatban.



94. ábra
A változók osztályozása többféle szempontból. (Kékkel az elvégezhető műveleteket tüntettük fel.)

95. megjegyzés
Külön említést érdemelnek azok a nominális változók, amelyeknek csak két értékük van. Ezek az úgynevezett dichotom vagy bináris változók. Ezekre egészen speciális elemzési módszereket dolgoztak ki.

Mivel az ilyen változók gyakran természetes módon rendezettek, – például amikor a két lehetséges érték igen/nem, van/nincs, pozitív/negatív – ezért bizonyos elemzésekben ordinális változóknak is tekinthetjük őket, így például beszélhetünk két tulajdonság megléte között fennálló pozitív vagy negatív korrelációról is.

96. megjegyzés
Fontos megjegyeznünk, hogy sok esetben egy változó típusa nem eleve adott, hanem mi dönthetjük el, hogy milyennek célszerű tekintenünk (vö. 32. megjegyzés). Sőt, már azt is mi döntjük el, hogy hogyan mérjünk meg egy adott mennyiséget, mik legyenek a megfigyelési egységek. A választásban az is szerepet játszik, hogy egyáltalán mi vizsgálható, mi milyen pontossággal mérhető. Természetesen a megfigyelési egységek megválasztásától függően a megfigyelt adatok és azok típusa, de még a minta elemszáma is különböző lehet az egyik vagy másik esetben. A választás azt is meghatározhatja, hogy mely statisztikai módszereket alkalmazhatjuk és melyeket nem, ugyanis a módszerek alkalmazhatósági feltételei is eltérőek lehetnek.

13.4. Adat típusok, a változók osztályozása

Mielőtt az induktív statisztika két legfontosabb fejezetére rátérnénk, tekintsük át az adatokkal kapcsolatban felmerülő lehetséges problémákat.

Egy vizsgálat során az összegyűjtött adatokat, legtöbbször egy olyan táblázatba írjuk be (általában valamilyen táblázatkezelő program segítségével), amelynek minden sora egy megfigyelési egységnek, oszlopai pedig az egyes mért vagy megfigyelt adatoknak felelnek meg. A sorokat eseteknek, az oszlopokat változóknak nevezzük (93. ábra).

Bár eredetileg adataink számértékek és szöveges adatok egyaránt lehetnek, az induktív statisztikában a valószínűségszámítás használata – amelyben a kvantitatív változók matematikai modellje a valószínűségi változó – megkívánja azt, hogy adataink számok, azaz kvantitatív változók legyenek. Ezért kell számokká alakítanunk a minőségi jellemzőket is (vö. 86. táblázat). Ez az átalakítás azonban, mint láthattuk új problémákat vet föl, nevezetesen **a mesterségesen generált számokkal bizonyos műveletek nem végezhetők el**, jobban mondva nem értelmezhetők. Egyes műveletek elvégzése pedig elkerülhetetlen ahhoz, hogy adatainkat az adott statisztikai eljárásnak megfelelő formára hozzuk, azaz transzformáljuk. (Csak egy másik példát is említve: a személyi azonosító számokból átlagot számolni nem nagyon értelmes dolog.) Így igen fontos annak ismerete, hogy adatainkkal milyen műveleteket végezhetünk. Ez az alapja a változók egyik fontos osztályozásának (94. ábra).

Eszerint az egyik nagy csoport a **minőségi** (kvalitatív) változó, amelyet kategoriálisnak is neveznek, a másik a **mennyiségi** (kvantitatív) változó, amely számszerű változót jelent. Ezek ismét két-két csoportra oszthatók: nominálisra és ordinálisra, valamint csak intervallumként illetve abszolút értelemben is használhatókra. A számszerűek esetében megkülönböztetünk még **diszkrét** és **folytonos** változókat, de ezt a problémát már megbeszéltük (vö. 6.1. rész).

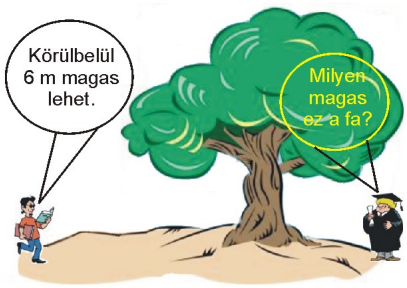
A ***nominális*** változó csak megnevez, csoportba sorol. Azt lehet megadni, hogy az adat egy csoportba beletartozik vagy nem, eleme vagy nem eleme egy adott halmaznak. Az ilyen adatokat gyakorisági adatoknak is nevezik. (Hány adat van egy csoportban?) A nevek, telefonszámok, postai irányítószámok, színekkel jellemzett tulajdonságok mind-mind ilyen típusú változók. Még ha számok is (mert például számokká alakítottuk őket), számolni valójában nem lehet velük (95. megjegyzés).

Az ***ordinális*** változó a nominálishoz igen hasonló, de értékeinek egyértelmű természetes sorrendje van, így az előzőkön túlmenően közöttük a kisebb vagy nagyobb (< >) reláció is értelmezhető. Ilyenkor a „hány adat van egy csoportban?” kérdést kiegészíthetjük a legalább, legfeljebb kifejezésekkel, ami a nevezettnél kisebb illetve nagyobb értékhez tartozó csoportok gyakoriságának összegzését jelenti. Számolni elvileg az ilyen változókkal sem lehetne, de sokszor mégis megtesszük (statisztikai szempontból gyakran megalapozatlanul). Ilyen változó például a vizsgaeredmény, amelyet egyes országokban betűkkel, más országokban számokkal adnak meg. Ez utóbbi esetben még átlagot is számolunk belőle. (Ha ugyanezt a „kiváló”, „jó”, „megfelelt” szavakkal fejezzük ki, eszünkbe sem jutna ilyen.) Ordinális változót használunk például akkor is, amikor egy betegség lefolyásának súlyosságát minősítjük (enyhe, közepes, súlyos).

Intervallumként használható változó esetén megadhatjuk, hogy például két érték mennyivel tér el egymástól. Az egyik mennyivel nagyobb, mint a másik, vagy fordítva, de a „kétszer akkora”, „harmadannyi” nem értelmezhető. Ilyen például a Celsius-fokokban mért hőmérséklet, mert skálájának 0 pontja önkényes. (Miért pont a víz fagyáspontjának hőmérsékletét választották?) A 40 °C-ról mondhatjuk, hogy 20 °C-kal melegebb a 20 °C-nál, de azt nem mondhatjuk, hogy kétszer melegebb. Az **ilyen változó esetén a legtöbb statisztikai módszer már biztosan alkalmazható** (96. megjegyzés).

Abszolút értelemben is használható az a változó, amelynél az arányok is értelmezhetők. A Kelvinben mért hőmérséklet már ilyen, de megemlíthetjük a kilogrammban mért tömeget is. Ezekkel a változókkal elméletileg korlátozás nélkül számolhatunk. Ritka az olyan statisztikai eljárás, amelyik csak ilyen adatokra alkalmazható. Példaként azért megemlíthetjük a relatív szórás fogalmát, amely egy aránnyal fejezhető ki, tehát ide tartozik (vö. 119. megjegyzés).

14.0. Becslés, statisztikai becslés, jó becslés



97. ábra
A hétköznapi becslés: az ember feltételezett testmagassága alapján adunk becslést a fa magasságára.

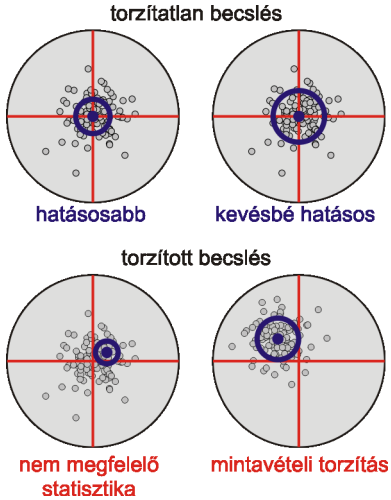
A becslés a hétköznapi életben a mérésnek egy kevéssé kifinomult változata. Ha valaki mér, akkor: „Egy nagyságot összehasonlít egy meghatározott egységgel, és az összehasonlítás eredményét számokban fejezi ki.” Ha becslést végez, akkor ugyanezt csak hozzávetőlegesen teszi. Tehát úgy tűnik, hogy a mérés valaminek a pontosabb, a becslés a kevésbé pontos meghatározása (97. ábra).

A statisztikai becslés mást jelent. Ugyanis az így kapott eredmény épphogy pontosabb, mint egy „közönséges” mérés eredménye. Ennek bemutatására példaként válasszuk a pulzusszám mérést. A pulzusszám valójában a szívverés frekvenciája, folytonos jellemző és csak az egyszerűség kedvéért, illetve megszokásból használunk diszkrét (egész) értékeket. Mértékegysége az 1/perc. Ezek figyelembevételével a továbbiakban csak a mérőszámokkal dolgozunk és tegyük fel azt a kérdést, hogy: ***mennyi az egészséges ember normális pulzusszáma***?

A kérdés megválaszolása érdekében az egészséges emberek közül választunk egyet és nyugodt körülmények között „nagyon pontosan” megmérjük a pulzusszámát. Mondjuk az eredmény 75 és úgy gondolhatjuk, hogy megkaptuk a „nagyon pontos” választ. Ha azonban valaki már egy kicsit statisztikával „fertőzött”, akkor kételkedése felébred és egyértelmű válasz helyett újabb kérdések merülnek fel benne: biztos, hogy ez a jó eredmény, nem hibáztam valahol? Ha még azt is tudja az illető, hogy a mérés eredménye a „véletlentől” is függ, tehát a legnagyobb igyekezettel sem tudunk „hibátlanul” mérni, akkor arra az elhatározásra jut, hogy újra mér. A **többszöri mérés** viszont nem más, mint **mintavétel**, és a kiválasztott mintaelemeknek valamilyen közös jellemzőjét (például az átlagát) fogjuk arra használni, hogy a feltett kérdésre választ adjunk.

A helyzet az, hogy a feltett kérdésben már benne van az, hogy bennünket egy sokaság jellemzője érdekel és valójában nem arra vagyunk kíváncsiak, hogy egy vagy több egészséges embernek mennyi a pulzusszáma. Ebben az esetben a sokaság nem egy konkrétan megfigyelhető halmazhoz köthető, hanem egy absztrakt végtelen halmazhoz, amely az összes lehetséges „normális” pulzusszámot tartalmazza.

A statisztikai becslés alapjait tulajdonképpen már tisztáztuk, nevezetesen amikor megfogalmaztuk a statisztika alaptételét. Ennek kapcsán azt is mondhatjuk, hogy az elméleti eloszlásfüggvény a tapasztalati eloszlásfüggvénnyel becsülhető (vö. 13.2. rész). De amennyiben ez igaz, akkor az elméleti eloszlásfüggvény számszerű jellemzői is (várhatóérték, variancia stb.) becsülhetők a tapasztalati eloszlásfüggvény megfelelő jellemzőivel. Általánosságban úgy fogalmazhatunk, hogy **egy számszerű jellemző elméleti értékének becslésére a $\xi_1, \xi_2, \ldots, \xi_n$ mintabeli változók egy függvényét használjuk**. Ezt a függvényt statisztikai függvénynek vagy röviden statisztikának nevezzük. Az adott jellemző közelítésére konstruált statisztika pedig a jellemző becslése.



98. ábra
A „jó” és kevésbé jó becslések bemutatása a céltáblán. A becsülendő jellemzőnek a céltábla közepe felel meg. A kék kör közepe mutatja a becslés várható értékét, sugara pedig arányos a szórásával.

A torzítást egyfelől a nem megfelelő statisztikai függvény (becslés) használata, másfelől az elhibázott mintavétel is okozhatja. Ilyen lehet például egy nem kellően átgondolt szisztematikus mintavétel (vö. 88. megjegyzés).

Mivel maguk a mintaelemek valószínűségi változók, így minden becslés is természetesen az, amelynek adott eloszlása van. Legyen ω az alapsokaság egy számszerű jellemzője és a $w(\xi_1, \xi_2, \ldots, \xi_n)$ az erre vonatkozó becslés. Ahhoz, hogy a becslést jónak mondjuk, a w függvénynek bizonyos jó tulajdonságokkal kell bírnia.

1. A w becslés torzítatlan, ha várhatóértéke megegyezik ω-val:

$$E(w) = \omega\,. \tag{85}$$

2. Ha ω-nak w_1 és w_2 is torzítatlan becslése, akkor a kisebb varianciájút nevezzük hatásosabbnak, azaz, ha

$$Var(w_1) < Var(w_2)\,, \tag{86}$$

akkor w_1 hatásosabb becslése ω-nak, mint w_2. A torzítatlan becslések közül a minimális szórásút hatásos becslésnek nevezzük. Két torzítatlan becslés közül tehát könnyű kiválasztani a jobbat. A döntés azonban nem ilyen egyértelmű akkor, ha az egyik becslés torzított (98. ábra). (Kisebb torzítás még elfogadható abban az esetben, ha ettől lényegesen kisebb szórású becsléshez jutunk.)

3. A $w(\xi_1, \xi_2, \ldots, \xi_n)$ becslést konzisztensnek mondjuk, ha a minta elemszámának növekedtével egyre csökken annak a valószínűsége ($n \to \infty$ esetén 0-hoz tart), hogy a $|w - \omega|$ eltérés egy előre megadott küszöbértéknél nagyobb legyen.

4. A $w(\xi_1, \xi_2, \ldots, \xi_n)$ becslés elégséges, ha ugyanabból a mintából nem lehet több ismeretet nyerni ω -ra vonatkozóan. Ezen azt értjük, hogy a statisztika a minta eloszlásának adott jellemzőjére vonatkozóan minden ismeretet magába sűrít. Egy másik megfogalmazás szerint – ami más megvilágításba helyezi az elégséges becslés fogalmát – belátható, hogy amennyiben w ω -nak hatásos becslése, akkor az mindig egy elégséges becslés függvényeként adható meg.

14.1. A becslés pontossága, hibája

Egy becslés hibáján a becslés (w) és a becsülendő jellemző (ω) közötti eltérést, azaz $|w-\omega|$ -t értjük (99. megjegyzés). Ugyanúgy, mint a becslés maga, a hiba is valószínűségi változó, mintáról mintára változik, ezért a becslés pontosságát a hiba eloszlásával, illetve annak valamely jellemzőjével számszerűsíthetjük. A szokásos jellemző az átlagos négyzetes eltérés (mean squared error, *MSE*), azaz az eltérés négyzetének várhatóértéke:

$$MSE = E\left[(w-\omega)^2\right]. \qquad (87)$$

99. megjegyzés
Ha a becsülendő jellemző a medián, aminek a (35) definíciós összefüggés alapján nemcsak egyetlen érték felelhet meg, hanem egy intervallum is, ilyenkor célszerű az egyértelművé tétel például oly módon, hogy az intervallum közepét tekintjük mediánnak. (A mindennapi gyakorlatban a két szélső érték számtani közepét szokták használni.)

Minél kisebb az átlagos négyzetes eltérés, annál pontosabb a becslés. Egy becslés pontossága valójában két, egymástól logikailag ***független*** tényező eredője. Egyik a véletlen hiba vagy véletlen ingadozás, a másik a tendenciózus hiba vagy torzítás. Torzításnak tekintjük a hibát, ha a jellemző becsült értéke szisztematikusan kisebb vagy nagyobb magánál a becsülendő jellemzőnél. A becslés véletlen hibáját rendszerint a becslés szórásával adjuk meg, amelyet nem szórásnak, hanem standard hibának (standard error, *SE*) nevezünk (100. megjegyzés):

$$SE(w) = \sqrt{Var(w)} = \sqrt{E\left[\left(w - E(w)\right)^2\right]}. \qquad (88)$$

100. megjegyzés
Megkülönböztetésképpen ha a ***minta*** szórásáról beszélünk, akkor a megfelelő angol kifejezés figyelembevételével (standard deviation) *SD*-t fogunk használni.

A becslés torzítását (*bias*) a becslés várhatóértékének a becsülendő jellemzőtől való eltérésével mérjük:

$$bias(w) = E(w-\omega) = E(w) - \omega. \qquad (89)$$

Viszonylag egyszerű algebrai átalakításokkal belátható, hogy az átlagos négyzetes eltérés éppen megegyezik a standard hiba és a torzítás négyzetösszegével (Mintafeladat):

$$MSE = SE^2(w) + bias^2(w). \qquad (90)$$

Mintafeladat

Igazoljuk a (90) összefüggést, nevezetesen, hogy az átlagos négyzetes eltérés a standard hiba és a torzítás négyzetösszege!

Megoldás: Induljunk ki a (87) definíciós összefüggésből és a $(w - \omega)$ különbséghez adjuk hozzá a $0 = -E(w) + E(w)$ kifejezést (ami természetesen nem jelent érdemi változást), majd végezzük el a négyzetre emelést:

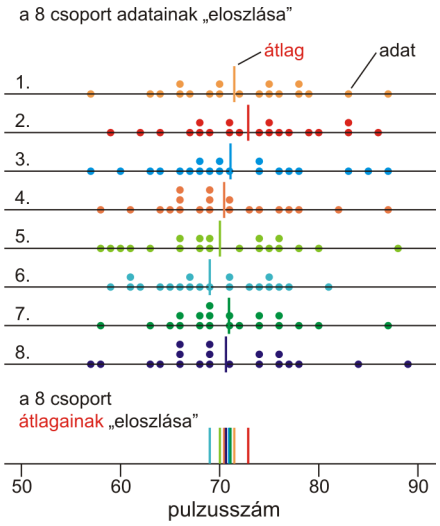
$$E\left[\left(\left(w - E(w)\right) + \left(E(w) - \omega\right)\right)^2\right] = E\left[\left(w - E(w)\right)^2 + \left(E(w) - \omega\right)^2 + 2\left(w - E(w)\right)\left(E(w) - \omega\right)\right].$$

A szögletes zárójelen belül a várható érték képzése a (74) összefüggés alapján tagonként végezhető el. Látható, hogy az első tag várható értéke éppen $Var(w) = SE^2(w)$-vel egyenlő (vö. (88)), a második tag a torzítás négyzete, ami egy szám ezért várható értéke önmaga (vö. (89)). A harmadik tag várható értéke pedig 0, hiszen az első és az utolsó tényező egy-egy szám (vö. (73/1)), a $(w - E(w))$ tényező várható értéke pedig 0 (vö. 62b. megjegyzés), így a 0-val szorozás 0-t ad eredményül.

Minél kisebb az átlagos négyzetes eltérés, annál pontosabb a becslés, tehát a (90) összefüggés szerint akkor a legjobb a becslés, ha mindkét tag, azaz a véletlen hiba és a torzítás is kicsi. Azonban míg a véletlen hiba számszerűen mérhető, addig a torzítás többnyire nem mérhető becslési hiba.

Előfordulhat, hogy egy jellemzőre nem tudunk torzítatlan becslést adni, de olyat igen, amelynek **torzítása a minta elemszámának növekedtével egyre kisebb lesz,** azaz $n \to \infty$ esetén 0-hoz tart. Az ilyen becslést aszimptotikusan torzítatlannak nevezzük. (A statisztikában egy tulajdonságra általában is akkor mondjuk, hogy „aszimptotikus”, ha az egyre nagyobb elemszámú mintákra egyre jobban teljesül.) Amennyiben hasonló állítás megfogalmazható a standard hibára nézve is, akkor bizonyítható, hogy a becslés konzisztens is (101. megjegyzés).

101. megjegyzés
Azt szokás mondani, hogy a konzisztencia amolyan minimális követelmény egy becsléssel szemben. Ha még ezt sem teljesíti, akkor nem tekinthető jó becslésnek.



102. ábra
Nyolc, 20 fős tanulócsoport hallgatóinak pulzusszáma és a mintaátlagok szemléltetése.

103. megjegyzés
A mintaátlag várható értékének (91) meghatározásakor a (73/1) és a (74) tulajdonságokat, a mintaátlag szórásnégyzetének (92) meghatározásakor pedig a (73/2) és – a mintaelemek függetlensége mellett – a (75) tulajdonságot használtuk ki.

104. megjegyzés
Mindezen túl a centrális határeloszlás tétel alkalmazásával (vö. 11.1. rész) bizonyítható, hogy a mintaátlag eloszlása a minta elemszámának növekedtével – a vizsgált változó eloszlásától függetlenül – egy normális eloszláshoz tart, tehát aszimptotikusan normális eloszlású.

105. megjegyzés
Itt láthatóan bajba kerülünk az elnevezésekkel, ugyanis eddig az „elméleti” és a „tapasztalati” szavakkal utaltunk arra, hogy a sokaságról, vagy a mintáról van szó. A várható érték és átlag illetve a variancia és szórásnégyzet, pedig nagyjából szinonimákként használhatók. Ha azonban az ”eltérés négyzetek átlagát” tekintjük, mint általános meghatározást, akkor nem világos, hogy minek, mitől való eltéréséről beszélünk. A (94) összefüggésben épp ez a kettősség nyilvánul meg, nevezetesen a mintaelemek eltérését mérhetjük a mintaátlagtól, de a sokaság átlagától is.

14.2. A mintaátlag és a mintaszórás néhány fontos tulajdonsága

A leggyakrabban használt becslő függvény a mintaátlag vagy ***tapasztalati átlag*** (tapasztalati várhatóérték) $(\bar{x})$ illetve $(\bar{\xi})$, amellyel az alapsokaság átlagát (μ -t), azaz az ***elméleti átlagot*** (elméleti várhatóértéket) becsüljük (vö. (27)). (A kétféle jelölést az indokolja, hogy szeretnénk megkülönböztetni az adott mintából kiszámolt konkrét mintaátlagot $(\bar{x})$ és a mintákból kiszámolható mintaátlagot, mint valószínűségi változót $(\bar{\xi})$ (vö. (84)).

A tulajdonságok egy része szembetűnő, ha több mintát és a mintaátlagokat is szemléltetjük. Ennek érdekében térjünk vissza a pulzusszám méréshez és tekintsük a 102. ábrán feltüntetett adatokat. Megfigyelhető, hogy az átlag mint középérték a minta változására nem nagyon érzékeny, hiszen a számolás az összes mintaelem figyelembe vételével történik, így – főleg nagyobb elemszámú minták esetén – egy -egy adatnak csak kevés módosító szerep jut. Ennek az a következménye, hogy a különböző mintákból számolt átlagok csak kevéssé térnek el egymástól. Azt is mondhatjuk, hogy a mintaátlagok jóval kevésbé „szóródnak”, mint az adatok.

Mit jelent ez kvantitatívabban? Tudjuk, hogy (reprezentatív minta esetében) a mintaelemek mint valószínűségi változók eloszlása az alapsokaság eloszlásával megegyezik (vö. 13.2. rész). Így, ha a sokaság átlagát μ-vel, szórásnégyzetét σ^2-tel jelöljük, akkor a $\xi_1, \xi_2, \ldots, \xi_n$ mintára igaz, hogy $E(\xi_i) = \mu$ és $Var(\xi_i) = \sigma^2$. Mindennek figyelembevételével határozzuk meg a mintaátlag várhatóértékét és szórásnégyzetét (103. megjegyzés):

$$E(\bar{\xi}) = E\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n}\sum_{i=1}^{n}E(\xi_i) = \frac{1}{n}n\mu = \mu \ , \qquad (91)$$

$$Var(\bar{\xi}) = Var\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}Var(\xi_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \ . \qquad (92)$$

Látható, hogy a mintaátlag várhatóértéke megegyezik az alapsokaság megfelelő jellemzőjével, az elméleti átlaggal, tehát a **mintaátlag** mint becslés **torzítatlan** (vö. (85)). Az is megfigyelhető, hogy $n \to \infty$ esetén a **mintaátlag varianciája**, illetve az ebből vont négyzetgyök, azaz a mintaátlag **standard hibája** is (vö. (88)) 0-hoz tart:

$$SE(\bar{\xi}) = \frac{\sigma}{\sqrt{n}} \ . \qquad (93)$$

E két feltétel együttes teljesülése viszont azt jelenti (mint azt a 14.1. részben az aszimptotikus torzítatlan esetben már említettük), hogy a **mintaátlag** mint becslés nemcsak torzítatlan, hanem **konzisztens** is (104. megjegyzés).

A másik igen gyakran használt becslő függvény a mintaszórás vagy ***tapasztalati szórás*** (vö. (42)), – a fent már említett kétféle jelöléssel – (s_x) illetve (s_ξ), amellyel az alapsokaság szórását (σ -t), azaz az ***elméleti szórást*** becsüljük. Itt azonban azzal kell szembesülnünk, hogy míg az **elméleti szórás** az **elméleti átlagtól** való eltéréseket méri, a **tapasztalati szórás** – az elméleti átlag ismeretének hiányában – a hasonló eltéréseket **„csak”** a **tapasztalati átlagtól** tudja mérni. Elvileg azonban ugyanabból a konkrét mintából $(x_1, x_2, \ldots, x_n)$ egy olyan szórásnégyzet is meghatározható, amelyik az elméleti átlagtól való eltéréseket méri (105. megjegyzés). A kétféle szórásnégyzetre az átlag minimum tulajdonsága alapján (vö. 10.0. rész második mintafeladat) fennáll a következő egyenlőtlenség:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \geq \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2 \ . \qquad (94)$$

Ez az összefüggés viszont az összes lehetséges mintára igaz. Így a minimum tulajdonságból az is következik, hogy a tapasztalati szórásnégyzet mint valószínűségi változó (s_ξ^2) minden esetben szisztematikusan kisebb vagy egyenlő, mint a becsülendő elméleti szórásnégyzet (σ^2), így várhatóértéke nem lehet egyenlő az elméleti szórásnégyzettel:

$$E(s_\xi^2) \neq \sigma^2 \ \text{(vö. (85))}. \qquad (95)$$

Ezek szerint a ***tapasztalati szórásnégyzet*** csak **torzított becslést** ad az ***elméleti szórásnégyzetre***.

A kérdés az, hogy miként lehet korrigálni ezt a torzítást. A következő mintafeladatban – a 14.1. rész mintafeladatában alkalmazott lépésekhez hasonlóan – belátjuk, hogy a tapasztalati és az elméleti szórásnégyzet között fennáll az alábbi összefüggés:

$$\sigma^2 = s_\xi^{\,2} + \frac{\sigma^2}{n}\,, \tag{96}$$

Mintafeladat

Igazoljuk a (96) összefüggést.

Megoldás: Első lépésként írjuk fel a megfelelő szórásnégyzeteket (vö. (92)).

1. a mintaelemek varianciája: $Var(\xi_i) = E\left[(\xi_i - \mu)^2\right] = \sigma^2$,
2. a mintaátlag varianciája: $Var(\bar{\xi}) = E\left[(\bar{\xi} - \mu)^2\right] = \frac{\sigma^2}{n}$,
3. a mintaelemek tapasztalati szórásnégyzete: $E\left[(\xi_i - \bar{\xi})^2\right] = s_\xi^{\,2}$,

ahol felhasználtuk, hogy $E(\xi) = E(\xi_i) = E(\bar{\xi}) = \mu$ (vö. (91)). Ezután az 1. kifejezést alakítsuk át úgy, hogy a $(\xi_i - \mu)$ különbségből vonjuk le a mintaátlagot majd rögtön adjuk is hozzá (ami összességében természetesen nem okoz változást). Így felírható a következő egyenlőség:

$$E\left[(\xi_i - \mu)^2\right] = E\left[\left((\xi_i - \bar{\xi}) + (\bar{\xi} - \mu)\right)^2\right] .$$

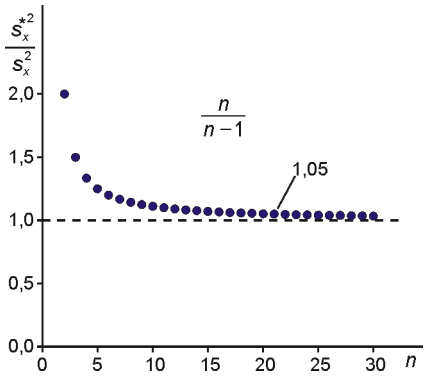
A jobb oldalon a négyzetre emelés elvégzése után csak a négyzetes tagok maradnak meg (vö. 14.1. rész mintafeladat), hiszen a függetlenség és az átlag $E(\xi_i - \bar{\xi}) = 0$ tulajdonsága miatt (vö. 62. megjegyzés) a vegyes szorzat 0-t ad eredményül. A kapott várható értékeket a fenti 3. illetve 2. kifejezésekkel összevetve éppen a bizonyítandó egyenlőséget kapjuk.

Ebből kisebb átalakítással a tapasztalati szórásnégyzet is kifejezhető:

$$s_\xi^{\,2} = \frac{n-1}{n}\sigma^2\,, \tag{97}$$

illetve a (41) összefüggés alapján egy konkrét minta esetén egy új mennyiség, a korrigált tapasztalati szórásnégyzet is bevezethető:

$$s_x^{*2} \equiv s_x^{\,2}\,\frac{n}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 . \tag{98}$$



106. ábra
A korrigált és a korrigálatlan tapasztalati szórás hányadosa a minta elemszámának függvényében. $n > 21$ esetén az eltérés már 5%-nál kisebb.

A (97) összefüggés azt mutatja, hogy bár a ***tapasztalati szórásnégyzet*** torzított becslése az elméleti szórásnégyzetnek, **aszimptotikusan torzítatlan**, hiszen $n \to \infty$ esetén a σ^2 előtt álló hányados 1-hez tart. A (98) összefüggésből pedig az derül ki, hogy a korrigált tapasztalati szórásnégyzet eleve torzítatlan, ugyanis:

$$E(s_\xi^{*2}) = \sigma^2 \quad \text{(vö. (97) és (85))}. \tag{99}$$

A két szórásnégyzet viszonyát jól szemlélteti a 106. ábra, ahol a kettő arányát a minta elemszámának függvényében tüntettük fel (107. megjegyzés).

107. megjegyzés
A variancia korrigálatlan és korrigált becslésének a viszonya azért is érdekes, mert a korrigált becslés torzítatlan ugyan, de mivel szisztematikusan mindig nagyobb a korrigálatlan becslésnél, ezért a véletlen ingadozások miatt a standard hibája is nagyobb. Ennek ellenére a gyakorlatban mégis inkább a korrigált becslést használják.

A korrigált szórás kapcsán egy új, a későbbiek során még sokszor előforduló mennyiség, a szabadságfok (ν) bevezetésére nyílik lehetőség. Ez pedig a (98) összefüggésben az $(n-1)$ kifejezés. A szabadságfok számszerű jellemzők becslésével kapcsolatos statisztikai fogalom, amely szoros összefüggésben van a minta elemszámával, n-nel, de láthatóan nem mindig egyenlő vele. A számítások kezdetekor az n adatból álló minta szabadságfoka is n. Ha egy mintából valamely jellemzőt úgy kell becsülnünk, hogy ahhoz ugyanebből a mintából már előzetesen meghatározott jellemzőket fel kell használnunk, akkor annyit kell levonnunk az eredetileg n szabadságfokból, ahány korábbi (egymástól független) jellemzőt a becslés közben felhasználtunk. Mivel a tapasztalati szórás becslésénél az n adaton kívül az ugyanabból a mintából már meghatározott átlagot is fel kell használnunk, ezért a tapasztalati szórás szabadságfoka $\nu = n - 1$.

Általánosan egy statisztika szabadságfokát úgy definiáljuk, hogy a minta elemszámából (n) levonjuk az adott statisztika kiszámításhoz szükséges azon számszerű (egymástól független) jellemzők számát (k), amelyeket ugyanebből a mintából már meghatároztunk. Bonyolultabb esetekben a szabadságfokot külön képlet adja meg.

14.3. $\bar{\xi}$ és s_ξ^2 eloszlása normális eloszlású sokaság esetén

Sok esetben úgy járunk el, hogy feltesszük, hogy a sokaság eloszlása jól közelíthető egy nevezetes eloszlással, például normális eloszlással, majd ezután ennek jellemzőit használjuk az egyébként ismeretlen sokaság jellemzésére. A statisztikában a normális eloszlás azért játszik kulcsfontosságú szerepet, mert a vizsgált változók nagyon gyakran – legalábbis közelítőleg – normális eloszlásúak (vö. 77. megjegyzés), valamint azért, mert sok statisztikai eljárás csak normális eloszlású változókkal működik helyesen (108. megjegyzés).

108. megjegyzés
Az ilyen közelítés akkor is jól használható, ha a változóra vonatkozó matematikai feltételek szigorúan nem teljesülnek. Tudjuk például, hogy a normális eloszlás negatív számokon is értelmezve van, mégsem vonjuk kétségbe használhatóságát a kizárólag pozitív számokon értelmezhető testmagasságra. Ebben az esetben egyszerű a magyarázat: a gyakorlatban a nagyon kis valószínűségek a 0 valószínűségtől nem különböztethetők meg.

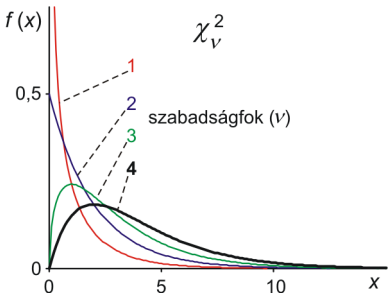
Mivel esetünkben a minta – a 13.2. rész elején elmondottak alapján – független normális eloszlású valószínűségi változók sorozata, ezért **a mintaátlag** ($\bar{\xi}$), ami a mintaelemek összegéből adódik, **maga is normális eloszlású**. Ha $N(\mu, \sigma)$-val jelöljük a sokaság eloszlását, akkor a tapasztalati átlag eloszlása a (91) és (93) összefüggések alapján:

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right). \tag{100}$$

Kevésbé egyszerű a minta szórásnégyzet (s_ξ^2) eloszlásának meghatározása. Ennek érdekében először be kell vezetnünk egy új eloszlást, amely a normális eloszlásból származtatható. Legyen $\xi_1, \xi_2, \ldots, \xi_n$ n darab független standard normális eloszlású ($N(0,1)$, vö. 12.1. rész) valószínűségi változó. Ekkor az

$$\eta^2 = \xi_1^{\,2} + \xi_2^{\,2} + \ldots + \xi_n^{\,2} \tag{101}$$

valószínűségi változó eloszlása a $\nu = n$ szabadságfokú χ^2-eloszlás (109. ábra), négyzetgyöke pedig a χ-eloszlás.



109. ábra
Az 1, 2, 3 és 4 szabadságfokú χ^2-eloszlás sűrűségfüggvényének grafikus szemléltetése. $E(\xi) = \nu$, $Var(\xi) = 2\nu$ ($\nu \geq 2$); $Mo = \nu - 2$ ($\nu \geq 3$).

Amennyiben a $\xi_1, \xi_2, \ldots, \xi_n$ n elemű minta $N(\mu, \sigma)$ eloszlású, akkor a μ-vel eltolt $\zeta_i = \xi_i - \mu$ valószínűségi változók $N(0, \sigma)$ eloszlásúak lesznek (vö. 12.1. rész). Mivel az eltolási transzformáció a szórásnégyzetet nem befolyásolja, ezért a transzformált minta szórásnégyzetének eloszlása az eredetivel megegyező lesz. Egy további nyújtási, illetve zsugorítási transzformáció elvégzésével az új változó $N(0;1)$ eloszlásúvá alakítható. Mindezek ismeretében, ha nem is közvetlenül a szórásnégyzetnek (s_ξ^2 -nek), de egy viszonylag egyszerű transzformáltjának eloszlása meghatározható. Mivel a szórásnégyzet (41) összefüggésében a ζ_i-hez hasonló különbségek négyzetének összege szerepel, így belátható (a bizonyítást itt is mellőzzük), hogy az

$$\frac{n}{\sigma^2} s_\xi^{\,2} = \frac{n-1}{\sigma^2} s_\xi^{*2} \tag{102}$$

transzformált szórásnégyzet (illetve korrigált szórásnégyzet, vö. (98)) $\nu = n-1$ szabadságfokú χ^2-eloszlású valószínűségi változó (110. megjegyzés).

110. megjegyzés
Mint azt a szabadságfok bevezetésénél (14.2. rész) láthattuk, a szórásnégyzet szabadságfoka eleve $\nu = n-1$, így nem kell meglepődnünk ezen az eredményen sem.

14.4. Egy valószínűség becslése

Példaként becsüljük meg annak a valószínűségét, $P(A)$-t, hogy Magyarországon egy véletlenszerűen kiválasztott egyénnek „A” vércsoportú vére van. Ebben az esetben a sokaság a Magyarországon élő emberek összessége a megfelelő vércsoport adatukkal együtt. A becsléshez ebből választunk ki egy n elemű mintát. Tegyük fel, hogy a vércsoport meghatározások után k esetben kapunk „A” vércsoportú eredményt és $(n-k)$ esetben egyebet. Tudjuk, hogy az ehhez hasonló ismételt alternatívák modellezésére a binomiális eloszlás használható. Amennyiben a konkrét k érték helyett a neki megfelelő κ valószínűségi változót tekintjük, akkor ennek várhatóértéke és varianciája (vö. (48), (49)):

$$E(\kappa) = nP(A)\ , \qquad Var(\kappa) = nP(A)\big(1 - P(A)\big)\,. \tag{103}$$

Ennek alapján a k/n relatív gyakoriságra:

$$E\left(\frac{\kappa}{n}\right) = \frac{1}{n} E(\kappa) = P(A)\,, \tag{104}$$

$$Var\left(\frac{\kappa}{n}\right) = \frac{1}{n^2} Var(\kappa) = \frac{1}{n} P(A)\big(1 - P(A)\big) \leq \frac{1}{4n}\,, \tag{105}$$

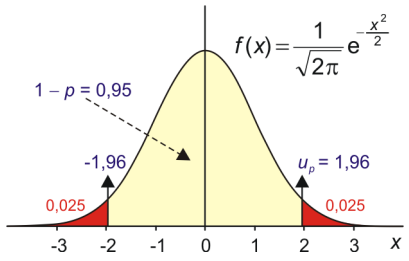
111. ábra
Példa „pontbecslés”-re: a körözött személy körülbelül 175 cm magas;
és „intervallum becslés”-re: a körözött személy mintegy 170-180 cm magas.

112. megjegyzés
A megoldás az lehet, hogy egyetlen becsült érték helyett megadunk egy olyan intervallumot, amely az adott elméleti értéket nagy valószínűséggel tartalmazza. Problémánk most már csak az, hogy hol vonjuk meg az intervallum határait.

113. megjegyzés
Az intervallum határainak megvonásánál a következő szempontokat kell figyelembe vennünk.

Ha nagyon tág határokat adunk meg, akkor várható ugyan (nagy a valószínűsége), hogy ezek a „valódi” értéket (ω-t) közrefogják, de a nagyon tág határoknak többnyire kevés a gyakorlati haszna.

Ha ezzel szemben szűkítjük a határokat, fokozottan növekszik annak a kockázata, hogy azok már nem fogják közre az adott jellemző értékét (ω-t), tehát következtetésünk megbízhatósága csökken.



114. ábra
A konfidencia határok meghatározásának grafikus szemléltetése standard normális eloszlású valószínűségi változó esetén (p = 0,05, u_p = 1,96, 95%-os konfidencia szint) (115. megjegyzés).

115. megjegyzés
A gyakorlati alkalmazások során, mivel 1,96 ≈ 2 kevésbé akkurátusan azt szokás mondani, hogy a 95%-os konfidencia szint esetén $u_p = 2$.

ahol a 13.2. rész mintafeladatához hasonlóan írtuk fel az egyenlőtlenséget. E két összefüggés, (104) és (105) azt jelenti, hogy a k/n relatív gyakoriság torzítatlan és konzisztens becslése a $P(A)$ valószínűségnek (vö. 14.0. rész).

Mivel a valószínűség általában nem ismert, a gyakorlatban a relatív gyakoriságot használjuk helyette. Fontos hangsúlyozni azonban, hogy a minta számszerű jellemzői eltérhetnek a sokaság jellemzőitől. A feltételes mód azért indokolt, mert a sokaság jellemzőit csak kivételes esetekben ismerjük, így a legtöbb esetben nem tudjuk, hogy a jellemzők valójában eltérnek-e vagy sem.

Az eddigi becslések mindegyike pontbecslés volt (111. ábra), ami azt jelenti, hogy a **becsült** tapasztalati **érték** és a „valódi” **elméleti érték között általában van egy véletlen eltérés, amelynek nagyságáról semmit sem tudunk mondani**. Amit ilyen esetben hiányolunk, az a megbízhatóság vagy bizonyosság, idegen szóval konfidencia. Ezt szeretnénk a következőkben számszerűen is jellemezni (112. megjegyzés).

14.5. Konfidencia intervallum (tartomány)

Legyen ω az alapsokaság egy számszerű jellemzője és $\xi_1, \xi_2, \ldots, \xi_n$ egy n elemű minta. A legtöbb esetben lehetőség nyílik olyan w_1 és w_2 statisztikák konstruálására is, hogy a (w_1, w_2) intervallum – előre megadott – nagy valószínűséggel **lefedje** az ω jellemzőt, azaz teljesüljön a

$$P(w_1 \le \omega \le w_2) = 1 - p \qquad (106)$$

egyenlőség, ahol p tetszőlegesen (kicsire) megválasztható valószínűség, amelytől a w_1 és w_2 statisztikák függenek (113. megjegyzés). Ilyenkor ω-t egy intervallummal becsüljük, ezért a becslésnek ezt a módját intervallum becslésnek nevezzük (111. ábra).

A (w_1, w_2) véletlen helyzetű intervallumot konfidencia intervallumnak, az $1-p$ valószínűséget, ami a lefedettség bizonyosságának mértékét fejezi ki, és amit legtöbbször %-ban adunk meg, konfidencia szintnek, míg az intervallum kezdő és végpontját konfidencia határoknak nevezzük.

Néhány speciális esetben a **normális eloszlású sokaság μ és σ paramétereire vonatkozóan a konfidencia intervallum** elég egyszerűen meghatározható .

1. Tegyük fel, hogy ***σ ismert***, (ami a gyakorlatban elég ritkán fordul elő, de elméleti jelentősége miatt foglalkoznunk kell vele) és ***μ lehetséges konfidencia intevallumaira vagyunk kíváncsiak***. Az n elemű minta átlagának standardizálása (vö. (71)) után a transzformált u valószínűségi változó már $N(0;1)$ eloszlású:

$$u = \frac{\overline{\xi} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\overline{\xi} - \mu)}{\sigma} . \qquad (107)$$

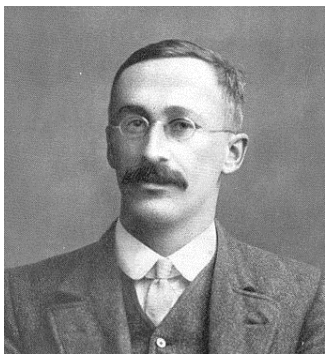
Erre vonatkozóan a (106) összefüggés alapján megadható a $(-u_p, u_p)$ konfidencia intervallum:

$$P(|u| \le u_p) = 2\int_0^{u_p} f(x)\,\mathrm{d}x = 2\frac{1}{\sqrt{2\pi}}\int_0^{u_p} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = 1 - p \ , \qquad (108)$$

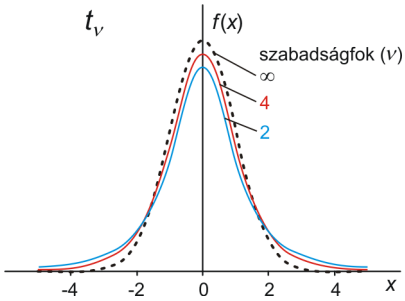
ahol az integrállal a sűrűségfüggvény ismeretében a konfidencia határok közötti görbe alatti területet számoljuk ki (114. ábra). u_p-t aszerint tudjuk változtatni, hogy az a kívánt $(1-p)\cdot 100\%$ konfidencia szintnek megfeleljen. A visszatranszformálás elvégzése után (vö. (107)) most már μ-re vonatkozóan is megkaphatjuk a konfidencia intervallumot:

$$\left(\overline{\xi} - u_p\frac{\sigma}{\sqrt{n}},\ \overline{\xi} + u_p\frac{\sigma}{\sqrt{n}}\right). \qquad (109)$$

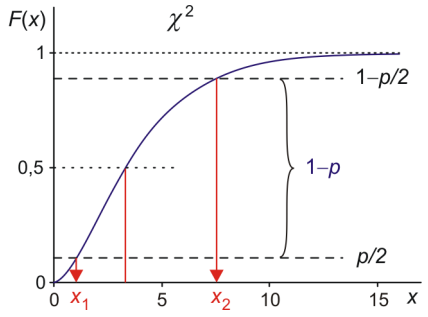
Amennyiben p = 0,05 (95%-os a konfidencia szint), akkor ezt úgy kell érteni, hogy ha például 20-szor veszünk mintát és számolunk belől konfidencia intervallumot, akkor nagyjából 19-szer kapunk olyat, amely lefedi a becsülendő jellemzőt, de egyszer olyat is kaphatunk, ami nem fedi le (vö. 91. megjegyzés).



William Sealy Gosset (1876-1937) angol vegyészmérnök és statisztikus, a *t*-eloszlás kidolgozója, melyet **Student** álnéven publikált.



116. ábra
A 2, 4 és ∞ szabadságfokú *t*-eloszlás sűrűségfüggvényének grafikus szemléltetése. $E(\xi) = 0$, $(\nu > 1)$. $Me = Mo = 0$; $Var(\xi) = \nu/(\nu-2)$, $(\nu > 2)$.



117. ábra
A konfidencia határok kijelölésének általános módszere a 4 szabadságfokú χ^2-eloszlás eloszlásfüggvényén bemutatva.

Láthatjuk, hogy amennyiben a konfidencia intervallumokat a transzformált eloszlásban interkvantilis terjedelemként kezeljük, akkor teljesen mindegy, hogy az eloszlás szimmetrikus vagy nem, hiszen $(1-p)100\%$ konfidencia szint esetén azok az x_1, x_2 értékek lesznek a konfidencia határok, amelyekre $F(x_1) = p/2$, és $F(x_2) = 1-p/2$ (118. megjegyzés).

118. megjegyzés
Figyeljünk arra, hogy a táblázatokban és a számítógépes programokban is az az elfogadott konvenció, hogy a fenti ábrához képest éppen fordítva adják meg az értékeket. Nevezetesen az x_1 helynek felel meg a $\chi^2_{1-p/2}$, és az x_2-nek a $\chi^2_{p/2}$ érték.

Ha – mint az leggyakrabban lenni szokott – egyetlen mintából számolunk konfidencia intervallumot, akkor ezzel **az egyedi intervallummal kapcsolatban valószínűségi kijelentésnek már nincs helye**. Persze azért **bízunk** benne, hogy ez az egy lefedi a becsülendő jellemzőt, hiszen tudjuk, hogy a hibalehetőség csak 5%. Ezért használjuk a valószínűség helyett a konfidencia kifejezést.

2. Tegyük fel, hogy ***σ nem ismert***, és ebben az esetben is ***μ lehetséges konfidencia intevallumait szeretnénk megadni***. Tekintsük az alábbi statisztikát (t), amelyet egy n elemű mintából határoztunk meg:

$$t = \frac{\sqrt{n}(\bar{\xi}-\mu)}{s_\xi^*} = \frac{\frac{\sqrt{n-1}}{\sigma}\sqrt{n}(\bar{\xi}-\mu)}{\frac{\sqrt{n-1}}{\sigma}s_\xi^*} = \frac{u\sqrt{n-1}}{\frac{\sqrt{n-1}}{\sigma}s_\xi^*}. \qquad (110)$$

Ez a (107) összefüggéstől annyiban tér el, hogy itt σ helyett a mintából meghatározható (tapasztalati) korrigált szórás (s_ξ^*) szerepel, majd a számlálót és a nevezőt is megszoroztuk $\sqrt{n-1}/\sigma$-val. A (107) összeüggéssel való összevetés után láthatjuk, hogy a számláló az $N(0;1)$ eloszlású u valószínűségi változó $\sqrt{n-1}$-szerese. A nevező pedig a (102) összefüggés jobb oldalának négyzetgyöke, tehát $\nu = n-1$ szabadságfokú χ-eloszlású valószínűségi változó. Mivel a számláló és a nevező függetlenek egymástól, belátható (amit most sem részletezünk), hogy t, $\nu = n-1$ szabadságfokú Student–, vagy *t*–eloszlású valószínűségi változó (116. ábra). Végső soron ez az eloszlás is a normális eloszlásból származtatható, végtelen szabadságfok esetén egzaktul visszaadja az $N(0;1)$ eloszlást, egyébként pedig sűrűségfüggvényének konkrét alakja a szabadságfoktól is függ.

Ugyanúgy, mint az előző esetben erre vonatkozóan is megadható a $(-t_p, t_p)$ konfidencia intervallum:

$$P(|t| \le t_p) = 2\int_0^{t_p} f_{n-1}(x)\,\mathrm{d}x = 1-p \quad \text{(vö. (108))}, \qquad (111)$$

ahol f_{n-1} a minta elemszámánál eggyel kisebb szabadságfokú t-eloszlás sűrűségfüggvénye. A visszatranszformálás eredményeként megkapjuk μ-re vonatkozóan is a konfidencia intervallumot:

$$\left(\bar{\xi} - t_p\frac{s_\xi^*}{\sqrt{n}},\ \bar{\xi} + t_p\frac{s_\xi^*}{\sqrt{n}}\right). \qquad (112)$$

3. Az előző két pontban (ismert és ismeretlen σ esetén) μ-re vonatkozóan adtunk meg szimmetrikus konfidencia intervallumokat, σ -ra vonatkozóan azonban másként kell eljárnunk. A (102) összefüggés alapján (μ ismeretétől függetlenül) tudjuk, hogy az ns_ξ^2/σ^2 kifejezés $\nu = n-1$ szabadságfokú χ^2-eloszlású valószínűségi változó. Míg az előző két pontban a transzformált változó (u, illetve t) eloszlása szimmetrikus volt, ez nem áll fenn a χ^2-eloszlásra. Így ebben az esetben a konfidencia intervallumot sem tudjuk szimmetrikusan kijelölni (117. ábra). Eléggé kézenfekvően adódik azonban egy olyan választás, amelyre teljesül az alábbi feltétel:

$$P\left(\chi^2_{1-\frac{p}{2}} \le \frac{ns_\xi^2}{\sigma^2} \le \chi^2_{\frac{p}{2}}\right) = 1-p \text{, vagy } P\left(\frac{ns_\xi^2}{\chi^2_{\frac{p}{2}}} \le \sigma^2 \le \frac{ns_\xi^2}{\chi^2_{1-\frac{p}{2}}}\right) = 1-p, \qquad (113)$$

amelyek ekvivalensek egymással, és ahol χ^2_p (az eloszlásfüggvényeknél megszokottól eltérően) azt jelenti, hogy az eloszlásban az ennél ***nagyobb*** értékek előfordulásának valószínűség p (118. megjegyzés). Így σ -ra vonatkozóan az $(1-p)100\%$ konfidencia szintnek megfelelő konfidencia intervallum (négyzetgyököt vonva):

$$\left(\frac{\sqrt{n}s_\xi}{\chi_{\frac{p}{2}}}, \frac{\sqrt{n}s_\xi}{\chi_{1-\frac{p}{2}}}\right). \qquad (114)$$

Mintafeladat

Adjunk becsléseket az egészséges ember normális pulzusszámára vonatkozóan az alábbi 20 elemű minta alapján (az értékek mind (1/perc)-ben értendők).

66	56	89	63	66	69	71	68	58	69	78	66	64	84	74	76	69	77	74	76

Megoldás: **1.** Első lépésként számítsuk ki a mintaátlagot (vö. (27)): $\bar{x} = 70{,}6$ (1 tizedes jegyre kerekítve);

majd adjuk meg a standard hibáját (vö. (93)): $SE(\bar{x}) = 1{,}8$ (σ becslésére a korrigált tapasztalati szórásnégyzet négyzetgyökét használtuk ($s_x^* = 8{,}1$) vö. (98)). Ezzel megadtuk a normális pulzusszám **pontbecslését** (119. megjegyzés). (Ennél a becslésnél fontos észrevétel az is, hogy amennyiben a korrigált tapasztalati szórásnégyzet helyett a korrigálatlannal számolnánk ($s_x = 7{,}9$), az eredmény az 1 tizedes jegyre való kerekítés miatt nem változna meg.)

Ha **feltesszük, hogy a változó normális eloszlású**, amelynek paraméterei μ és σ, akkor az előzőek alapján **konfidencia intervallumokat** is meg tudunk határozni. A **konfidencia szint mindegyik esetben legyen 95%-os**.

2. Amennyiben σ ismeretlen, a μ-re vonatkozó konfidencia intervallum meghatározásához $\bar{x}$ és $SE(\bar{x})$ -on kívül szükségünk van a $\nu = n - 1 = 19$ szabadságfokú $t_{0,05}$ értékre is, amit vagy számítógépes program vagy táblázat segítségével határozunk meg: $t_{0,05} = 2.09$ (2 tizedes jegyre kerekítve). Mindezek ismeretében a (112) összefüggés szerint a konfidencia intervallum: (66,8; 74,4) (ismét 1 tizedes jegyre kerekítve). (A statisztikai táblázatok az utolsó oldalon találhatók.)

3. Tételezzük fel ezután azt az esetet, hogy σ ismert és értéke ugyanakkora, mint a korrigált tapasztalati szórás. Ilyenkor a $t_{0,05}$ érték helyett, az $u_{0,05}$ értékre van szükségünk, amely a 114. ábrán már szerepelt: $u_{0,05} = 1.96$. (Ez az érték a t táblázat $\nu = \infty$ szabadságfokú sorából is kiolvasható.) Most a (109) összefüggés felhasználásával kapjuk meg a keresett konfidencia intervallumot: (67,1; 74,1).

4. Végül határozzuk meg a konfidencia intervallumot σ-ra vonatkozóan is a (113) és a (114) összefüggések alapján. $n = 20$; ($\nu = 19$); $s_x^2 = 62{,}3$; $\chi^2_{0,025} = 32{,}85$; $\chi^2_{0,975} = 8{,}91$ ismeretében: (6,2; 11,8).

119. megjegyzés
Van olyan eset, amikor a szórás helyett a relatív szórást adjuk meg, ami a szórás és az átlag hányadosa. (Ezt természetesen csak olyan változó esetében tehetjük meg, ahol az arányok is értelmezhetők (vö. 13.4. rész).) Ilyen megadási formát akkor érdemes választanunk, ha azt akarjuk kifejezni, hogy mekkora a megfigyelt értékek bizonytalansága. Nem mindegy ugyanis, hogy például a ±10 egységnyi ingadozás 50 vagy 5000 körül történik. Az első esetben jelentősnek mondható a 20%-os ingadozás, míg a második esetben a 0,2%-os valószínűleg elhanyagolható.

A mintafeladatból kiderül, hogy a pontbecslés mellett – amennyiben van rá lehetőség – igen hasznos a konfidencia intervallum megadása is. Ennek figyelembevételével a fenti eredmények (70,6 ± 3,8); (70,6 ± 3,5); illetve (8,1 – 1,9 és 8,1 + 3,7). Ebből kiderül, hogy míg (az adott feltételek mellett) a mintaátlag szimmetrikus eloszlású, hiszen a konfidencia határok a pontbecslés körül szimmetrikusan helyezkednek el, addig a tapasztalati szórás nem az. **A konfidencia intervallum** tehát **kifejezheti az eloszlás ferdeségét is**. Az „egyetlen” számként megadott standard hiba viszont erre nem képes. Ennek látszólag ellent mond a pontbecslésnél is szokásos 70,6 ± 1,8 (*SE*) vagy 70,6 ± 8,1 (*SD*) (vö. 100. megjegyzés) megadási forma, ami azért inkább csak szimmetrikus eloszlásnál használatos. Ilyenkor azonban nem szabad elmulasztani annak jelzését, hogy eredményünkben a becsült jellemző standard hibáját (*SE*), vagy mintabeli szórását (*SD*) tüntetjük fel.

A becslések megadásánál fontos, hogy egyértelmű legyen az eredmény. **Intervallum becslés esetén a konfidencia szint megadása elengedhetetlen**. Emellett **mindkét esetben célszerű megadni a minta elemszámát** is. A minta elemszámának ismerete azért fontos, mert n növelésével egyrészt csökkenthetjük a pontbecslés hibáját, másrészt anélkül szűkíthetjük a konfidencia intervallumot, hogy a konfidencia szintet csökkentenünk kellene (vö. (109), (112), valamint a 113. megjegyzés).

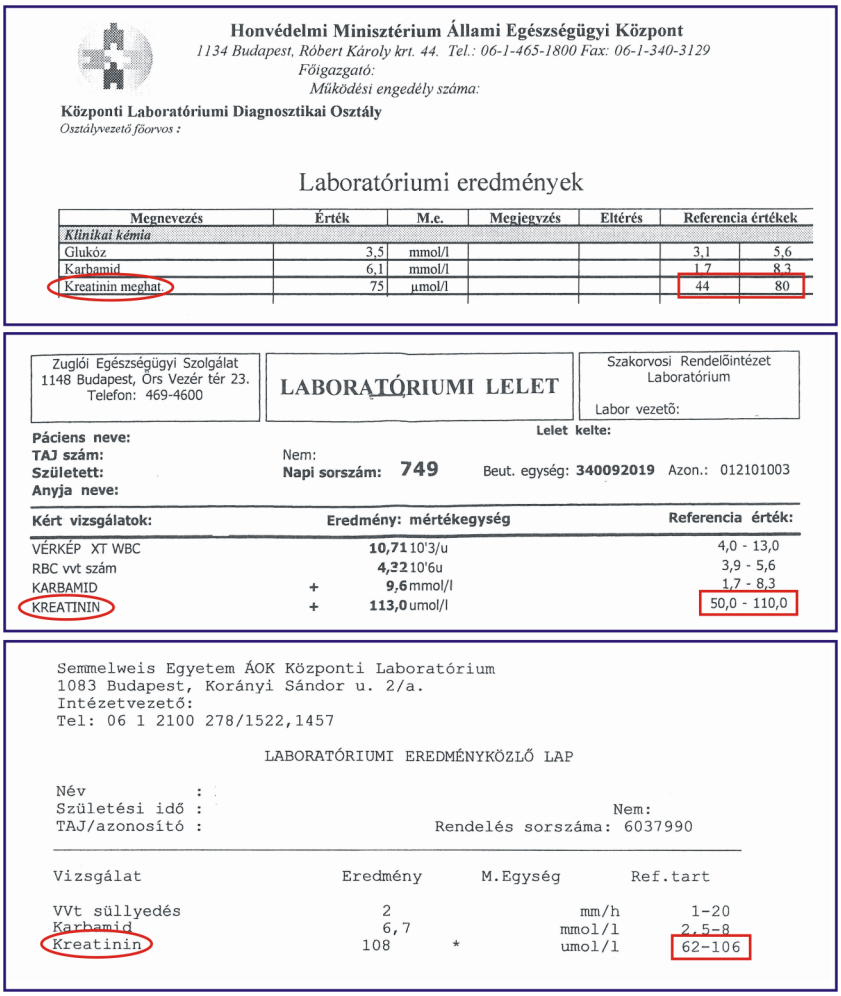
14.6. Referencia tartomány (intervallum)

120. megjegyzés
A „referencia tartomány” helyett gyakran használják a „normál tartomány” vagy „normál érték” kifejezéseket is. Szabatosabb és pontosabb a „referencia tartomány”" használata, mert amíg a referencia sokaság jól meghatározható, addig a „normál” sokaság igen nehezen. Ha egy vizsgálat eredményét különböző populációkon mért eredményekkel vetjük össze, könnyen kiderül, hogy ugyanaz az eredmény az egyik esetben „normál”-nak felel meg, míg a másikban nem szükségképpen az. Tipikus példa erre a várandós nőknél mérhető kémiai jellemzők eltérései, ezért számukra több vizsgálatnál is speciális referencia tartományt kell készíteni.

Laboratóriumi leleteken gyakran olvasható ez a kifejezés és arra szolgál, hogy segítse az orvost annak eldöntésében, hogy egy diagnosztikai adat normális-e vagy sem. A „jó” referencia tartomány meghatározása sok esetben egyáltalán nem könnyű feladat. Általában úgy szokás eljárni, hogy első lépésben meg kell határozni azt az alapsokaságot, amelyre majd a referencia tartomány vonatkoztatható. Ilyen sokaság lehet például az adott területen élő 20 és 40 év közötti, „egészséges” emberek valamilyen jellemzőjének összessége. Ezután egy nagy elemszámú mintán (például $n \approx 300$, de ennek meghatározása külön feladat) elvégzik a vizsgálatot, majd a nyert adatokból – a konfidencia intertvallum meghatározásánál használt általános módszerhez hasonlóan (vö. 117. ábra) – meghatározzák az alsó 2,5, valamint a felső 97,5 percentiliseket (vö. 10.0. rész) és ezeket tekintik a referencia tartomány határainak (120. megjegyzés).

121. megjegyzés
A referencia tartományok használatában még a torzítás sem jelent problémát, mert amennyiben minden adat szisztematikusan el van tolódva, akkor ugyanennyivel a referencia tartomány is eltolódik. Ezt esetenként meg is figyelhetjük, ha különböző laboratóriumokban elvégzett vizsgálatok eredményét hasonlítjuk össze. A referencia tartományok ugyanarra a változóra nézve kissé eltérhetnek egymástól. Ennek oka például az lehet, hogy az alkalmazott mérési módszerek, illetve a mérőberendezések különbözőek (122. ábra).

Amennyiben az adatok eloszlása normális eloszlással közelíthető, akkor **a mintaátlagból és -szórásból meghatározott $\bar{x} \pm u_{0,05} s_x^*$ intervallum** lesz a referencia tartomány, amelyben a minta elemeinek 95%-a található ($u_{0,05} \approx 2$, vö. 114. ábra, 115. megjegyzés). Ezt a gyakorlatban úgy lehet felhasználni, hogy amennyiben egy laboratóriumi érték a referencia tartományon belüli, akkor arról 95%-os bizonyossággal azt mondhatjuk, hogy akár az „egészséges” állapotnak is megfelelhet, tehát önmagában nem utal kóros elváltozásra. A bizonyosság a 14.5. részben leírtakhoz hasonlóan értendő (121. megjegyzés).



Honvédelmi Minisztérium Állami Egészségügyi Központ
1134 Budapest, Róbert Károly krt. 44. Tel.: 06-1-465-1800 Fax: 06-1-340-3129
Főigazgató:
Működési engedély száma:

Központi Laboratóriumi Diagnosztikai Osztály
Osztályvezető főorvos :

Laboratóriumi eredmények

Megnevezés	Érték	M.e.	Megjegyzés	Eltérés	Referencia értékek	
Klinikai kémia						
Glukóz	3,5	mmol/l			3,1	5,6
Karbamid	6,1	mmol/l			1,7	8,3
Kreatinin meghat.	75	µmol/l			44	80

Zuglói Egészségügyi Szolgálat
1148 Budapest, Örs Vezér tér 23.
Telefon: 469-4600

LABORATÓRIUMI LELET

Szakorvosi Rendelőintézet
Laboratórium
Labor vezető:

Páciens neve:
TAJ szám:
Született:
Anyja neve:
Nem:
Napi sorszám: **749**
Lelet kelte:
Beut. egység: **340092019** Azon.: 012101003

Kért vizsgálatok:		Eredmény: mértékegység	Referencia érték:
VÉRKÉP XT WBC		**10,71** 10'3/u	4,0 - 13,0
RBC vvt szám		**4,22** 10'6u	3,9 - 5,6
KARBAMID	+	**9,6** mmol/l	1,7 - 8,3
KREATININ	+	**113,0** umol/l	50,0 - 110,0

Semmelweis Egyetem ÁOK Központi Laboratórium
1083 Budapest, Korányi Sándor u. 2/a.
Intézetvezető:
Tel: 06 1 2100 278/1522,1457

LABORATÓRIUMI EREDMÉNYKÖZLŐ LAP

Név :
Születési idő :
TAJ/azonosító :
Nem:
Rendelés sorszáma: 6037990

Vizsgálat	Eredmény		M.Egység	Ref.tart
VVt süllyedés	2		mm/h	1-20
Karbamid	6,7		mmol/l	2,5-8
Kreatinin	108	*	umol/l	62-106

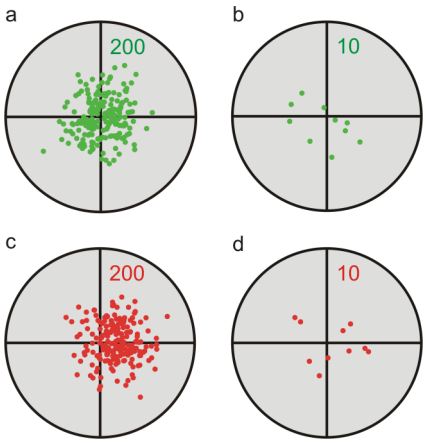
122. ábra
Különböző laboratóriumokban készült leletek részletei, amelyeken jelentősen eltérő referencia tartományok figyelhetők meg (121. megjegyzés).

15.0. Statisztikai hipotézisvizsgálat (feltevésvizsgálat)

A ***becslések*** témakörében azt a fontos észrevételt alkalmaztuk, hogy egy reprezentatív ***minta és az a sokaság***, amelyikből ez a minta származik, ***hasonlóságot mutat*** (vö. 13.2. rész). Ezt arra használtuk, hogy a sokaságra vonatkozó „Mennyi? Mekkora? stb.” típusú kérdésekre – a sokaság teljes ismerete nélkül – választ tudtunk adni.

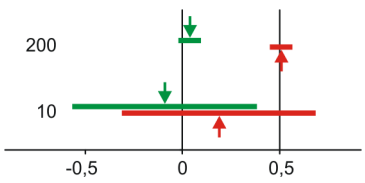
A ***hipotézisvizsgálat*** esetében ***fordított a helyzet***, itt ugyanis éppen ***a hasonlóság tényére vagyunk kíváncsiak***, azaz a mintából becsült számszerű jellemzők alapján szeretnénk eldönteni azt a kérdést, hogy **az adott minta** valóban **a megadott sokaságból származik-e**? A döntés „Igen” vagy „Nem”, de olyan választ is adhatunk, hogy a rendelkezésünkre álló adatok nem elegendők ahhoz, hogy a vizsgált kérdést megnyugtatóan eldöntsük.

Ennyi bevezető után vegyünk egy szemléletes példát, ami ugyan nem túlzottan életszerű, de talán a megértést jobban elősegíti. Tegyük fel, hogy van két puskánk és mindkettővel céltáblára lövünk. (A lövések a puskákban nem okoznak elváltozást, tehát puskánként lényegében azonos feltételek mellett történnek.) A zöld puska zöld lövedékeket lő ki, a piros puska meg pirosakat. Mindkét esetben az összes



123. ábra
A zöld és a piros puskával elvégzett 200, illetve 10 lövésből álló lövéssorozatok eredménye a céltáblákon. (A 10-es sorozat a 200-as sorozat első 10 lövését jelenti.)

124. megjegyzés
Ilyen típusú döntések sokasága vár a gyakorló orvosra. Nevezetesen az említett problémának az orvosi megfelelője például olyasmi lehet, hogy a vizsgálati eredmények alapján meg tudjuk mondani azt, hogy betegünk azt általunk feltételezett betegségben szenved-e vagy valami más baja van.

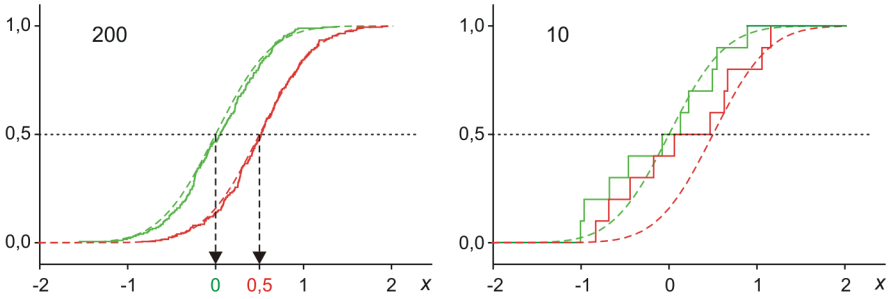


127. ábra
A kétféle puskából kilőtt lövedékek (cm-ben mért) x koordinátájára vonatkozó átlagok és a hozzájuk tartozó 95%-os konfidencia intervallumok szemléltetése a 200, illetve 10 lövésből álló lövéssorozatok esetében.

leadható (végtelen számú) lövés a hozzájuk tartozó koordinátákkal együtt jelenti az alapsokaságot. Egy lövéssorozat pedig, ugyancsak a megfelelő koordinátákkal együtt jelent egy mintát. Tegyük fel továbbá, hogy a zöld puska „pontosan” lő, a piros pedig kicsit jobbra, „félre hord”.

A 123. ábrán, a céltáblákon a zöld és a piros puskából kilőtt, 200 lövésből, illetve 10 lövésből álló lövéssorozatok eredményét tüntettük fel. A 200-as sorozatok eredményéből már ránézésre is megállapítható, hogy a zöld lövedékek a céltábla közepe körül szóródnak, míg a piros lövedékek attól kissé jobbra. Így ha valaki színtévesztő, és a színek alapján nem is látja a különbséget, akkor is meg tudja mondani, hogy a felső céltábla találatai a zöld, az alsó pedig a piros puskához tartoznak. **Tehát csupán a koordináta adatok ismeretében** (a szín ismerete nélkül) **dönteni tudunk abban a kérdésben, hogy melyik puskából adták le a lövéseket** (124. megjegyzés). A 10-es sorozatok esetében a helyzet egyáltalán nem ilyen egyértelmű. Azt gondolhatjuk, hogy ebben az esetben a szín ismeretének hiányában az előbbi kérdést nem tudjuk eldönteni.

Továbblépésként egyszerűsítsük a problémát és foglalkozzunk kizárólag a becsapódott lövedékek x koordinátájával. Először a „zöld” és a „piros” sokaság, valamint a hozzájuk tartozó kétféle minta eloszlásfüggvényeit hasonlítsuk össze.



125. ábra
A zöld és a piros puskából kilőtt lövedékek (cm-ben mért) x koordinátájának eloszlásfüggvényei. A szaggatott vonal a sokaságokra, a folytonos vonal a 200, illetve a 10 elemű mintákra vonatkozik.

A 125. ábrán látható, hogy míg a 200 elemű minta eloszlásfüggvénye mindkét puska esetében jól megközelíti a megfelelő elméleti eloszlásfüggvényt, addig ez nem mondható el a 10 elemű minták eloszlásfüggvényeiről. Nézzük meg, hogy hogyan tükröződik mindez az eloszlásokhoz tartozó számszerű jellemzőkön.

	várhatóérték		ennek standard hibája		medián	
sokaság	0,00	0,50	0,00	0,00	0,00	0,50
200 elemű minta	0,04	0,51	0,03	0,03	0,05	0,52
10 elemű minta	-0,09	0,19	0,21	0,22	0,03	0,26

126. táblázat
A kétféle puskából kilőtt lövedékek x koordinátájának várhatóértéke, ezek standard hibája és a mediánok, a sokaságra és a két mintára vonatkozóan (cm-ben kifejezve).

A 126. táblázatból első látásra kitűnik, hogy a „zöld” várhatóértékek és a mediánok is, mindegyik esetben kisebbek, mint a „pirosak”. Mindezek ismeretében kissé elhamarkodottan, azt a következtetést is levonhatnánk, hogy a piros puska valóban jobbra, félre hord, tehát még akár a 10-es sorozat alapján is választ tudunk adni arra a kérdésre, hogy melyik puskából adták le a lövéseket. Akkor viszont miért gondolhattuk azt az imént, hogy csupán a koordináta értékeket nézve a 123b. és 123d. ábrák alapján ez a kérdés eldönthetetlen.

A magyarázatot a standard hibák különbözősége szolgáltatja. Határozzuk meg ugyanis a 95%-os konfidencia intervallumokat a (112) összefüggésnek megfelelően. A 127. ábrán éppen az figyelhető meg, hogy míg a 200-as minták esetében a zöld és a piros intervallum jól elkülönül, addig a 10-es minta esetében jelentős átfedést mutatnak. Így például hiába tudjuk azt, hogy az alsó piros intervallum definíció szerint 95%-os eséllyel lefedi a 0,5-et, ami az ábra szerint most éppen be is következett, azt is látjuk, hogy ugyanez az intervallum a 0-t is lefedi. Ennek alapján pedig elég nehéz eldönteni, hogy melyiket fedi le jobban.

15.1. A hipotézisvizsgálat főbb lépései, döntés a konfidencia intervallum ismeretében

Maradjunk a piros puska példájánál és felejtsük el, amit a korábbi feltételezés alapján tudunk, nevezetesen, hogy kicsit jobbra, „félre hord”. Prekoncepció nélkül vizsgáljuk meg az a kérdést, hogy ***ez a puska vajon „félre hord-e” vagy „pontosan” lő*** ? Elsőre úgy gondolhatnánk, hogy a vagylagosság miatt a két megfogalmazás teljesen egyenértékű, tehát mindegy, hogy melyik hipotézist vizsgáljuk. Mindkettőre teljesül, hogy amennyiben az egyik igaz, akkor a másik automatikusan hamis.

Figyelmesebben tanulmányozva a két lehetséges hipotézist, észrevehetünk egy fontos különbséget közöttük. Míg pontos lövés csak „egyféle” van, addig pontatlan „sokféle”. (Az egyszerűség kedvéért ismét csak az x koordinátákat vesszük figyelembe.) A koordináták alapján a **pontos lövés** azt jelenti, hogy jóllehet a lövések szóródhatnak, de **várhatóértékük** 0 (a céltábla közepe), tehát **egy jól meghatározott szám**. A **pontatlanok várhatóértéke** pedig 0-tól különböző, de **nem tudjuk, hogy mekkora**. Pontatlanul lőni lehet: kicsit, nagyon, jobbra, balra stb.

Hangsúlyoznunk kell, hogy ezen a ponton nem az a fontos, hogy a meghatározható jellemző 0 vagy nem 0, hanem az, hogy egyértelmű vagy nem. Ezek után arra a kérdésre, hogy melyik hipotézist érdemes egyáltalán megvizsgálnunk már sokkal könnyebb válaszolni, hiszen a nem egyértelmű hipotézissel kapcsolatban sokkal nehezebb bármilyen egyértelmű döntést hozni. Így az **egyértelmű hipotézist**, esetünkben azt, hogy ***a piros puska „pontosan” lő***, előnyben részesítjük, és ezt **tesszük vizsgálatunk tárgyává**. Ezt nevezzük nullhipotézisnek és H_0-lal jelöljük. Az ezzel ellentétes hipotézis, nevezetesen az, hogy ***a piros puska „félre hord”***, az alternatív hipotézis, amelynek szokásos jelölése H_1.

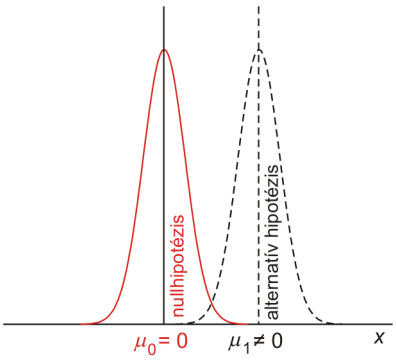
A döntés tehát a nullhipotézis elfogadását, vagy elvetését jelenti. A döntés meghozatalához azonban a nullhipotézis még további pontosításra szorul. Le kell fordítanunk a valószínűségszámítás nyelvére. Ezért további egyszerűsítésként tegyük fel, hogy a lövedékek becsapódásának x koordinátája **normális eloszlású** és mondjuk azt is tudjuk, hogy ennek **szórása** (cm-ben kifejezve) σ_0 **= 0,5** (128. megjegyzés). Amit nem ismerünk, az kizárólag az eloszlás várhatóértéke, de erről meg a nullhipotézisnek megfelelően feltesszük, hogy 0-val egyenlő. Ezzel nem tettünk mást, mint számszerűsítettük a nullhipotézist. Így ***a piros puska „pontosan” lő*** nullhipotézis helyett azt mondhatjuk, hogy az ***adott piros puskából érkező lövedékek x koordinátájának várhatóértéke, $\mu_0 = 0$*** (eloszlása $N(0;0{,}5)$, 129. ábra). Ezt kell ellenőriznünk a birtokunkba jutó adatok (minta) alapján.

Amennyiben a mintánk végtelen elemszámú lenne, azaz ismernénk a piros puskával leadható összes lövés x koordinátáját, akkor ismernénk az alapsokaságot, illetve annak eloszlását, tehát nem kellene mást tennünk, mint a sokaságra jellemző eloszlás várhatóértékét (μ-t) össze kellene hasonlítanunk 0-val. Ebben a helyzetben nincs szükségünk mérlegelésre. Egyszerűen azt mondhatjuk, hogy ha $\mu = 0$, akkor elfogadjuk, ha $\mu \neq 0$, akkor elvetjük a nullhipotézist.

Mivel ez az ideális eset a valóságban gyakorlatilag sohasem fordul elő, ezért a döntési mechanizmus is sokkal bonyolultabb. A 126. táblázat szerint a 200 elemű minta átlaga 0,51, a 10 eleműé pedig 0,19, tehát mindkettő eltér 0-tól, aminek két oka lehet. Az eltérés lehet **látszólagos**, azaz **a különbség csupán a véletlen műve**, ekkor a nullhipotézist igaznak tekintjük, tehát elfogadjuk. Lehet azonban **valódi** is, azaz a különbség abból ered, hogy az adott minta nem a nullhipotézisnek megfelelő, $\mu_0 = 0$ várhatóértékű populációból lett véletlenül kiválasztva, hanem egy másikból, olyanból, amelynek a várhatóértéke, $\mu_1 \neq 0$ (vö. 15.0. rész, 129. ábra). Ekkor a nullhipotézist hamisnak tekintjük, tehát elvetjük és ezzel együtt az alternatív hipotézist fogadjuk el. No, de mi alapján döntsünk?

Tudjuk, hogy egy mintából meghatározott, jól megválasztott konfidencia intervallum a várhatóértéket adott bizonyossággal lefedi (normális eloszlásra: vö. (109) és (112)). Ezért úgy járhatunk el, hogy először megválasztjuk a számunkra „szükséges” bizonyosságot, azaz azt a konfidencia szintet, amely esetén a lefedés már **„számunkra meggyőző, elégséges bizonyosságú”**. Ezután megnézzük, hogy az ehhez tartozó konfidencia intervallum közrezárja-e a 0-t. Ha igen, akkor elfogadjuk, ha nem, akkor elvetjük a nulhipotézist. Tehát a **döntést egy minta alapján, egy előre megválasztott és rögzített konfidencia szint mellett hozzuk meg.** (130. megjegyzés)

128. megjegyzés
Természetesen ezeket a feltevéseket (hipotéziseket) is mind ellenőrizni kell egy „igazi” hipotézisvizsgálat során, de erre még visszatérünk.(vö. 15.7. és 15.8. rész valamint a 194. megjegyzés.)



129. ábra
A nullhipotézis és az alternatív hipotézis szemléltetése sűrűségfüggvényekkel.

130. megjegyzés
Kissé leegyszerűsítve azt mondhatjuk, hogy a büntető ügyekben a bírósági tárgyalás során is „hipotézisvizsgálat” történik. A bíró dönt arról, hogy a vádlott elítélhető-e vagy sem. Tehát van egy eldöntendő kérdés, amire csak igen-nem válasz adható. Számos jogrendben alkalmazzák az ártatlanság vélelmét, ami annyit tesz, hogy mindaddig, amíg a vádlottról be nem bizonyosodik, hogy bűnös, ártatlannak kell tekinteni. Ez a „nem ítélhető el”, (mert nincs bizonyítva a bűnössége) állítás az alaphelyzet, vagy, ha tetszik, kiindulási hipotézis.

A vád képviselőjének a feladata a bizonyítékok felvonultatása annak igazolására, hogy a vád megalapozott. A védelem képviselője a felhozott bizonyítékok hitelét, megbízhatóságát próbálja gyengíteni. A bíró végül értékeli, mérlegeli a bizonyítékokat és dönt. A döntés maga a kiindulási hipotézis, nevezetesen a „nem ítélhető el” állítás elfogadását, vagy elvetését jelenti. Bárhogyan is dönt a bíró, összesen négyféle kimenetel valósulhat meg:

bűnösség / ítélet	ártatlan	bűnös
felmentő	helyes	helytelen
elmarasztaló	helytelen	helyes

131. megjegyzés
Amennyiben nem rendelkezünk azzal az ismerettel, hogy $\sigma = 0{,}5$, akkor a konfidencia határokat a (109) összefüggés helyett a (112) összefüggés felhasználásával határozhatjuk meg, ahol a korrigált tapasztalati szórásokat a mintákból becsüljük.

A 200-as mintára ($s^* = 0{,}49$; $t_{0{,}05} = 1{,}97$) így a konfidencia intervallum 0,44 és 0,58 (nem zárja közre a 0-t);

a 10-es mintára ($s^* = 0{,}71$; $t_{0{,}05} = 2{,}26$) így a konfidencia intervallum –0,32 és 0,7 (közre zárja a 0-t). (Ezek az intervallumok vannak feltüntetve a 127. ábrán is piros színnel.)

A végeredményt tekintetve tehát itt ugyanazt kapjuk, mint az ismert σ-jú esetben.

Érdemes megfigyelni, hogy milyen döntést hozhatunk a két különböző elemszámú minta alapján az eredeti, ***a piros puska „pontosan" lő*** nullhipotézisről. Válasszuk a konfidencia szintet **95%-osra**. Ez azt jelenti, hogy 100 hasonló mintavétel esetén csak 5-ször fordulhat elő, hogy az ennek megfelelő konfidencia intervallum a „véletlen" folytán nem zárja közre a piros puskával leadható összes lövés x koordinátájának várhatóértékét, μ-t, tehát **elég biztos**, hogy közrefogja. A (109) összefüggés szerinti konfidencia határok ($\sigma = 0{,}5$; $u_{0{,}05} = 1{,}96$) a **200-as minta** esetén 0,44 és 0,58; tehát **nem zárja közre a 0-t**; míg a 10-es minta esetén –0,12 és 0,5; tehát közre zárja a 0-t (131. megjegyzés).

Hogyan értelmezhető ez a kétféle eredmény? Amikor jó nagy értékűre (95%) választottuk a konfidencia szintet, ezzel azt akartuk biztosítani, hogy az az **esemény**, hogy μ a konfidencia intervallumon belül van, **legyen csaknem biztos**, és ezzel együtt az **ellentett eseményt**, nevezetesen azt, hogy μ a konfidencia intervallumon kívül van, **minősítsük gyakorlatilag lehetetlennek** (vö. 3.2. rész).

Így amikor a 200 elemű minta esetében **az általunk „lehetetlennek" minősített esemény mégis bekövetkezik**, és így ellentmondásra jutunk, akkor ez csak azt jelentheti, hogy az indirekt bizonyítások logikájához hasonlóan **a nullhipotézis nem lehet „igaz"**, emiatt **el kell vetnünk**. Ennek az a következménye, hogy automatikusan az alternatív hipotézis lép életbe, ami azzal ekvivalens, hogy ***a piros puska „félre hord"***. Úgy tűnik, a minta 200 eleme elegendő bizonyítékot szolgáltatott ehhez a döntéshez (vö. 130. megjegyzés). Persze a konfidencia szintet sohasem választhatjuk 100%-nak, ezért a tévedés sohasem zárható ki.

Mi a helyzet a 10 elemű minta estében? Amennyiben az előző gondolatmenetet követjük, itt a csaknem biztos esemény következett be, tehát nincs ellentmondás adataink és a a nullhipotézis között, így nincs indokunk a nullhipotézis elvetésére sem, a nullhipotézist tehát el kell fogadnunk, ***a piros puska „pontosan" lő***. **Pontosabban fogalmazva a minta alapján nem állíthatjuk az ellenkezőjét** (vö. 123d. ábra). Bár úgy tűnik, hogy a logika szabályai szerint jártunk el, a 200-as minta eredményének ismeretében mindenképpen furcsa helyzet állt elő.

Annak érdekében, hogy ez a furcsaság még szembetűnőbb legyen vizsgáljuk meg azt az esetet is, amikor abból indulunk ki, hogy ***a piros puska „félre hord"***. Ez az előzőek szerint nem lehet nullhipotézis, mert nem egyértelmű. Ha azonban a valószínűségszámítás nyelvén egyértelművé tesszük, akkor megfogalmazható egy ilyen értelmű nullhipotézis is: az ***adott piros puskából érkező lövedékek x koordinátájának várhatóértéke, $\mu_0 = 0{,}5$*** (tehát „félre hord").

132. megjegyzés
A 130. megjegyzésre hivatkozva ez felel meg annak az esetnek, amikor a bűnösséget nem sikerül bizonyítani, és a felmentő ítéletet nem bűncselekmény hiányában, hanem bizonyítottság hiányában mondják ki.

Láthatjuk, hogy a 10 elemű mintából az imént meghatározott konfidencia intervallum ezt az értéket is lefedi (vö. 12.0. rész). (Ezt mutatja ismeretlen σ esetén a 127. ábrán a hosszabb piros vonal.) Ebből viszont az következik, hogy ezt a nullhipotézist is el kell fogadnunk. Tehát ugyanabból a mintából kiindulva azt is elfogadtuk, hogy ***a piros puska „pontosan" lő***, meg azt is, hogy ***a piros puska „félre hord"***. A nyilvánvaló ellentmondás feloldása érdekében kissé módosítanunk kell az eddigi egyértelmű alternatív (elvetés vagy elfogadás) döntéshozatalt.

Az előző 15.0. részben már említettük, hogy van olyan eset is, amikor a rendelkezésünkre álló adatok nem elegendők ahhoz, hogy a vizsgált kérdést megnyugtatóan eldöntsük. Nyugodtan kijelenthetjük, hogy amennyiben a piros puska csak egy kicsit hord félre, ez a kis különbség nem biztos, hogy kiderül a 10 elemű mintából, így ilyenkor ez a minta nem alkalmas a kérdés eldöntésére (132. megjegyzés). Megállapíthatjuk tehát, hogy kis különbségek kimutatására csak a nagyobb elemszámú minták alkalmasak.

15.2. Statisztikai próbák

Mint láthattuk, a hipotézisvizsgálatot az előző 15.1. részben bemutatott módon is el lehet végezni, az egyszerűbb kezelhetőség érdekében azonban inkább statisztikai próbákat használunk. A statisztikai próbákról általánosságban elmondható, hogy alapsokaságok eloszlásaival kapcsolatos hipotézisek ellenőrzésére valók. Tegyük fel például, hogy ismerjük az eloszlás típusát, de nem ismerjük annak paramétereit. Ilyenkor azt vizsgálhatjuk, hogy az egyik ismeretlen paraméter egyenlő -e egy előre megadott számmal. Foglakozhatunk olyan típusú problémákkal is, hogy két eloszlás azonosnak vagy különbözőnek tekinthető, tehát azt vizsgáljuk, hogy mennyire hasonlóak (vö. 1.0. és 13.2. részek). A statisztikai próbák igen **sokfélék** aszerint, hogy **mi az ellenőrizendő hipotézis**, mik az **alkalmazhatóság fel-**

tételei, és mi a **végrehajtás módja**; valamennyinek **közös** azonban a **gondolatmenete**. Ezért kiválasztjuk a legegyszerűbbet, az *u*-próbát, amelynek alkalmazhatósága ugyan meglehetősen korlátozott, de ezen keresztül talán könnyebb megérteni a többit is.

Abból indulunk ki, hogy a **mintákból meghatározott paramétereknek is van valamilyen eloszlásuk** (vö. 14.0. rész), hiszen ők maguk is valószínűségi változók, ha másik mintát választunk, a paraméterek is mások lesznek. Egy ilyen eloszlás konkrét alakja függ

- egyrészt az eredeti változó eloszlásától,
- másrészt attól, hogy melyik becsült paraméterről, vagy általánosabban mondva, melyik statisztikáról van szó,
- harmadrészt pedig a minta elemszámától, pontosabban az azzal szorosan összefüggő szabadságfoktól.

Az összevethetőség kedvéért használjuk a korábbi példát, konkrétabban azt a kérdést, hogy ***a piros puska „félre hord-e”***, az eredeti feltételekkel: a lövedékek becsapódásának x koordinátája normális eloszlású és azt is tudjuk, hogy ennek szórása $\sigma_0 = 0{,}5$. Ellenőrizni akarunk egy, a sokaság várhatóértékére vonatkozó nullhipotézist, H_0-t, nevezetesen azt, hogy $\mu = \mu_0$, vagy másképpen $\mu - \mu_0 = 0$.

Erről a nullhipotézisről az n elemű minta alapján akarunk döntést hozni. Mivel a sokaság várhatóértékét, μ-t nem ismerjük, csak annak becsült értékét a minta átlagát, $\bar{x}$ -t, így az eredetileg megfogalmazott H_0 nullhipotézis direkt módon nem ellenőrizhető. Helyette csak az $\bar{x} \approx \mu_0$, vagy $\bar{x} - \mu_0 \approx 0$ marad, noha erről meg tudjuk, hogy **nem teljesülését a véletlen is okozhatja**.

Amennyiben ugyanis **a mintaátlag standard hibája**, $\sigma_0/\sqrt{n}$ **elég nagy**, akkor elfogadható az az érvelés, hogy **a nullhipotézis** valójában **igaz**, csak a nagy hiba ezt elrejti. Ha azonban ez **a hiba kicsi**, akkor egyáltalán nem lehetünk annyira biztosak a dolgunkban, tehát **az eltérés** akár **„valódi”** is lehet. Emiatt a döntéshez az $\bar{x} - \mu_0$ eltérést a mintaátlag standard hibájához, $\sigma_0/\sqrt{n}$ -hez kell hasonlítanunk. A kettő hányadosa mondja meg, hogy a hibához képest az eltérés nagy vagy kicsi. Ez kisebb átalakítás után:

$$u^* = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0} \quad . \tag{115}$$

Amíg $\bar{x}$ nem egy konkrét számérték, addig valószínűségi változóként kezelhetjük. A 14.5. részben már láthattuk, hogy amennyiben egy ilyen változón standardizálási transzformációt hajtunk végre (a kiindulási feltételek mellett), akkor a transzformált u valószínűségi változó már $N(0;1)$ eloszlású (vö. (107)):

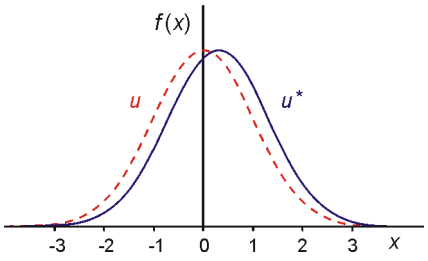
$$u = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma_0} \quad . \tag{116}$$

Láthatjuk, hogy a két kifejezés szembetűnően hasonló. Ezért amennyiben igaz a nullhipotézis, azaz $\mu = \mu_0$, akkor u^* is szükségképpen $N(0;1)$ eloszlású kell, hogy legyen (133. ábra). Ebben az esetben a (108) összefüggés alapján meg tudunk adni egy konfidencia intervallumot (114. ábra):

$$P(|u^*| \le u_p) = 1 - p \quad \text{vagy , ami ezzel ekvivalens,} \tag{117}$$

$$P(|u^*| > u_p) = p \quad . \tag{118}$$

A visszatranszformálást most nem kell elvégeznünk (az x-re vonatkozó konfidencia intervallum pillanatnyilag nem érdekel bennünket), mert a döntést u^* ismeretében is meg tudjuk hozni. Válasszuk ugyanis u_p-t olyan nagyra, hogy a fenti ***u*** abszolút értéke még ennél is nagyobb* esemény bekövetkezésének p valószínűsége gyakorlatilag elhanyagolható legyen. Ha ezután egy konkrét n elemű mintából a (115) összefüggés szerint meghatározzuk u^*-ot és $|u^*| > u_p$, akkor egy olyan esemény következett be, amelyet gyakorlatilag bekövetkezhetetlennek tartottunk, ami csak azt jelentheti, hogy az eredeti $\mu = \mu_0$ feltételezésünk nem helyes, ezért ezt a nullhipotézist elvetjük. Ilyenkor azt mondjuk, hogy a sokaság tényleges μ várhatóértéke és a feltevésben szereplő μ_0 közötti **eltérés** „valódi”, azaz szignifikáns. A szignifikáns szó ebben a szövegkörnyezetben csak azt jelenti, hogy a **„vajon nem véletlen-e”** kérdésre **nemmel** válaszoltunk (134. megjegyzés).



133. ábra
u és u^* közötti különbség szemléltetése. A (115) és (116) összefüggések alapján megfigyelhető, hogy u és u^* csak az eltolás mértékében különbözik egymástól. Míg u-ról biztosan tudjuk, hogy $N(0;1)$ eloszlású, de μ-t nem ismerve nem tudjuk meghatározni értékeit (szaggatott piros görbe). u^* viszont μ_0 ismerete miatt pontosan meghatározható (folytonos kék görbe), de eloszlása csak abban az esetben egyezik meg az $N(0;1)$ eloszlással, ha a nullhipotézis igaz, azaz $\mu = \mu_0$.

134. megjegyzés
A „szignifikáns” sohasem jelenti azt, hogy az eltérés létezését bizonyítottuk; a „nem szignifikáns” pedig még kevésbé azt, hogy biztosan nincs eltérés. Továbbá abból, hogy egy eltérés szignifikáns, még nem következik, hogy az szakmailag is releváns. Arra a kérdésre pedig, hogy mi az ok, a szignifikancia vizsgálat sohasem ad feleletet.

A döntéshez tartozó (önkényesen, de kicsire megválasztott) p valószínűséget szignifikancia szintnek nevezzük, ami a **nullhipotézis elvetésekor bekövetkező esetleges tévedés mértékét jellemzi**. (Például, ha $p = 0{,}05$, akkor 100 független, de hasonló próba elvégzésekor körülbelül 5 esetben tévedünk.) Más szavakkal, a nullhipotézist akkor vetjük el, ha igen kevéssé valószínű az, hogy a H_0-nak pusztán a véletlen folytán ennyire ellentmondó mintát kapjunk a mintavétel során.

Az $|u^*| \le u_p$ esetben, az adott szignifikancia szint mellett, semmi sem indokolná a nullhipotézis elvetését. (Természetesen más szignifikancia szint esetén a döntésünk is módosulhat.) Az elfogadásnak azonban nem egyértelmű a jelentése, amennyiben ugyanis $\mu - \mu_0 \neq 0$, de az eltérés kicsiny, attól az $|u^*| \le u_p$ esemény még nagy valószínűséggel bekövetkezhet (vö. 132. megjegyzés).

Mintafeladat

Használjunk u-próbát annak a kérdésnek az eldöntésére, hogy ***a piros puska „félre hord-e" vagy „pontosan" lő***, a korábbi 200 és 10 elemű minták alapján.

Megoldás: Tegyük fel, hogy a próba használhatóságának feltételei teljesülnek: a lövedékek becsapódásának x koordinátája normális eloszlású, aminek ellenőrizhetőségére még visszatérünk (vö. 15.8. rész) és azt is tudjuk, hogy szórása $\sigma_0 = 0{,}5$.

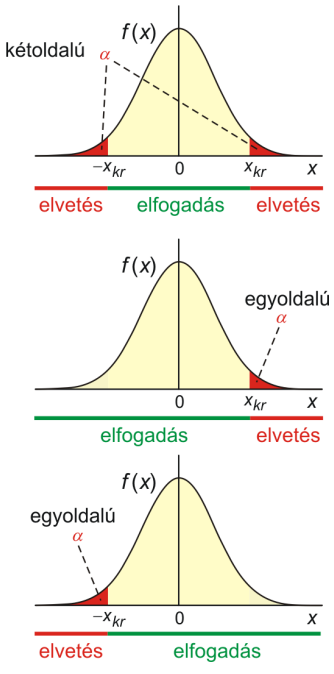
Nullhipotézis: ***a piros puska „pontosan" lő***. Pontosabban megfogalmazva: ***a piros puskából érkező lövedékek x koordinátájának várható értéke, $\mu_0 = 0$***

A (115) összefüggés szerint $u^*_{n=200} = 14{,}4$; $u^*_{n=10} = 1{,}2$ (az átlagok, vagyis a tapasztalati várható értékek a 126. táblázatból kiolvashatók: 0,51 és 0,19). Válasszuk a szignifikancia szintet $p = 0{,}05$-nak, az ehhez tartozó $u_p = 1{,}96$, ezért a ***200-as minta esetében elvetjük, a 10-es minta esetében elfogadjuk a nullhipotézist***.

A 200-as minta esetében a nullhipotézis elvetése azt jelenti, hogy ***a piros puska „félre hord"*** és ezt elég nagy biztonsággal (legalább 95%-os bizonyossággal) állíthatjuk. A 10-es minta esetében a nullhipotézis elfogadása azt jelenti, hogy ***a piros puska „pontosan" lő*** feltételezés **nem cáfolható** a minta alapján.

döntés / valóság	H_0 -t elfogadjuk	H_0 -t elvetjük
H_0 igaz	helyes döntés	elsőfajú hiba
H_0 hamis	másodfajú hiba	helyes döntés

135. táblázat
A helyes és hibás döntési lehetőségek bemutatása.



136. ábra
A számegyenes felosztása elfogadási és elvetési, vagy kritikus tartományra kétoldalú és egyoldalú u-próba esetén.

15.3. Hibalehetőségek, alternatív hipotézisek

Mivel p sohasem egyenlő 0-val, – ellenkező esetben ugyanis $u_p = \pm\infty$ lenne, és így $|u^*| > u_p$ sohasem, $|u^*| \le u_p$ pedig mindig teljesülne – ezért döntésünk megbízhatósága sem lehet 100%-os, tehát a tévedés lehetősége mindig fennáll. Kétféle hibát követhetünk el;

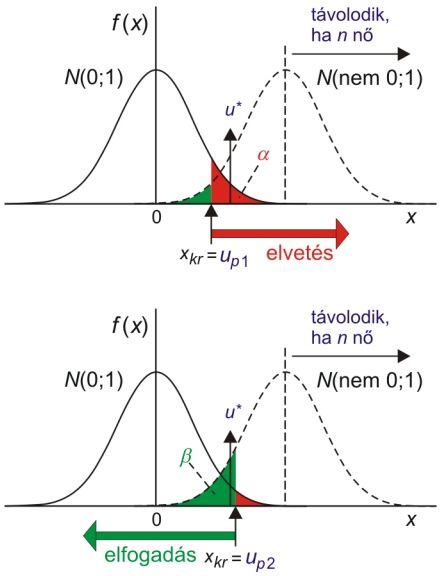
1. elsőfajú hiba: elvetjük a nullhipotézist, holott az igaz, illetve;
2. másodfajú hiba: elfogadjuk a nullhipotézist, pedig az hamis.

Sohasem tudhatjuk, azonban, hogy az adott hibát elkövettük-e vagy sem. Emiatt nem azt mondjuk, hogy a nullhipotézis igaz, hanem azt, hogy elfogadjuk, ami csak annyit jelent, hogy a megfigyelés (minta) nem mond ellent a nullhipotézisnek. Azt sem mondjuk, hogy a nullhipotézis hamis, hanem azt, hogy elvetjük (135. táblázat, vö. 130. megjegyzés; 18.3. rész; 204. ábra; 206. táblázat).

A nullhipotézis megfogalmazásakor egy **alternatív hipotézist, H_1-et** is meg kell adnunk, ami akkor lép életbe, ha a nullhipotézist elvetjük. Első gondolatunk az lehet, hogy ez teljesen felesleges, hiszen az ellentétes állítás megfogalmazása egyértelmű. Így amennyiben $H_0 : \mu - \mu_0 = 0$, akkor $H_1 : \mu - \mu_0 \neq 0$. Sok esetben ez igaz, tehát a ***H_0-t akkor is elvetjük, ha $\mu - \mu_0 < 0$, meg akkor is, ha $\mu - \mu_0 > 0$*** **.** Ekkor beszélünk kétoldalú próbáról.

Vannak azonban olyan esetek is, amikor csak az egyik irányú eltérés felel meg elvárásainknak. Ha például egy gyógyszer hatását vizsgáljuk, legyen az mondjuk egy vérnyomáscsökkentő vagy lázcsillapító, akkor a lehetséges két irányú változás közül csak az egyik, nevezetesen a csökkenés jelenti a gyógyszer hatásosságát. Ilyenkor eleve feltesszük, hogy vérnyomás-, illetve hőmérsékletnövekedés csakis véletlenül következhet be. Ezekben az esetekben a nullhipotézis ugyanaz marad ($H_0 : \mu - \mu_0 = 0$), de az alternatív hipotézis $H_1 : \mu - \mu_0 < 0$, tehát ***H_0-t csak akkor vetjük el, ha $\mu - \mu_0 < 0$*** **.** Ekkor beszélünk egyoldalú próbáról.

A két különböző esetben eszerint kell a számegyenest is felosztanunk elfogadási tartományra, illetve elvetési, vagy kritikus tartományra. Általánosságban azt mondhatjuk, hogy az elsőfajú hiba nagysága α, amely attól függ, hogy hol jelöljük ki a tartományok határait, azaz a kritikus értékeket, x_{kr} (136. ábra).



137. ábra
Az elsőfajú (α) és a másodfajú (β) hiba nagyságának szemléltetése egyoldalú u-próba esetén.

138. megjegyzés
β értékét nem tudjuk addig meghatározni, ameddig az alternatív hipotézis egy specifikus értékét nem rögzítettük. β helyett gyakran az $(1-\beta)$ mennyiséget használjuk, amit a próba erejének szokás nevezni. Ez gyakorlatilag azt adja meg, hogy a próba milyen mértékben képes H_1 H_0-tól való elkülönítésére (vö. 18.3. rész; szenzitivitás (*Se*) a (165) összefüggésben; 208. megjegyzés).

Másképpen mondva a próba ereje annak a valószínűsége, hogy egy különbséget – adott mintanagyság és szignifikancia szint mellett – egy statisztikai próba kimutat. A vizsgálatok tervezésének gyakorlatában sok esetben az erő nagyságának előre megszabott értékéből kiindulva határozzák meg a szükséges mintaelemszámot.

Kétoldalú u-próba esetén α-t úgy kaphatjuk meg, hogy az $N(0;1)$ eloszlás sűrűségfüggvényét az x_{kr} valamint az $-x_{kr}$ értéknél levágjuk és megnézzük, hogy mekkora terület marad a leeső görberészek alatt. Egyoldalú próba esetén csak az egyiknél, x_{kr}-nál, vagy $-x_{kr}$-nál vágunk (vö. 136. ábra). Ezek szerint a **szignifikancia szint** jelentése úgy is megfogalmazható, hogy az nem más, mint az **elsőfajú hiba még tolerálható maximális** nagysága. Tehát a szignifikancia szinttel felülről becsüljük az elsőfajú hiba nagyságát, azt adjuk meg vele, hogy ennél nagyobb elsőfajú hibát már nem tudunk elfogadni.

A másodfajú hiba nagysága β, de ez nem becsülhető α-hoz hasonló egyszerű módon, mert nem ismerjük az alternatív hipotézishez tartozó eloszlás várhatóértékét (137. ábra). (Elsőre csak annyit tudunk róla, hogy nem 0, 138. megjegyzés.)

Fontos hangsúlyoznunk, hogy **α-nak csak a hipotézis elvetésekor**, β -nak pedig csak a hipotézis elfogadásakor **van értelme**. Ugyanakkor nyilvánvaló, hogy x_{kr} változásával a kétfajta hiba nagysága ellentétesen változik, az egyik csökkenésével a másik növekszik és fordítva. Mivel közvetlenül csak az elsőfajú hiba nagysága (α) adható meg, ezért csak a szignifikancia szint alkalmas megválasztásával tudjuk figyelembe venni az ismeretlen nagyságú másodfajú hiba lehetséges elkövetésének kockázatát is (v. ö. 137. ábrán $p1$, illetve $p2$).

Amennyiben egyidejűleg szeretnénk mindkét hiba nagyságát csökkenteni, akkor ezt a minta elemszámának növelésével érhetjük el. Ha ugyanis összevetjük a (115) és (116) összefüggéseket, láthatjuk, hogy az u mint valószínűségi változó minden n-re $N(0;1)$ eloszlású marad, tehát a várhatóértéke mindig 0. Ezzel szemben u^* mint valószínűségi változó, és annak várhatóértéke is minden $\mu \neq \mu_0$ esetben eltávolodik 0-tól, ha n növekszik, hiszen egy nem 0 különbséget szorzunk egyre nagyobb számokkal (v. ö. 137. ábra).

A „jó” próbától elvárjuk, hogy mindkét hiba nagysága 0-hoz tartson, ha a minta elemszáma ∞-hez tart. Ekkor konzisztens próbáról beszélünk és az u-próba, mint láthatjuk, ilyen.

15.4. A statisztikai próbák elvégzésének gyakorlati kérdései

A 15.2. részben a döntést az $|u^*| > u_p$, illetve az $|u^*| \leq u_p$ alapján hoztuk meg. Ehhez arra volt szükségünk, hogy ismerjük az **adott *p*-khez tartozó u_p értékeket**. Ezeket **vagy számítógépes program, vagy táblázat segítségével kaphatjuk meg**. Egy ilyen táblázatban az $N(0;1)$ eloszlás megfelelő kvantilisei szerepelnek (139. táblázat, vö. 114. ábra).

p (egyoldalú)	0,4	0,25	0,1	0,05	0,025	0,01	0,005
p (kétoldalú)	0,8	0,5	0,2	0,1	0,05	0,02	0,01
u_p	0,250	0,674	1,282	1,645	1,960	2,326	2,576

139. táblázat
A $p \rightarrow u_p$ hozzárendelések egy-, illetve kétoldalú próba esetén.

Ilyenkor tehát előre megválasztjuk a p szignifikancia szintet, majd a táblázatból kapott értékkel, vagy értékekkel két részre osztjuk a számegyenest. Egyoldalú próba esetén vagy az $x > u_p$, vagy az $x < -u_p$ a kritikus tartomány, kétoldalú próba esetén pedig a kettő együtt ($x > u_p$ és $x < -u_p$). Természetesen mindkét esetben a komplementer halmaz az elfogadási tartomány (vö. 18. táblázat). Amennyiben u^* a kritikus tartományba esik, akkor a nullhipotézist el kell vetnünk az adott szignifikancia szinten, ellenkező esetben pedig el kell fogadnunk.

140. megjegyzés
p^*-ot úgy is tekinthetjük, hogy az egy mérőszám arra vonatkozóan, hogy a megfigyelt minta mennyire erős bizonyíték a H_0 ellen, a H_1 javára. (Ha mondjuk $p^* = 0{,}333$, akkor ez nem meggyőző bizonyíték H_0 ellen, hiszen ebben az esetben minden harmadik minta akkor is ellentmond H_0-nak, ha az igaz.)

Egyszerűen fogalmazva azt mondhatjuk, hogy azok az eredmények szignifikánsak, amelyekhez kis p^* érték tartozik.

A nullhipotézis elvetéséről illetve elfogadásáról közvetlenül a p érték segítségével is dönthetünk. Míg a táblázatokból csak bizonyos p értékekhez tartozó u_p-ket tudunk kiolvasni, a számítógépes programok segítségével bármely (p, u_p) értékpár meghatározható, így az $x \geq u^*$-hoz tartozó p^* is. Felhasználva azt, hogy a p szignifikancia szint a még tolerálható maximális elsőfajú hiba nagysága, ezért amennyiben $p^* < p$ a nullhipotézist elvetjük, ha $p^* \geq p$, akkor pedig elfogadjuk, de p^* megadásával árnyalhatjuk döntésünket (140. megjegyzés). p^* megadásának akkor van kiemelt szerepe, ha az a szignifikancia szinthez nagyon közel van. Például, ha $p = 0{,}2$, akkor $p^* = 0{,}199$ esetén elvetjük, $p^* = 0{,}201$ esetén megtartjuk a nullhipotézist, pedig a kettő közötti különbség gyakorlati szempontból elhanyagolható.

15.5. A statisztikai próbák fajtái és az ezzel kapcsolatos tudnivalók, illetve félreértések

Statisztikai próbából igen sokféle van (vö. 15.2. rész), így az ***első*** és legnehezebb ***probléma*** legtöbbször éppen **annak eldöntése, hogy az adott kérdésfeltevéshez** – a sok próba közül – **melyiket használjuk** (141. megjegyzés). A ***második probléma***, ami ezzel összefügg, a **nullhipotézis és az alternatív hipotézis pontos megfogalmazása**.

141. megjegyzés
A megfelelő statisztikai próba kiválasztási szabályait nem lehet egyértelműen meghatározni. Ugyanarra a kérdésfeltevésre válaszadásként többféle próba is szóba jöhet.

Már az előzőekben is láthattuk, hogy a nullhipotézis természetes nyelven történő megfogalmazása nem mindig eléggé egyértelmű, így az sokszor további pontosításra szorul, ami adott esetben a nullhipotézis számszerűsítését jelenti (vö. 15.1. rész). Például az, hogy egy kezelés hat egy változóra (mondjuk növeli a tünetmentes időszak hosszát), többféle módon is kifejezhető, pontosítható. A növekedés mérhető a várhatóértékkel ($E(\xi)$), a mediánnal (Me), de akár bizonyos hosszúságú tünetmentes időszakok gyakoriságával is.

Egyáltalán nem mindegy azonban, hogy melyik számszerű jellemzőt választjuk. Ha mondjuk két jellemző közül az egyik eloszlása jobbra elnyújtott, tehát $Me_1 < E_1(\xi)$, a másiké pedig balra, tehát $Me_2 > E_2(\xi)$ (vö. 10.1. rész), akkor előfordulhat olyan eset, hogy az $E_1(\xi) > E_2(\xi)$, de $Me_1 < Me_2$. Látható tehát, hogy menynyire körültekintően kell eljárnunk.

A ***harmadik probléma*** azzal függ össze, hogy az adott próbának **milyen alkalmazhatósági feltételei vannak, és azok teljesülését hogyan lehet ellenőrizni**.

Egy további csoportosítás aszerint történhet, hogy **mit hasonlítunk össze**. Nevezetesen a valószínűségi változóra vonatkozó valamilyen számszerű jellemzőt, például: várhatóértéket, varianciát, esetleg mediánt, vagy csak adott valószínűségeket, de összehasonlíthatjuk magukat az eloszlásokat is.

Egy következő csoportosítási lehetőség azon alapul, hogy **hány mintával** kell dolgoznunk. Például egy mintát hasonlíthatunk egy elméleti eloszláshoz, illetve kettő vagy kettőnél több mintát hasonlíthatunk egymáshoz. Általában feltesszük, hogy a minták függetlenek, de vannak párosított minták is. Ez utóbbi esetben a két minta vagy ugyanazoknak a megfigyelési egységeknek kétszeri megfigyeléséből, vagy valamilyen szempontból összetartozó párok (például páros szervek) megfigyeléséből származnak. Ilyenkor a párosítás miatt két minta helyett csak egyet, legtöbb esetben az adatpárok különbségét vizsgáljuk.

142. megjegyzés
A paraméter kifejezést itt abban az értelemben használjuk, hogy azok egyértelműen jellemzik az eloszlást (vö. 11.0. rész).

Egy másik csoportosítási lehetőség a paraméteres és nemparaméteres eljárások, ami a változó tulajdonságai alapján történik. Tágabb értelemben azok a ***paraméteres*** eljárások, amelyek csak **akkor működnek helyesen, ha a változók egy bizonyos típusú eloszlásúak** (például normális eloszlásúak; 142. megjegyzés).

Ezzel szemben a ***nemparaméteres*** eljárások a vizsgált **változók eloszlásától többé kevésbé függetlenül alkalmazhatók**. Ezeket a módszereket eloszlástól független módszereknek is nevezik.

A próbák elvégezhetősége érdekében gyakran transzformálni kell a változókat. Transzformációkról már esett szó (vö. 12.1. rész), itt még megemlítünk néhány továbbit.

Kategorizáló transzformációkkal folytonos eloszlású változókat ordinális vagy nominális változókká alakíthatunk át. Akkor hasznosak többek között az ilyen eljárások, ha valamely, általunk vizsgált jelenség egy folytonos mennyiségi jellemző változásával minőségileg is megváltozik (143. megjegyzés).

143. megjegyzés
Ilyen változó például az életkor, ahol a következő transzformációval, folytonos változóból bináris változóhoz juthatunk:
életkor < 18 év → gyermek (0),
életkor ≥ 18 év → felnőtt (1).

Rangsor transzformációt számszerű vagy ordinális változó esetén végezhetünk. Ilyenkor a minta elemeit nagyság szerint sorba rendezzük, majd a sorszámokat, vagy másképpen rangokat használjuk az eredeti értékek helyett. Előfordulhat, hogy a sorba rendezés nem egyértelmű, mert vannak azonos értékeink is. Ilyenkor az azonos érékekhez tartozó rangoknak kiszámítjuk az átlagát és ezeket a kapcsolt rangokat rendeljük az azonos értékekhez (144. megjegyzés). Bár rangsor transzformáción alapul számos nemparaméteres eljárás, mint később látni fogjuk a két kifejezés nem szinonimája egymásnak. Ettől függetlenül könnyen belátható, hogy mivel ez az eljárás az eredeti értékek közötti különbségeket már nem veszi figyelembe, ezért a rangok eloszlása nem fogja tükrözni az eredeti változó eloszlását. Így ebben az esetben érthetővé válik az eloszlástól független elnevezés is.

144. megjegyzés
A kapcsolt rangok használatával megőrizzük az adatoknak azt a tulajdonságát, hogy a transzformáció után sem lesz különbség az azonos értékek között továbbá azt is, hogy a rangok összege változatlan marad.
Példa kapcsolt rangok használatára: az 1,2; 3,8; 3,8; 5,2 adatsor rangjai nem az 1; 2; 3; 4, hanem 1; 2,5; 2,5; 4.

Felmerül a kérdés, hogyha vannak eloszlástól független próbák is, amelyeket

Karl Pearson (1857-1963) angol matematikai statisztikus, a modern statisztika alapjainak megteremtője.

minden esetben használhatunk, akkor egyáltalán mi az értelme a paramétereseknek. Azt mondhatjuk, hogy egy paraméteres próba, természetesen amennyiben az alkalmazhatósági feltételei teljesülnek, általában hatékonyabb, mint egy nemparaméteres. Hatékonyságon itt azt értjük, hogy ugyanazokkal a változókkal dolgozva, ugyanakkora első-, illetve másodfajú hiba mellett kisebb elemszámú mintából kapjuk meg az eredményt.

Ezzel a megállapítással függ össze egy másik félreértés, nevezetesen az, hogy minden paraméteres próbának megvan a maga nemparaméteres megfelelője, ami általánosságban nem igaz. A kellő körültekintés itt sem nélkülözhető, mert például, ha kicsit mások a hipotézisek, már nem használhatjuk automatikusan egyiket a másik helyett.

Egy hasonló probléma a robusztusság kérdése. Ez a fogalom azt jelenti, hogy amennyiben egy próba alkalmazhatósági feltételeinek nem mindegyike teljesül, akkor ez mennyiben befolyásolja az eredmény hitelességét. Azt mondhatjuk, hogy a robusztus módszerek általában paraméteresek, ahol eleve szigorúbbak a feltételek, de megvan az a jó tulajdonságuk, hogy akkor is helyesen működnek, ha a mintában kis százalékban előfordulnak olyan elemek is, amelyek miatt a feltételrendszer nem teljesül.

Lényegében minden próba egy jól megválasztott statisztikán alapul (v. ö. 14.0. rész). A jó statisztika érzékeny, azaz a H_0 és H_1 melletti eloszlásának minél inkább különböznie kell egymástól. A statisztika H_0 melletti eloszlását nulleloszlásnak nevezzük. Ennek alapján tudjuk meghatározni p^*-ot is, amely alapján a döntést meghozhatjuk (vö. 14.4. rész; 140. megjegyzés).

Egyes próbák elnevezésében éppen a nulleloszlásra történik utalás, ennek megfelelően beszélünk például t-próbáról, vagy χ^2-próbáról. Az elnevezések során gyakran szerzői nevek is előfordulnak, így találkozhatunk például a Student-féle t-próba vagy a Pearson-féle χ^2-próba kifejezéssel is. Láthatjuk tehát, hogy a próbák elnevezése nem egyértelmű, ugyanaz a próba több néven is szerepelhet, és azonos elnevezésű próbák is különböző módszereket jelenthetnek, így ismét a körültekintésre hívjuk fel a figyelmet.

A következőkben néhány konkrét, a gyakorlatban viszonylag gyakran használt statisztikai próbát mutatunk be.

15.6. A sokaság várhatóértékére vonatkozó statisztikai próbák

A 15.2. részben a próbák általános bemutatásához az u-próbát használtuk, amely talán a legegyszerűbb hipotézisvizsgálati módszer, és amely ebbe a csoportba tartozik, de mint említettük a gyakorlatban nincsen túl nagy jelentősége, hiszen a használhatóságának egyik feltétele, nevezetesen az, hogy a változó szórását eleve ismerjük, csak a legritkább esetben teljesül. Helyette használjuk a t-próbát, ahol a változó (vagy változók) ismeretlen szórását is a minta alapján becsüljük.

1. Egymintás t-próba, amelyben egy ismeretlen várhatóértéket (μ-t) hasonlítunk össze egy hipotetikus értékkel (μ_0-val).

Nullhipotézis: $H_0 : \mu = \mu_0$.

Feltétel: A változó legyen normális eloszlású (μ és σ ismeretlenek).

Statisztika (vö. 110):
$$t^* = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_x^*}. \qquad (119)$$

Nulleloszlás: $\nu = (n - 1)$ szabadságfokú t-eloszlás.

145. megjegyzés
A mintafeladatban p^*-ot számítógép segítségével határozhatjuk meg. t értékén kívül csak annyit kell tudnunk, hogy a szabadságfok $\nu = 5$ és, hogy kétoldalú a próba. (t-táblázat az utolsó oldalon található)
Mivel a nullhipotézist elfogadtuk a nagyobb szignifikancia szint választásának csak az az értelme, hogy ilyenkor a másodfajú hiba nagysága (β) kisebb lesz, (de nem tudjuk, hogy menynyi).
Látható, hogy a t-próba két dologban tér el az u-próbától:
1. a szórást a minta alapján becsüljük és
2. a kritikus értékeket nem az $N(0,1)$, hanem a t-eloszlás alapján határozzuk meg.

Mintafeladat

Egy gyógyszer rendszeres forgalmazásának egyik feltétele a 6 mg hatóanyag tartalom. Egy ellenőrzés alkalmával a tárlóból kivett néhány tabletta adatai a következők voltak (mg-ban): 6,05; 5,95; 5,75; 5,9; 5,95; 6,05. Megtiltható-e a további forgalmazás az ***eltérő*** hatóanyag tartalom miatt?

Megoldás: Normális eloszlást feltételezve (vö. 15.8. rész), használjunk egymintás t-próbát. $\boldsymbol{H_0 : \mu = 6}$.
$n = 6$; $\bar{x} = 5{,}94$; ($\mu_0 = 6$); $s_x^* = 0{,}11$; mindezek és a (119) összefüggés alapján $t^* = -1{,}28$. A feladat szövegében az „eltérő” szó utal arra, hogy az ellenhipotézis, $\boldsymbol{H_1 : \mu \neq 6}$, tehát kétoldalú próbát kell használnunk, mely szerint $p^* = 0{,}26$. Ha a szignifikancia szintet akár 0,25-nek választjuk, p^* még ennél is nagyobb, ezért a nullhipotézist elfogadjuk. Ezek szerint a vizsgálat eredménye nem szolgáltatott elegendő bizonyítékot ahhoz, hogy a forgalmazást megtiltsák (145. megjegyzés).

146. megjegyzés
A nullhipotézis lehet az is, hogy:
$H_0 : \mu_1 - \mu_2 = d$, ahol d egy adott hipotetikus érték és nem feltétlenül 0.

147. megjegyzés
a) A kétmintás t-próbának van olyan változata is, ahol az ismeretlen szórások egyenlőségét nem követeljük meg. Ennek a próbának a két szokásos elnevezése: Welch-próba, vagy „t-próba nem egyenlő varianciák esetére".
Bár a szórások összehasonlítására is vannak próbák (vö. 15.7. rész; F-próba), ha kétségeink vannak a szórások egyenlőségét illetően, inkább használjuk a Welch-próbát.
b) Amennyiben **párosított mintáink** vannak, akkor a „kétmintás" t-próba az adatpárok különbségére mint változóra vonatkozóan minden szempontból ugyanaz, mint az egymintás t-próba (vö. 15.5. rész és a 146. megjegyzés, valamint a 15.6. rész 3. mintafeladat).

148. megjegyzés
A szignifikancia szintek megválasztásának történeti hagyományai vannak. Pearson munkássága nyomán esett a választás az „általában elfogadott" 5%-os szintre, de ennek kitüntetett voltát semmilyen matematikai háttér nem támasztja alá, ezért célszerű a mindenkori p^* megadása is.

2. Kétmintás *t*-próba, amelyben két ismeretlen várhatóértéket (μ_1-t és μ_2-t) hasonlítunk össze egymással. Valójában azt vizsgáljuk, hogy a két minta származhat-e ugyanabból az alapsokaságból.

Nullhipotézis: $H_0 : \mu_1 = \mu_2$ (146. megjegyzés).

Feltételek: A változók legyenek normális eloszlásúak (μ_1, μ_2, σ_1, σ_2 ismeretlenek), de tudjuk, vagy feltételezzük, hogy $\sigma_1 = \sigma_2$, továbbá a két, n_1, illetve n_2 elemű minta legyen független (147. megjegyzés).

Statisztika:
$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^{*2}}{n_1} + \frac{s^{*2}}{n_2}}}, \qquad (120)$$

ahol s^{*2} a két minta korrigált tapasztalati szórásnégyzetének a szabadságfokokkal súlyozott átlaga.

Nulleloszlás: $\nu = (n_1 + n_2 - 2)$ szabadságfokú t-eloszlás.

Hasonló feltételek mellett lehetőség van ***kettőnél több várhatóérték*** összehasonlítására is. Amennyiben ezt páronként kétmintás t-próbával végeznénk, két probléma is felmerülne. Az egyik az, hogy a párok és így az elvégzendő próbák száma is rohamosan nőne: k minta esetén $k(k-1)/2$ lenne (ez 3 minta esetén 3 próbát jelentene, 4 esetén pedig már 6-ot). A másik a bizonyosság kérdése. Vegyük a 3 minta esetét és tegyük fel, hogy mindhárom esetben 5%-os szignifikancia szint mellett elvetettük a nullhipotézist, így esetenként legalább 95%-os a bizonyosság. Ez az érték, a próbák függetlensége miatt, a három esetre együttesen csak $0{,}95^3 \approx 0{,}86$, tehát körülbelül 14% az esetleges tévedés mértéke. Így a kettőnél több várhatóérték összehasonlítását általában nem t-próbákkal végezzük. A új módszer neve a varianciaelemzés, amelyet a 17.0. részben ismertetünk részletesen.

Mintafeladat

Azonos-e a férfiak és a nők pulzusszáma abban a populációban, amelyből az alábbi két minta származik? Az adatokat a táblázat egészre kerekítve, 1/perc egységekben tartalmazza:

$x_{\text{nők}}$	74	87	62	79	71	77			
$x_{\text{férfiak}}$	71	63	70	74	71	69	82	56	78

Megoldás: $\bar{x}_{\text{nők}} = 75$; $\bar{x}_{\text{férfiak}} = 70$. Az átlagok kiszámításából úgy tűnik, hogy a lányoknak magasabb a pulzusszáma. Vajon szignifikáns-e ez az eltérés vagy csak a véletlen okozza? A kérdés eldöntéséhez használjunk kétmintás t-próbát, természetesen föltéve, hogy a próba elvégzéséhez minden feltétel teljesül. $\boldsymbol{H_0 : \mu_{\text{nők}} = \mu_{\text{férfiak}}}$; $\boldsymbol{H_1 : \mu_{\text{nők}} \neq \mu_{\text{férfiak}}}$. (Az ellenhipotézis ismerete azért fontos, mert ebből derül ki, hogy kétoldalú próbát kell használnunk.)

Számítógép segítségével kiszámítjuk az adatainkhoz tartozó valószínűséget: $p^* = 0{,}296$. (A t-próba függvény paramétereinek beállításánál figyelembe vettük még az azonos szórásokat.) Mivel p^* sokkal nagyobb, mint a „szokásos" szignifikancia szintek (0,05; 0,02) (vö. 148. megjegyzés), ezért a nullhipotézist (H_0-t) elfogadjuk. Azt mondhatjuk tehát, hogy a mintáinkban megfigyelhető különbség nem szignifikáns. Így férfiak és a nők pulzusszáma ebben a populációban azonosnak tekinthető.

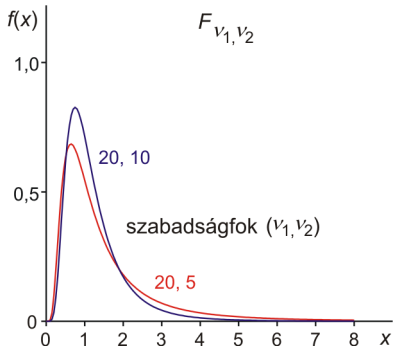
Mintafeladat

Hatásos-e az a lázcsillapító, amelynek beszedése után egy órával az alábbi táblázatban foglaltak szerint változott a vizsgált betegek testhőmérséklete? (A hőmérsékletek °C-ban értendő, tized fokra kerekítve.)

eset	1	2	3	4	5	6	7	8	9	10	11	12	13	14
beadás előtt	37,7	37,7	38,1	37,5	37,7	38,3	38,1	38,6	37,8	38,7	38,7	38,4	38,5	38,3
beadás után	37,2	38,9	36,5	37,3	37,5	38	37,3	38,3	38	37,9	37,6	38,1	37,6	38,3
különbség	-0,5	1,2	-1,6	-0,2	-0,2	-0,3	-0,8	-0,3	0,2	-0,8	-1,1	-0,3	-0,9	0

Megoldás: A kérdés eldöntéséhez tipikusan a párosított kétmintás t-próbát használhatjuk, amely mint említettük megegyezik az egymintással a különbségekre vonatkozóan. Így a különbségeknek kell normális eloszlásúnak lennie, amit ellenőrizni kell (vö. 15.8. rész). A nullhipotézis általános megfogalmazása az, hogy a lázcsillapító nem hatásos. $\boldsymbol{H_0 : \mu_{\text{előtte}} = \mu_{\text{utána}}}$ vagy, ami ezzel ekvivalens $\mu_{\text{előtte}} - \mu_{\text{utána}} = 0$. $\boldsymbol{H_1 : \mu_{\text{előtte}} > \mu_{\text{utána}}}$. (Ebből az ellenhipotézisből kiderül, hogy egyoldalú próbát kell használnunk.)

A megfelelő valószínűség: $p^* = 0{,}021$. Például 5%-os szignifikancia szint mellett a nullhipotézist elvethetjük, amiből arra következtethetünk, hogy a lázcsillapító hatásos. Ilyenkor p^* azt jelzi, hogy a nullhipotézis akár 2,5%-os szignifikancia szint mellett is elvethető, ami arra utal, hogy az esetlegesen bekövetkező tévedés mértéke sem nagyobb ennél (vö. 148. megjegyzés).



149. ábra
A (20, 5) valamint a (20, 10) szabadságfokú F-eloszlás sűrűségfüggvényének szemléltetése. $E(\xi) = n_1/(n_1 - 2)$, $(n_1 \geq 3)$; amely $n_1 \to \infty$ esetén 1 - hez tart.

Ronald Aylmer Fisher (1890-1962) angol statisztikus. Az ő tiszteletére kapta a próba és az eloszlás az F nevet.

15.7. A sokaság varianciájára vonatkozó statisztikai próbák

1. χ^2-próba egy variancia vizsgálatára, amelyben egy ismeretlen varianciát (σ^2-et) hasonlítunk össze egy hipotetikus értékkel (σ_0^2-tel).

Nullhipotézis: $H_0 : \sigma^2 = \sigma_0^2$.

Feltételek: A változó legyen normális eloszlású (μ és σ ismeretlenek).

Statisztika (vö. 102): $$\chi^{2^*} = \frac{(n-1)s_x^{*2}}{\sigma_0^2}. \qquad (121)$$

Nulleloszlás: $\nu = (n - 1)$ szabadságfokú χ^2-eloszlás.

Mivel magát a χ^2-próbát más vonatkozásban is fogjuk használni, így részletesebben ott térünk vissza az alkalmazására (vö. 15.8. rész).

2. *F*-próba, amelyben két ismeretlen varianciát (σ_1^2-et és σ_2^2) hasonlítunk össze egymással.

Nullhipotézis: $H_0 : \sigma_1^2 = \sigma_2^2$.

Feltételek: A változók legyenek normális eloszlásúak (μ_1, μ_2, σ_1, σ_2 ismeretlenek), továbbá a két, n_1, illetve n_2 elemű minta legyen független.

Statisztika (vö. 102): $$F^* = \frac{s_{x_1}^{*2}}{s_{x_2}^{*2}} \geq 1. \qquad (122)$$

Nulleloszlás: $\nu_1 = (n_1 - 1)$; $\nu_2 = (n_2 - 1)$ szabadságfokú F-eloszlás (149. ábra).

Bár ezt az eloszlást eddig még nem használtuk, belátható, hogy ez is a normális eloszlásból származtatható és a χ^2-eloszlás illetve a t-eloszlás további általánosításának tekinthető (a bizonyítást itt is mellőzzük). (Ha a (121) és (122) összefüggéseket összehasonlítjuk, azért az megfigyelhető, hogy $\sigma_1^2 = \sigma_2^2$ esetén, F két – a szabadságfokokkal elosztott – független χ^2-eloszlású valószínűségi változó hányadosából nyerhető.)

Mintafeladat

Az előző (15.6.) rész első mintafeladatában egy gyógyszer hatóanyag tartalmát ellenőriztük. Egy új gyártósor beállítása után azt figyelték meg, hogy az ezen készített tabletták hatóanyag tartalma sokkal jobban ingadozik, mint az előzőn készítetteké. Így felmerült egy újabb kérdés, vajon szignifikáns-e ez az eltérés.

A régi és új adatokat a táblázat mg egységekben tartalmazza:

x_1	6,05	5,95	5,75	5,9	5,95	6,05
x_2	6,15	5,65	6,20	5,85	5,80	6,15

Megoldás: A szórások összehasonlításából ($s_1^* = 0{,}11$; $s_2^* = 0{,}23$) az látszik, hogy az új gyártósoron gyártott tabletták szórása több, mint kétszeres a korábbinak. Normális eloszlást és független mintákat feltételezve (vö. 15.8. rész), használjunk F-próbát. $\boldsymbol{H_0 : \sigma_1^2 = \sigma_2^2}$; $\boldsymbol{H_1 : \sigma_1^2 \neq \sigma_2^2}$. (Kétoldalú próba.)

$F^* = 4{,}37$ (mindig a nagyobb tapasztalati szórásnégyzetet kell osztani a kisebbel, csak így kapunk 1-nél nagyobb értéket, de ezt a számítógépes programok automatikusan figyelembe veszik) a számítógép segítségével meghatározott $p^* = 0{,}13$, amely lényegesen nagyobb mint a „szokásos” szignifikancia szintek (0,05; 0,02), ezért a nullhipotézist elfogadjuk. Ezek szerint a szórásokban tapasztalt különbség a látszat ellenére nem szignifikáns (150. megjegyzés).

150. megjegyzés
Bár esetünkben a két szabadságfok megegyezett, amennyiben táblázatot használunk a döntéshozatalhoz, vigyáznunk kell a szabadságfokok sorrendjére, mert a táblázat nem szimmetrikus. Így a számláló és a nevező szabadságfoka nem cserélhető fel.

Hasonló feltételek mellett lehetőség van ***kettőnél több variancia*** összehasonlítására is, de ezzel itt nem foglalkozunk. Amire viszont felhívjuk a figyelmet az az, hogy F-próbát fogunk használni az előző (15.6.) részben már említett ***varianciaelemzésre*** is (vö. 17.0. rész).

15.8. Eloszlásokra vonatkozó statisztikai próbák

Ebben a részben sokszor fog szerepelni a χ^2-próba elnevezés, mert a megfelelő statisztikák nulleloszlása vagy egzaktul megegyezik a χ^2-eloszlással, vagy legalább aszimptotikusan ahhoz tart (vö. 14.1. rész; 109. ábra). Így a megkülönböztetés kedvéért itt is illik megmondani, hogy éppen milyen célra használjuk a χ^2-próbát. Három egymáshoz hasonló módszerről lesz szó: illeszkedésvizsgálat, függetlenségvizsgálat, homogenitásvizsgálat.

Az eddigi próbák mindegyikének feltételei között szerepelt, hogy a változó

legyen normális eloszlású. Az ***illeszkedésvizsgálat*** épp ilyen kérdések ellenőrzésére való. Általánosságban azt mondhatjuk, hogy egy ismeretlen eloszlást (ebből vettük a mintát) hasonlítunk össze egy hipotetikus eloszlással. Amikor a hipotetikus eloszlás típusán túlmenően, annak paramétereit is ismerjük, akkor tiszta illeszkedésvizsgálatról, ennek hiányában becsléses illeszkedésvizsgálatról beszélünk. Ilyenkor a paramétereket a minta alapján becsüljük.

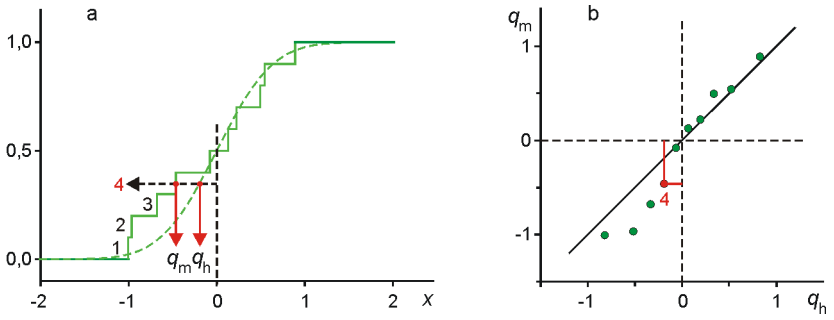
0. Kvantilis-kvantilis ábra, (QQ ábra), amely nem is próba, hanem egy grafikus módszer illeszkedésvizsgálat céljára. (Ezért kapta a 0. sorszámot.)

Nullhipotézis: Az ismeretlen eloszlás, amelyből a minta származik, azonos a hipotetikus eloszlással.

Feltételek: A többi hasonló célú módszerrel ellentétben, kis elemszámú minták esetén is alkalmazható.

Példaként tekintsük a 125. ábra jobb oldali részén látható, a korábbiakban már többször is szereplő zöld lövedékek x koordinátájának hipotetikus, és a 10 elemű mintára vonatkozó, „valódi" eloszlásfüggvényeit. Ezt tüntettük fel ismét a 151a. ábrán is. Az összehasonlítást már akár ennek az ábrának a segítségével is elvégezhetjük, de az illeszkedés jobban megítélhető a QQ ábra alapján.

A módszer lényege az, hogy ha a minta valóban az adott eloszlásból származik, akkor a minta eloszlásfüggvényéhez, illetve a hipotetikus eloszlásfüggvényhez tartozó kvantilisek, q_m és q_h nem lehetnek messze egymástól. Ha ezután a q_m-eket a q_h-k függvényében ábrázoljuk, akkor ezek a pontok az $f(x) = x$ egyenes közelében helyezkednek el (151b. ábra; 152. megjegyzés).



151. ábra
a) A zöld puskából kilőtt lövedékek (cm-ben mért) x koordinátájának eloszlásfüggvényei. A szaggatott vonal a sokaságra, a folytonos vonal a 10 elemű mintára vonatkozik. b) Az ebből konstruált QQ ábra.

152. megjegyzés
Hasonló módon készíthető ábra a valószínűségekre vonatkozóan is (PP ábra), amikor a távolságokat a függőleges, azaz a valószínűség tengely mentén mérjük. A pontoknak ilyenkor is az $f(x) = x$ egyenes közelében kell elhelyezkedniük.

A módszer működik két tapasztalati eloszlásfüggvénnyel is, ezért homogenitásvizsgálatra is alkalmas.

A nullhipotézist akkor vetjük el, ha a QQ ábra szemrevételezésekor a pontoknak az egyenestől való eltérése nem tűnik véletlennek.

1. χ^2-próba illeszkedésvizsgálatra, amely első közelítésben diszkrét változók elemzésére alkalmas, mivel gyakoriságokat hasonlít össze. Folytonos változók esetében tehát először mindig osztályokat kell megadnunk (vö. 7.1. rész) és az ezekhez tartozó gyakoriságokat használjuk a vizsgálat során.

Nullhipotézis: Az ismeretlen eloszlás, amelyből a minta származik, megegyezik a hipotetikus eloszlással. Becsléses illeszkedésvizsgálat esetén ez utóbbi paramétereit a minta alapján becsüljük.

Feltételek: A próba aszimptotikus volta miatt csak nagy elemszámú mintákra alkalmazható. (Például $n > 50$, de minél nagyobb a minta elemszáma annál nagyobb a bizonyosság is.) Ezen túlmenően a változó lehetséges értékeit úgy kell osztályokba sorolni, hogy **egy osztályon belül a hipotetikus eloszlásnak megfelelő várható gyakoriság legalább 5 legyen** (153. megjegyzés).

Statisztika:
$$\chi^{2^*} = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}, \qquad (123)$$

ahol k az osztályok száma, f_i a mintából származó (**megfigyelt**), e_i pedig a hipotetikus eloszlásnak megfelelő **várható** gyakoriság az i-edik osztályban.

Nulleloszlás: Aszimptotikusan χ^2-eloszlás; a tiszta esetben $\nu = (k - 1)$, a becsléses esetben $\nu = (k - m - 1)$ szabadságfokú, ahol k az osztályok, m pedig a mintából becsült paraméterek száma.

153. megjegyzés
A feltétel teljesülését esetenként a szomszédos osztályok összevonásával érhetjük el. (154. megjegyzés)
Amennyiben ez az út nem járható, akkor a Fisher-féle egzakt próbát alkalmazzuk.

154. megjegyzés
Az osztályhatárok különböző megválasztása nagymértékben befolyásolhatja az eredményeket.

Az $F(x)$ eloszlásfüggvény ismeretében e_i egyszerűen kiszámítható:

$$e_i = n\left(F(x_{\text{felső}}) - F(x_{\text{alsó}})\right), \qquad (124)$$

ahol n a minta elemszáma, $x_{\text{alsó}}$ és $x_{\text{felső}}$ pedig az i-edik osztály alsó és felső határát jelenti (154. megjegyzés).

Mintafeladat

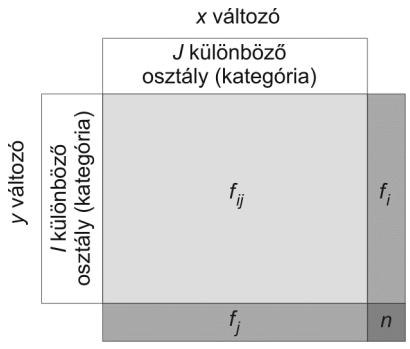
Azt akarjuk ellenőrizni az alábbi kockadobások eredménye alapján, hogy cinkelt-e a dobókocka.

a dobás eredménye	1	2	3	4	5	6
gyakorisága	13	8	7	5	6	11

Megoldás: Tiszta illeszkedés vizsgálatról van szó, azt akarjuk ellenőrizni, hogy a fenti gyakoriságok ellentmondanak-e annak a nullhipotézisnek, hogy a hipotetikus eloszlás egyenletes (vö. 36. ábra). Erre a célra használjunk χ^2-próbát. Mivel a dobások száma $n = 50$, az egyenletes eloszlásnak megfelelő várható gyakoriság mindegyike $e_i = 8{,}33$. A (123) összefüggés alapján $\chi^{2*} = 5{,}68$. (Tökéletes egyezés esetén ez az érték 0 lenne.) Mivel $5{,}68 < \chi^2 = 11{,}07$ (5%-os szignifikancia szintet választva a $6 - 1 = 5$ szabadságfokú eloszlás esetén), ezért a nullhipotézist elfogadjuk. Azt mondhatjuk tehát, hogy a dobókocka akár cinkelt is lehet, de ezt a fenti adatok alapján nem tudtuk bizonyítani.

Érdekes eredményre juthatunk, ha feltesszük, hogy a dobások megismétlése során ugyanezek a gyakoriságok jönnek ki. A két sorozatot egynek tekintve ($n = 100$; $e_i = 16{,}66$; a gyakoriságok a táblázatban megkétszereződnek) $\chi^{2*} = 11{,}36$ és $11{,}36 > \chi^2 = 11{,}07$, tehát szintén 5%-os szignifikancia szint mellett most a nullhipotézist elvetjük, ami már azt támasztja alá, hogy a dobókocka valóban cinkelt lehet.

A 12.0. részben két valószínűségi változó függetlenségéről már szót ejtettünk. Megállapítottuk, hogy a függetlenség úgy is megfogalmazható, hogy az egyik változó (például y) bármely értéke mellett, a másik változó (x) eloszlása ugyanolyan marad, egyedül a normálási tényező módosul (v. ö. 79a. ábra; (66), (67), (68) összefüggések). Ez folytonos változó esetén azt is jelenti, hogy függetlenül az egyik változó értékeitől, a másik változó eloszlására jellemző sűrűségfüggvények egymással és a peremeloszlás sűrűségfüggvényével is meg fognak egyezni, ha a normálástól eltekintünk. Ezt a tulajdonságot használjuk ki akkor, amikor két változóra vonatkozóan ***függetlenségvizsgálatot*** hajtunk végre. Természetesen ilyenkor a mintavétel csak a megfigyelési egységek kiválasztását jelenti és minden ilyen egységhez a két változónak megfelelően két adat tartozik. Ha az eset-változó táblázatra gondolunk (v. ö. 93. ábra), akkor annak mindig két oszlopa lesz a vizsgálatunk tárgya.



155. ábra
A kontingencia tábla általános szerkezete. f_{ij} a két változó megfelelő osztályaihoz (kategóriáihoz) tartozó gyakoriságok; f_i , f_i a megfelelő peremeloszlásokhoz tartozó gyakoriságok; n a minta elemszáma.

156. megjegyzés
A latin „contingens” = esetleges, a szükségszerű ellentétéből származó szó.

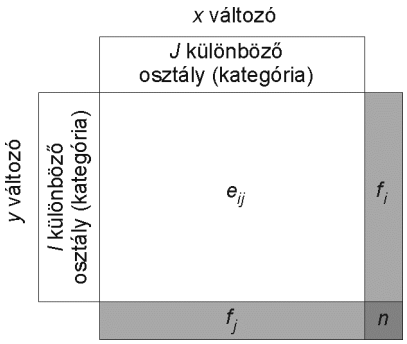
2. χ^2-próba függetlenségvizsgálatra, amelyben először mindkét változó értékkészletét az illeszkedésvizsgálatnál már alkalmazott módon osztályokba kell sorolnunk. Diszkrét változók esetén ezek általában adottak, folytonos változók esetén tetszőlegesen választhatók (v. ö. 154. megjegyzés). A módszer kategoriális változók elemzésére is alkalmas (vö. 32. megjegyzés), hiszen végül csak gyakoriságokat hasonlítunk össze. Az osztályok (kategóriák) száma I az egyik és J a másik változóra vonatkozóan ($I \neq J$ általában). A mintavétel és az osztályokba sorolás után nyert f_{ij} **megfigyelt** gyakoriságokat egy I sorból és J oszlopból álló táblázatban foglalhatjuk össze. Ezt kiegészítve, az $(I + 1)$-edik sorban, illetve $(J + 1)$-edik oszlopban a megfelelő oszlop- és sorösszegeket tüntetjük fel, amelyek az adott peremeloszláshoz tartozó (f_j, f_i) gyakoriságokat adják meg. Az összes gyakoriság, bárhogy is adjuk őket össze, természetesen a minta elemszámát, n-et adja meg:

$$\sum_{i=1}^{I} f_{ij} = f_j \,, \qquad \sum_{j=1}^{J} f_{ij} = f_i \,, \qquad (125)$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = \sum_{i=1}^{I} f_i = \sum_{j=1}^{J} f_j = n \,. \qquad (126)$$

Az így nyert táblázatot kontingencia táblának nevezzük (155. ábra; 156. megjegyzés).

A próba elvégzéséhez a függetlenség fennállása esetén **várható** e_{ij} gyakoriságokat is meg kell határoznunk. Ehhez az események függetlenségére vonatkozó (17) összefüggést alkalmazzuk, de a valószínűségek helyett a relatív gyakoriságokkal számolunk: $P(A) \approx f_i/n$; $P(B) \approx f_j/n$. Ennek alapján $e_{ij}/n = (f_i/n)(f_j/n)$,



157. ábra
A várható gyakoriságok táblázata. Szerkezetében megegyezik a kontingencia táblával. A különbség csak annyi, hogy az f_{ij} megfigyelt gyakoriságok helyett az e_{ij} várható gyakoriságok szerepelnek benne (v. ö. (127) összefüggés).

ahonnan e_{ij} az alábbi összefüggés szerint meghatározható:

$$e_{ij} = \frac{f_i f_j}{n} \ . \qquad (127)$$

Ily módon minden sorban és oszlopban az adott peremeloszlásoknak megfelelő eloszlásokhoz jutunk (157. ábra). Ezt kell összevetnünk a kontingencia táblában szereplő megfigyelt gyakoriságokkal.

Nullhipotézis: A két változó (x, y) független egymástól.

Feltételek: Az illeszkedésvizsgálathoz hasonlóan elég nagy mintával kell dolgoznunk ahhoz, hogy az e_{ij} várható gyakoriság a táblázat mindegyik cellájában legalább 5 legyen (v. ö. 153. megjegyzés).

Statisztika:
$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \ , \qquad (128)$$

ahol f_{ij} a kontingencia tábla **megfigyelt** gyakoriságait, míg e_{ij} a **várható** gyakoriságokat jelenti.

Nulleloszlás: Aszimptotikusan $\nu = (I-1)(J-1)$ szabadságfokú χ^2-eloszlás, ahol I és J az egyik (y), illetve a másik (x) változó szerinti osztályok számát jelöli.

Mintafeladat

Döntsük el a mellékelt kontingencia tábla alapján, hogy független-e a biofizika vizsgaeredmény a csoportbeosztástól?

osztályzat

csoport	1	2	3	4	5	Σ
A	23	12	5	15	3	58
B	21	13	9	12	3	58
C	22	14	9	9	4	58
D	20	12	16	10	2	60
E	25	8	12	14	2	61
F	27	5	9	14	6	61
G	32	4	8	11	5	60
H	27	7	13	6	7	60
Σ	197	75	81	91	32	476

Megoldás: A függetlenségvizsgálatra alkalmazzunk χ^2-próbát. A nullhipotézis az, hogy a vizsgaeredmény és a csoportbeosztás (azaz a két változó) független egymástól. Ezután ellenőrizzük a kirótt feltételek teljesülését. Mivel a legkisebb peremeloszlás gyakoriságokból meghatározott várható gyakoriság (58·32/476 = 3,9) kisebb, mint 5, ezért a 4-es és 5-ös osztályzatokat összevonjuk. Az új kontingencia tábla alapján elkészítjük a várható gyakoriságok táblázatát is:

osztályzat

csoport	1	2	3	4-5	Σ
A	23	12	5	18	58
B	21	13	9	15	58
C	22	14	9	13	58
D	20	12	16	12	60
E	25	8	12	16	61
F	27	5	9	20	61
G	32	4	8	16	60
H	27	7	13	13	60
Σ	197	75	81	123	476

osztályzat

csoport	1	2	3	4-5	Σ
A	24,0	9,1	9,9	15,0	58
B	24,0	9,1	9,9	15,0	58
C	24,0	9,1	9,9	15,0	58
D	24,8	9,5	10,2	15,5	60
E	25,2	9,6	10,4	15,8	61
F	25,2	9,6	10,4	15,8	61
G	24,8	9,5	10,2	15,5	60
H	24,8	9,5	10,2	15,5	60
Σ	197	75	81	123	476

Ezek ismeretében, a számítógépes program segítségével meghatározott $p^* = 0{,}18$, amely lényegesen nagyobb, mint a szokásos szignifikancia szintek (például 0,05). Amennyiben a (128) összefüggés szerint meghatározott $\chi^{2*} = 26{,}7$ értéket veszszük alapul, és ezt hasonlítjuk össze az 5%-os szignifikancia szinthez tartozó $\chi^2_{\nu=21} = 32{,}7$ értékkel, ugyanarra az eredményre jutunk, nevezetesen a nullhipotézist elfogadjuk. Ezek szerint a táblázatban szereplő csoportonkénti nagy eltérések a függetlenséget szignifikánsan nem érintik.

Ennek a résznek az elején említett harmadik módszer, a ***homogenitásvizsgálat*** visszavezethető a függetlenségvizsgálatra.

158. megjegyzés
Láthatjuk, hogy a homogenitás- és a függetlenségvizsgálat lényegében csak a mintavétel módjában különbözik egymástól.
A k változónak annyi osztálya lesz, ahány összehasonlítandó mintánk van. Az x változó értékeit pedig a függetlenségvizsgálatnál már ismertetett módon kell osztályokba sorolni.

3. χ^2-próba homogenitásvizsgálatra, amelyben a probléma úgy vetődik fel, hogy: két vagy több (x változójú) független minta származhat-e ugyanabból az ismeretlen eloszlású alapsokaságból?

Nullhipotézis: Az ismeretlen alapsokaságok, amelyekből a minták származnak mind azonos eloszlásúak.

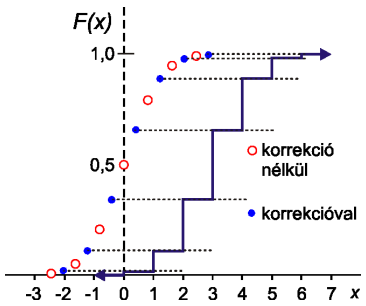
Első lépésként ezeket az alapsokaságokat sorszámmal látjuk el. Ily módon egy adott minta minden eleméhez azt a k számot is hozzárendeljük, amelyik annak az alapsokaságnak a sorszáma, amelyből a minta származik. Így már minden megfigyelési egységhez az x és k változónak megfelelően két adat tartozik. Könnyen belátható, hogy az eredeti nullhipotézis, ekvivalens azzal, hogy x és k független egymástól. Ettől kezdve **a feltételek, a statisztika és a nulleloszlás** – egy fontos kivételtől eltekintve – **ugyanazok**, mint a függetlenségvizsgálat esetében. A különbség csupán annyi, hogy itt nem egyetlen mintánk van, hanem kettő vagy több (158. megjegyzés).

159. megjegyzés
Az előjel próba elnevezés onnan származik, hogy a próba eredetileg a $Me = 0$ hipotézis ellenőrzésére szolgált, és ilyenkor a számoláshoz csak a mintabeli értékek előjelét használjuk.

160. megjegyzés
A próba megfogalmazható medián helyett tetszőleges kvantilisre is.

161. megjegyzés
Nagy minták esetén a binomiális eloszlás jól közelíthető Poisson- vagy normális eloszlással (vö. 11.1. rész).

162. megjegyzés
A (129) összefüggésben a 0,5-es eltolással (folytonossági korrekcióval) a binomiális eloszlás diszkrét és a normális eloszlás folytonos volta közötti eltérést tudjuk korrigálni (163. ábra).



163. ábra
A standardizálási transzformációnak alávetett (6, 0,5) paraméterű binomiális eloszlás eloszlásfüggvénye folytonossági korrekció nélkül (piros karika) és folytonossági korrekcióval (kék pötty).

15.9. A sokaság mediánjára vonatkozó statisztikai próba (előjel próba)

Ez a próba (159. megjegyzés) és az ezt követő próbák ***nemparaméteresek***, tehát ***eloszlástól független*** módszerek.

Ezzel a próbával egy ismeretlen mediánt (Me-t) hasonlítunk össze egy hipotetikus értékkel (Me_0-lal) (160. megjegyzés).

Nullhipotézis: H_0: $Me = Me_0$.

Feltétel: A változó legyen folytonos.

Statisztika: Az Me_0-nál kisebb, illetve nagyobb mintaelemek száma közül a kevesebb (k).

Nulleloszlás: (n,p) paraméterű binomiális eloszlás, ahol n az Me_0-tól különböző mintaelemek száma, és $p = 0{,}5$. (Az Me_0-lal megegyező értékeket nem vesszük figyelembe.)

A medián egy mintában definíció szerint olyan érték, amelynél kisebb, illetve nagyobb értékek ugyanolyan gyakorisággal fordulnak elő. Ezért, ha véletlenül kiválasztunk egy mintaelemet, akkor $p = 0{,}5$ a valószínűsége annak, hogy az a mediánnál kisebb. Annak a valószínűségét, hogy n-szer megismételve a kiválasztást éppen k-szor kapjunk a mediánnál kisebb értéket, az (n,p) paraméterű binomiális eloszlás segítségével határozhatjuk meg (161. megjegyzés). Az eloszlásfüggvény k-hoz tartozó függvényértéke az egyoldalú próbára vonatkozó p^*-ot adja meg. Mivel az eloszlás szimmetrikus, ezért ennek kétszeres ($2p^*$) a kétoldalú próbára vonatkozó megfelelő érték.

Amennyiben a binomiális eloszlást normális eloszlással közelítjük, akkor az alábbi standardizálási transzformációval (vö. (71), (48), (49)) standard normális eloszlású változóhoz juthatunk (162. megjegyzés).

$$\zeta = \frac{k - np + 0{,}5}{\sqrt{np(1-p)}} . \qquad (129)$$

Mintafeladat

Az újszülöttek érkezése a terhesség 40. hetében várható (a teljes népességre vonatkozó várható érték és a medián is ebbe a tartományba esik). Egy kisebb városban, az utóbbi hónapok születési adatai alapján (lásd a táblázatot) mondhatjuk-e azt, hogy az újszülöttek, valamilyen okból kifolyólag a vártnál szignifikánsan korábban születnek meg.

hét	34.	35.	36.	37.	38.	39.	40.	41.	42.
születések száma	1	1	1	1	2	2	3	3	1

Megoldás: Mivel a minta átlaga 38,7, mediánja 39, ezért azt gondolhatjuk, hogy valóban korábban születnek az újszülöttek. Bár a születésig eltelt idő folytonos változó, csak itt hetekben mérjük, az eloszlása nem szimmetrikus, hiszen gyakoribb a koraszülés, mint a túlhordott terhesség. Emiatt a változó nem lehet normális eloszlású, tehát a kérdés eldöntésére a t-próba nem alkalmas. Ilyenkor használhatjuk például a mediánra vonatkozó eloszlástól független előjel próbát.

H_0 : $Me = 40$; H_1 : $Me < 40$. (Egyoldalú próba.) A 40-nél kisebb mintaelemek száma 8, a nagyobbaké 4 (3 egyenlő vele), ezért $k = 4$, amelyhez a megfelelő (12, 0,5) paraméterű binomiális eloszlásból meghatározhatjuk a kívánt valószínűséget $p^* = 0{,}19$. Ez az érték az 5%-os szignifikancia szintnél (de még 10%-osnál is) lényegesen nagyobb, ezért a nullhipotézist elfogadjuk. Nem mondhatjuk tehát, hogy a mellékelt minta alapján az adott városban az újszülöttek a vártnál szignifikánsan korábban születnek meg.

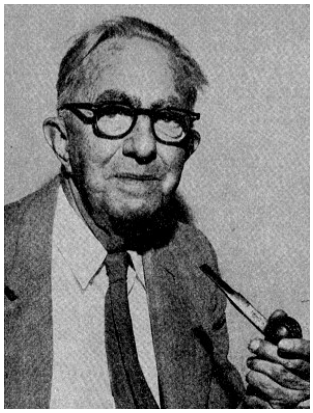
Amennyiben a (129) összefüggés szerint számolunk, akkor $\zeta = -1{,}5/1{,}732 = -0{,}866$. Az ehhez tartozó valószínűséget a standard normális eloszlás eloszlásfüggvényének felhasználásával határozhatjuk meg: $p^* = 0{,}19$. Mivel a binomiális eloszlásból, illetve a standard normális eloszlásból meghatározott két p^* érték közötti eltérés csak a negyedik tizedes jegyben nyilvánul meg, így azt mondhatjuk, hogy esetünkben a két módszer ugyanarra az eredményre vezetett.

15.10. Rangpróbák

164. megjegyzés
Rangsor transzformációt ordinális változó esetén is végezhetünk, így ezek a próbák ilyen típusú változókra is alkalmazhatók.

Ezekben a próbákban az eloszlástól való függetlenséget úgy érjük el, hogy a jellemző statisztikát nem az eredeti minta elemeiből számoljuk, hanem azokon először rangsor transzformációt hajtunk végre (vö. 15.5. rész), és a rangokat használjuk az eredeti értékek helyett (164. megjegyzés).

165. megjegyzés
A szimmetrikussági feltételből következik, hogy a medián megegyezik a várható értékkel, így akár arra vonatkozóan is felírhatnánk a nullhipotézist.



Frank Wilcoxon (1892-1965) Írországban született amerikai kémikus és statisztikus, több statisztikai próba megalkotója.

166. megjegyzés
A folytonossági korrekció itt is alkalmazható: mindkét esetben a (129) összefüggéshez hasonló 0,5-es eltolással.

1. Wilcoxon-féle előjeles rangpróba, amellyel – úgy mint az előjel próbában – egy ismeretlen mediánt (*Me*-t) hasonlítunk össze egy hipotetikus értékkel (Me_0-lal), de a próba párosított minták különbségeire is alkalmazható (vö. 147b. megjegyzés).

Nullhipotézis: H_0: $Me = Me_0$.

Feltétel: A változó legyen folytonos, eloszlása pedig legyen szimmetrikus (165. megjegyzés).

Statisztika (I.): Több lehetőség is adódik, de mindegyiknek az az alapja, hogy a mintaelemek Me_0-tól való eltéréseinek abszolút értékein rangsor transzformációt hajtunk végre. Ezután az egyik lehetséges statisztika **az eredetileg pozitív eltérésekhez tartozó rangok összege**.

Nulleloszlás (I.): Nincs külön neve, és bonyolultsága miatt itt nem is részletezzük, de alkalmas számítógépes programok kiszámolják a megfelelő p^* valószínűségeket. Aszimptotikusan már nem túl nagy elemszámú mintákra is a

$$\mu = \frac{n(n+1)}{4}, \qquad \sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{130}$$

paraméterű normális eloszlással szokás közelíteni.

Statisztika (II.): Egy másik lehetőség az lehet, ha a „Statisztika (I)” szerinti rangsor transzformáció után a rangoknak az eredeti különbségeknek megfelelően előjelet adunk és kiszámítjuk ezen **előjeles rangok átlagát** ($\overline{R}$ -ot; a (27) összefüggés alapján) valamint korrigált szórását (s_R^*-ot; a (98) összefüggés alapján), majd ezekből határozzuk meg az alábbi statisztikát (vö. (119)):

$$t^* = \frac{\sqrt{n}\,\overline{R}}{s_R^*}. \tag{131}$$

Nulleloszlás (II.): Aszimptotikusan $\nu = (n - 1)$ szabadságfokú *t*-eloszlás (166. megjegyzés).

Mintafeladat

A kórház egyik szárnyának felújításakor a betegeket át kellett költöztetni egy másik szárnyba. A kérdés az volt, hogy van-e hatása a költözésnek a betegek mentális állapotára? Ennek eldöntésére az ápoló személyzet egy 10-es szubjektív skálán minősítette a betegek állapotát a költöztetés előtt és után. Az alábbi táblázatba foglalt eredmények (14 beteg adatai) alapján milyen következtetést vonhatunk le?

költözés előtt	9	8	1	5	2	8	5	8	5	9	9	6	3	7
költözés után	3	8	3	1	3	2	9	9	8	4	8	3	1	2
különbség	-6	0	2	-4	1	-6	4	1	3	-5	-1	-3	-2	-5

Megoldás: Mivel a mentális állapot 8 esetben romlott és az összegzett pontérték is jelentősen csökkent (21 ponttal), ezért azt gondolhatnánk, hogy a költöztetés rontotta a betegek mentális állapotát. Mivel a *t*-próba feltételei biztosan nem teljesülnek, ezért azt nem, de a Wilcoxon-féle előjeles rangpróbát alkalmazhatjuk.

H_0 (a különbségekre vonatkozóan)**: $Me = 0$; H_1 : $Me < 0$.** (Egyoldalú próbát használunk, hiszen a mentális állapot javulása nem okoz semmilyen kockázatot.) A megfelelő p^* valószínűséget egzakt módon megkaphatjuk a megfelelő számítógépes program segítségével: $p^* = 0{,}06$.

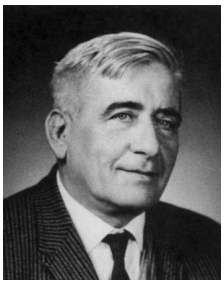
Amennyiben a (131) összefüggés szerint számolunk, akkor először a különbségek abszolút értékein rangsor transzformációt hajtunk végre, majd visszaírjuk a megfelelő előjeleket. (A 0-khoz nem rendelünk rangot, és a továbbiakban nem is vesszük figyelembe őket.)

különbség	1	1	-1	2	-2	3	-3	-4	4	-5	-5	-6	-6
előjeles rang (*R*)	2	2	-2	4,5	-4,5	6,5	-6,5	-8,5	8,5	-10,5	-10,5	-12,5	-12,5

Ezután t^* meghatározásához az előjeles rangok átlaga: $\overline{R} = -3{,}38$; korrigált szórásuk: $s_R^* = 7{,}45$; $n = 13$; tehát $t^* = -1{,}64$. A $\nu = 12$ szabadságfokú *t*-eloszlás alapján ehhez $p^* = 0{,}06$, amely az egzakt módon kapott értékkel megegyezik (eltérés csak a további tizedes jegyekben tapasztalható). 5%-os szignifikancia szinten tehát a nullhipotézist elfogadjuk. Mindezek alapján azt mondhatjuk, hogy a költöztetés szignifikánsan nem rontotta a betegek mentális állapotát.

Ilyen esetekben azonban, amikor p^* ennyire közel van a választott szignifikancia szinthez, igen körültekintően kell eljárnunk a következtetés levonásakor, hiszen egy kicsit nagyobb minta akár ellenkező eredményt is szolgáltathatott volna.

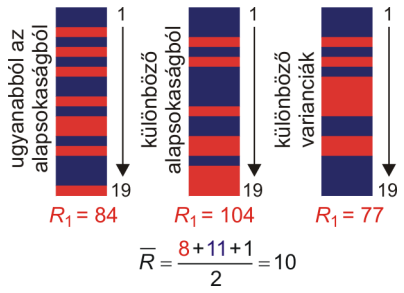
A következő részben ismertetendő próba kidolgozásában Wilcoxonnak szintén jelentős szerepe volt, de a félreértések elkerülése végett a próbát másként leíró statisztikusok neve alapján a Mann–Whitney-féle próba elnevezést használjuk.




Henry Berthold Mann (1905-2000) Bécsben született amerikai matematikus és statisztikus; valamint Donald Ransom Whitney (1915-2007) amerikai statisztikus.

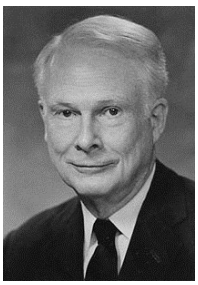
167. megjegyzés
Itt és az előző próbában is (v.ö (130)) a normális eloszlással való közelítés hátterében a centrális határeloszlás tétel érvényre jutása fedezhető fel (vö. 11.1. rész), hiszen mindkét statisztikában független valószínűségi változók (rangok) összege szerepel (vö. 78. ábra).

168. megjegyzés
Mivel a rangok eredetileg számtani sorozatot alkotnak 1-től N-ig, amelynek összege $N(N+1)/2$, ezért az átlag ennek N-ed része: $(N+1)/2$.



169. ábra
A rangok lehetséges eloszlásai az egyesített mintákban. (1. minta: piros; 2. minta: kék.)

170. megjegyzés
A (136) összefüggés már a folytonossági korrekciót is tartalmazza (vö. (129)).

William Henry Kruskal (1919-2005) amerikai matematikus és statisztikus; valamint
Wilson Allen Wallis (1912-1998) amerikai közgazdász és statisztikus.

171. megjegyzés
a) Egyszerűen belátható, hogyha R_{ji} helyére a rangok átlagát, $(N+1)/2$-t helyettesítünk, akkor H -ra éppen 0-t kapunk.
b) A χ^2-eloszlás megjelenése itt már nem meglepő, hiszen az R_j-k közelítőleg normális eloszlásúak (vö. 167. megjegyzés), és H-ban ezek négyzetösszege szerepel (vö. 14.3. rész).

2. Mann–Whitney-féle próba, amellyel a kétmintás t-próbához hasonlóan (vö. 15.6./2. rész), azt vizsgálhatjuk, hogy két minta származhat-e ugyanabból az alapsokaságból. Számításainkban azonban a rangpróbákra jellemző módon nem az eredeti minta elemeit, hanem az azokhoz rendelt rangokat használjuk.

Nullhipotézis: H_0: a két változó eloszlása megegyezik.

Feltétel: A két független változó legyen folytonos, sűrűségfüggvényeik pedig legyenek azonos alakúak (eltolással egymásba átvihetők). (Csak megjegyezzük, hogy ilyenkor a varianciák szükségképpen megegyeznek.)

Statisztika (I.): Itt szintén több lehetőségünk van. Az egyik az, hogy először a két mintát egyesítjük, ezután az eredeti értékeken rangsor transzformációt hajtunk végre, majd kiszámítjuk **az egyik mintához tartozó rangok összegét**:

$$\sum_{i=1}^{n_1} R_{1i} \text{ -t}, \qquad \text{vagy} \qquad \sum_{i=1}^{n_2} R_{2i} \text{ -t}. \tag{132}$$

Nulleloszlás (I.): Az U-eloszlás, amelyről csak annyit mondunk, hogy alkalmas számítógépes programmal kiszámolhatók a megfelelő p^* valószínűségek. Aszimptotikusan – például az 1-es mintára vonatkozóan – a

$$\mu = n_1 \frac{n_1 + n_2 + 1}{2}\,, \qquad \sigma = \sqrt{n_1 n_2 \frac{n_1 + n_2 + 1}{12}} \tag{133}$$

paraméterű normális eloszlással szokás közelíteni (167. megjegyzés).

Statisztika (II.): A nagyobb elemszámú mintákra vonatkozóan abból indulhatunk ki, hogy az átlagos rangszám ($\overline{R}$) mindig az első és az utolsó sorszám összegének a fele. Esetünkben:

$$\overline{R} = \frac{n_1 + n_2 + 1}{2} = \frac{N+1}{2}\,, \tag{134}$$

ahol n_1, illetve n_2 a két minta elemeinek száma, $N = n_1 + n_2$ az együttes elemszám (168. megjegyzés). Amennyiben a két minta ugyanabból az alapsokaságból származik (169. ábra), akkor a rangok összegének várhatóértéke mintánként az átlagos rangszám és az aktuális minta elemszámának a szorzata:

$$E\left(\sum_{i=1}^{n_1} R_{1i}\right) = n_1 \overline{R}\,, \qquad E\left(\sum_{i=1}^{n_2} R_{2i}\right) = n_2 \overline{R}\,. \tag{135}$$

Ezek szerint, ha az 1-es mintában a rangok összegén a (133) összefüggések szerinti standardizálási transzformációt hajtunk végre, akkor ez a statisztika:

$$\zeta = \frac{\sum_{i=1}^{n_1} R_{1i} - \mu + 0{,}5}{\sigma}\,. \tag{136}$$

Nulleloszlás (II.): Aszimptotikusan standard normális eloszlás, amely gyakorlatilag ekvivalens a fenti közelítéssel (170. megjegyzés).

3. Kruskal–Wallis-féle próba, amely az előző próbához hasonlóan, ismét a rangokat használva, kettőnél több (k) minta esetében teszi fel azt a kérdést, hogy a minták származhatnak-e ugyanabból az alapsokaságból.

Nullhipotézis: H_0: mindegyik változó eloszlása azonos.

Feltétel: Mind a k változóra megegyezik az előző próba feltételeivel.

Statisztika: Először az összes mintát egyesítjük, ezután az eredeti értékeken rangsor transzformációt hajtunk végre, majd az eredeti mintánként kiszámítjuk a hozzájuk tartozó rangok összegét (R_j-t): $R_j = \sum_{i=1}^{n_j} R_{ji}$.

Ebből a statisztika:

$$H = \frac{12 \sum_{j=1}^{k} \frac{R_j^2}{n_j}}{N(N+1)} - 3(N+1)\,, \qquad \text{ahol } N = \sum_{j=1}^{k} n_j\,. \tag{137}$$

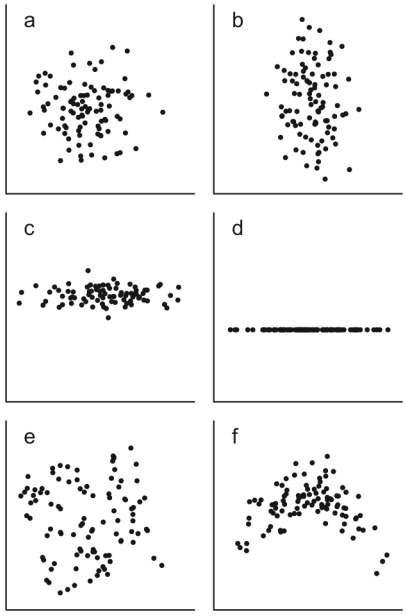
Nulleloszlás: Számítógépes program segítségével kiszámolhatók a megfelelő p^* valószínűségek. Aszimptotikusan $\nu = (k-1)$ szabadságfokú χ^2-eloszlás (171. megjegyzés).

16.0. Korreláció és regresszió számítás

172. megjegyzés
A két vagy több változó, két vagy több jellemző megfigyelését jelenti ugyanazokon a megfigyelési egységeken (vö. 83. ábra).

Ezt a problémakört a 12.2. részben már érintettük, amikor két valószínűségi változó függetlenségének vizsgálata után a függőség kritériumait tanulmányoztuk. Ott a témakörhöz kapcsolódó legfontosabb alapfogalmakat be is vezettük. Ebben a részben a gyakorlati szempontokat előtérbe helyezve kapcsolatot, összefüggést keresünk két vagy több változó között (172. megjegyzés).

A legáltalánosabb, bármilyen jellegű összefüggés az asszociáció. Az ilyen kapcsolat akár **nominális változók** között is fennállhat, amelynek hiányát például függetlenségvizsgálattal ellenőrizhetjük (vö. 15.8./3. rész). Amikor függetlenségvizsgálat céljából χ^2-próbát végzünk, a $\boldsymbol{\chi^2}$ mint statisztika éppen **az asszociáció mértékére jellemző érték**. Csak akkor vetjük el a „két változó független" nullhipotézist, ha ez az érték elég nagy. Mivel χ^2 nagysága nincs korlátozva, a különböző esetekben kapott értékek közvetlenül nem hasonlíthatók össze. Amennyiben alkalmas transzformációval χ^2-ből olyan mérőszámot nyerünk, amely például csak 0 és 1 között változik, akkor ezzel az asszociáció mértéke már általánosan, más esetekkel összehasonlítható módon is kifejezhető.



173. ábra
Korreláció mentes eseteket bemutató pontdiagramok ($r = 0$).

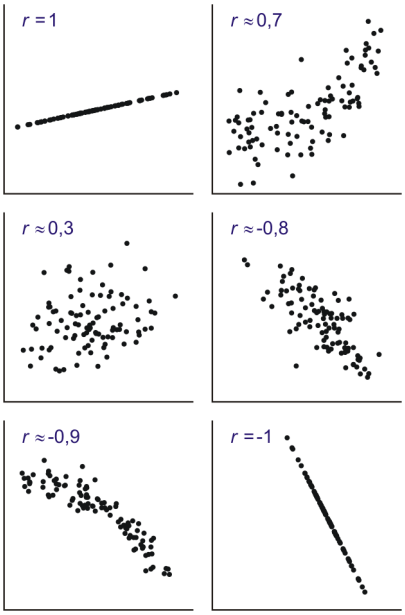
16.1. A korreláció mértéke, korrelációs *t*-próba

Két változó közötti **korrelációról** akkor beszélhetünk, ha azok **legalább ordinálisak**, azaz a mintákon belül meg tudunk állapítani valamilyen sorrendet. Ezenkívül feltesszük, hogy a két változó **szimmetrikus** viszonyban van egymással, tehát amennyiben van valamilyen összefüggés közöttük, akkor teljesen mindegy, hogy melyik a független és melyik a függő változó. A korreláció csak a **monotonitást** képes kifejezni; nevezetesen, ha a párosított mintákban a nagyobb értékekhez ***általában*** nagyobbak, a kisebb értékekhez pedig kisebbek tartoznak, akkor pozitív, fordított esetben negatív korrelációról beszélünk.

Számszerű adatok esetén a korreláció mértékét a **korrelációs együtthatóval** jellemezhetjük, amely –1 és +1 között változhat (vö. (81)(82)). A (82) összefüggés a következő módon is felírható:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} . \qquad (138)$$

A 173. ábrán olyan pontdiagramok láthatók, ahol a két változó között nincs korreláció, $r = 0$. Amint az a 173f. ábrán is megfigyelhető, a korrelálatlanság nem jelent feltétlenül függetlenséget. Itt ugyanis a két változó között van kapcsolat, de az nem monoton, (hiszen kezdetben növekvő, majd csökkenő tendenciát mutat) így nem beszélhetünk korrelációról sem. A 174. ábra a különböző mértékű korreláció néhány esetét mutatja be (monoton kapcsolatok). Az ábra arra is utal, hogy az $r = \pm 1$ akkor és csak akkor teljesül, ha a változók között lineáris kapcsolat van (vö. 12.2. rész). **Normális eloszlású változók** esetében, ha $r \neq 0$, akkor a kapcsolatok közül a **lineáris összefüggés** az egyedüli lehetőség, ha viszont $\boldsymbol{r = 0}$, akkor az egyértelműen **függetlenséget jelent** (vö. 81. megjegyzés).



174. ábra
Különböző mértékű korreláció bemutatása pontdiagramokon.

Hogyha a korreláció fennállását (lehet-e korreláció, vagy nem) az alapsokaságra vonatkozóan szeretnénk megvizsgálni, akkor ennek eldöntésére használhatjuk a korrelációs *t*-próbát:

Nullhipotézis: H_0: $R(\xi,\eta) = 0$, ahol $R(\xi,\eta)$ az alapsokaságra vonatkozó ismeretlen korrelációs együttható.

Feltétel: A két változó együttes eloszlása normális eloszlás. Minden megfigyelési egység legfeljebb egyszer van kiválasztva a mintákban és egymással nincsen kapcsolatuk (függetlenség).

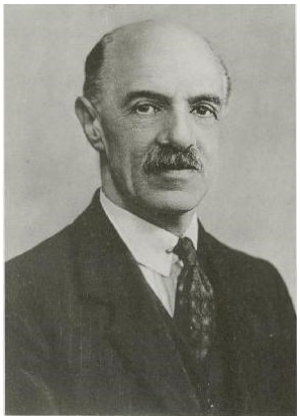
Statisztika: $$t^* = r\sqrt{\frac{n-2}{1-r^2}} , \qquad (139)$$

ahol r a mintából számolt korrelációs együttható, n pedig a minta elemszáma.

Nulleloszlás: $\nu = (n - 2)$ szabadságfokú t-eloszlás.

A döntés a t-próbáknál ismertetett módon történik (vö. 15.6. rész).

16.2. Rangkorreláció



Charles Edward Spearman (1863-1945) angol pszichológus és statisztikus.

Bár számszerű adatok esetén az r **korrelációs együttható**, mindig meghatározható, láthattuk, hogy valójában **csak a lineáris kapcsolatok jellemzésére alkalmas igazán** (vö. 16.1. rész). Például abszolút értéke még a legszorosabb nem lineáris kapcsolat esetén sem érheti el a maximális 1-et. Tudjuk, hogy r alkalmas az alapsokaságra vonatkozó korrelációs együttható, $R(\xi,\eta)$ becslésére, de például ordinális változók esetén nem is határozható meg. Így célszerű bevezetni egy másfajta statisztikát, olyat, amely ezeket a hiányosságokat kiküszöböli. Az egyik lehetőség a Spearman-féle rangkorrelációs együttható (r_S), amely a két változó közötti korreláció mértékét a rangok alapján jellemzi. Ennek meghatározása érdekében először a párosított minták eredeti (x_i, y_i) értékekein rangsor transzformációt hajtunk végre, majd az így kapott rangokat (R_{xi}, R_{yi}) az alábbi összefüggésbe helyettesítjük:

$$r_S = \frac{\sum_{i=1}^{n}(R_{xi}-\overline{R_x})(R_{yi}-\overline{R_y})}{\sqrt{\sum_{i=1}^{n}(R_{xi}-\overline{R_x})^2\sum_{i=1}^{n}(R_{yi}-\overline{R_y})^2}}. \qquad (140)$$

Látható, hogy ez az összefüggés megegyezik (138)-cal, amennyiben azt a rangokra alkalmazzuk (175. megjegyzés). A rangkorrelációs együttható (r_S), r-hez hasonlóan szintén –1 és +1 között változhat, és a jelentése is nagyon hasonló. A 0-hoz közeli értékek gyenge, a –1-hez közeliek erős negatív, a +1-hez közeliek erős pozitív korrelációs kapcsolatot jelentenek.

175. megjegyzés
A jobb megkülönböztethetőség kedvéért a (138) szerinti „közönséges” korrelációs együtthatót gyakran Pearson-féle korrelációs együtthatónak nevezik.

16.3. A regresszió jelentése a gyakorlatban

A regresszió szó eredeti értelmét tekintve a kapcsolatoknak azt a fajtáját jelenti, amikor valamit valamire **visszavezetünk**. Ebben az esetben az összefüggés nem szimmetrikus, a függő és független változó nem cserélhető fel (176. megjegyzés). Már az 1.1. részben megállapíthattuk, hogy általánosságban véve a „változások” determinisztikus és statisztikus része együtt van jelen. Egy modell alapján jósolt determinisztikus kapcsolatot általában egy függvénnyel írhatunk le, de a mérésekből, megfigyelésekből nyert adatok a megfelelő függvényértékektől mindig eltérnek. Emiatt az eredeti függvénykapcsolat „gyengített” formában jelenik meg, a független változó ugyan befolyásolhatja a függő változót, de sohasem határozza meg egyértelműen, hiszen a függő változó aktuális értéke a „véletlentől” is függ.

176. megjegyzés
A függő változót magyarázott, a független változót magyarázó változónak is szokás nevezni. Ezek az elnevezések arra utalnak, hogy a magyarázó változók változásai magyarázatot adhatnak a magyarázott változó változásaira. Más szavakkal mondva a független változó változásaira lehet **visszavezetni** a függő változó determinisztikus változásait, vagyis megtudhatjuk, hogy mi a változások eredete.

Általánosan fogalmazva a regresszió számítás feladata az, hogy **keressük meg azt a függvényt**, amely egy vagy több független és egy függő változó között teremt kapcsolatot és amely az említett „befolyásolásra” kellően jellemző. Erre az egyik lehetőség az, hogy **a determinisztikus és a statisztikus részt szétválasztjuk**, ami természetesen csak bizonyos feltételek teljesülése esetén tehető meg. Jelöljük a független változókat $X_1, X_2, \ldots, X_N$-nel, a függő változót pedig Y-nal. Általánosan felírható, hogy

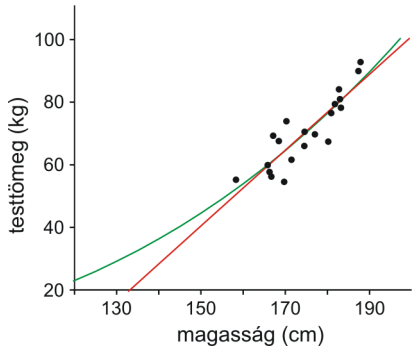
$$Y = f(X_1, X_2, \ldots, X_N) + H, \qquad (141)$$

ahol H a hibatag. Ebben az összefüggésben az X_i-k általában nem valószínűségi változók, vagy ha azok, akkor a megfelelő értékei elhanyagolhatóan kis hibával rendelkeznek. Így Y is csak az additív hibatag miatt válik valószínűségi változóvá. Ha a hibatag várhatóértéke 0 ($E(H) = 0$), akkor

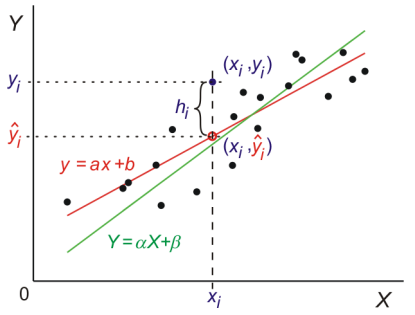
$$E(Y) = f(X_1, X_2, \ldots, X_N), \qquad (142)$$

azaz $f(X_1, X_2, \ldots, X_N)$ az Y várhatóértékével egyenlő. Ezek szerint az f függvény az Y **regressziója** (vö. 12.2. rész). Így amennyiben az X_i-k konkrét értékeit ismerjük, akkor a regresszió segítségével **becslést** adhatunk Y konkrét értékére is. A kérdés „már csak” az, hogy hogyan keressük meg a megfelelő f függvényt, ami természetesen nem mindig egyszerű és egyáltalán nem egyértelmű feladat.

Mielőtt nekilátnánk a függvény keresésének, mindenek előtt el kell döntenünk, hogy melyik változó a függő és melyek a függetlenek. Ennek érdekében azt kell megfontolnunk, hogy mi mire vezethető vissza, milyenek lehetnek a véletlen hatások, melyik értékeket tudjuk nagy pontossággal és melyeket tudjuk kevésbé pontosan meghatározni. Továbbá, hogy vannak-e ismereteink a függvényre vonatkozóan



177. ábra
A magasság és a testtömeg közötti kapcsolat pontdiagramja és az összefüggés két fajta regresszióval történő összekapcsolása.
Zöld görbe: köbös összefüggés; piros görbe: lineáris összefüggés. A két függvény görbéje alig tér el egymástól a vizsgált tartományon belül.



178. ábra
A „valódi” regressziós egyenes (zöld) és annak becslése (piros), azaz a pontokra legjobban illeszkedő egyenes. Látható, hogy a regresszió x_i-hez a becsült $f(x_i) = \hat{y}_i$-t rendeli hozzá.

179. megjegyzés
Ehhez hasonló problémával már találkoztunk a 10.0. rész 2. mintafeladatában, amikor az átlag minimum tulajdonságát bizonyítottuk. Ott is az eltérések négyzetösszegének minimumát kerestük. Az eljárás során az $Ax^2 + Bx + C = 0$ alakú másodfokú egyenletek megoldó képletéből olvastuk ki a megoldást:
$x_{\min} = -B/2A$.

a korábbi tapasztalatainkból és hogy mit mutat az adatok grafikus ábrázolása? Csak mindezek után tudunk javaslatot tenni a „legjobb” jelöltekre. A lineáris függvény egyszerűségénél fogva igen sokszor jön számításba, még akkor is, ha elvileg van nála alkalmasabb jelölt. A 177. ábra éppen ezt mutatja, hogy egy elvileg köbös összefüggés lineáris függvénnyel is igen jól közelíthető az adott tartományban. A háttérismeretek nélkül nem is dönthető el, hogy valójában melyik a „jobb” regresszió.

16.4. Lineáris regresszió

Ebben az esetben a keresett f függvény lineáris, ezért a (141) összefüggés alapján:

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_N X_N + \beta + H\,; \qquad (143)$$

vagy csak egyetlen független változóra felírva:

$$Y = \alpha X + \beta + H\,, \qquad (144)$$

ahol α a lineáris függvény meredeksége, β ugyanennek a 0-ban felvett függvényértéke, továbbá a korábbi jelölésnek megfelelően H a hibatag. A továbbiakban csak ezzel az egyszerűbb esettel foglalkozunk (egyszerű lineáris regresszió). Ilyenkor az f függvény grafikonja a regressziós egyenes.

Az f függvényt akkor használhatnánk arra, hogy az X konkrét értékei alapján Y konkrét értékeit közelítőleg megkapjuk, ha az α, β paraméterek értékeit ismernénk. Mivel **az α, β paramétereket eleve sohasem ismerjük**, így értéküket **csak becsülni tudjuk**, a becslés alapja pedig a rendelkezésünkre álló (x_i, y_i) adatpárok összessége, amelyet pontdiagramon tudunk szemléltetni. A feladat tehát az, hogy találjuk meg a pontdiagramra **legjobban illeszkedő egyenest** (178. ábra).

Ilyen egyenes többféleképpen is megadható. Ezek megtalálásához vezető egyik alapelv a legkisebb négyzetek módszere. Jelöljük az egyelőre ismeretlen, legjobban illeszkedő egyenes meredekségét a-val, tengelymetszetét pedig b-vel. Azokra az (x_i, y_i) koordinátájú pontokra, amelyeken az egyenes éppen átmegy: $y_i = ax_i + b = (\hat{y}_i)$; amelyeken nem megy át: $y_i \neq ax_i + b$. Az első esetben a regressziós egyenes hiba nélkül adja meg az első koordinátából a másodikat, a második esetben az eltérés, azaz a pontnak az illesztett egyenestől az y tengellyel párhuzamosan mért távolsága: $h_i = |y_i - (ax_i + b)|$ a reziduum vagy maradék, amellyel a hiba nagysága jellemezhető. A legkisebb négyzetek módszere alapján a becslés akkor a legjobb, ha a **reziduumok négyzetösszege minimális** (179. megjegyzés):

$$\sum_{i=1}^{n} h_i^2 = \min. \qquad (145)$$

A feltételnek eleget tevő a, b becsült paraméterekre a következő összefüggések adódnak :

$$a = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}\,, \qquad b = \bar{y} - a\bar{x}\,, \qquad (146)$$

ahol s_{xy} a kovarianciát, s_x, s_y a megfelelő szórásokat, r a korrelációs együtthatót, a felülvonás pedig az átlagokat jelöli (vö. (82), (42), (27)).

Mintafeladat

Mutassuk meg, hogy a (146/2) összefüggés a 178. megjegyzésben leírtak szerint egyszerűen meghatározható.

Megoldás: A (145) összefüggés szerint:

$$\sum_{i=1}^{n} h_i^2 = \sum\left[y_i - (ax_i + b)\right]^2 = \sum\left[y_i^2 + (a^2x_i^2 + b^2 + 2ax_ib) - 2y_iax_i - 2y_ib\right] \longrightarrow \underbrace{nb^2}_{A} + \underbrace{a2\sum x_i b - 2\sum y_i b}_{B}\,.$$

A nyíl után már csak a b-t is tartalmazó tagokat írtuk le. Tekintsük a $Ab^2 + Bb + C = 0$ másodfokú egyenletet, majd a megfelelő együtthatókat helyettesítsük a $b = -B/2A$ összefüggésbe:

$$b = \frac{2\sum y_i - a2\sum x_i}{2n} = \bar{y} - a\bar{x}\,.$$

Az egyszerűsítés elvégzése után a kívánt összefüggéshez jutottunk.

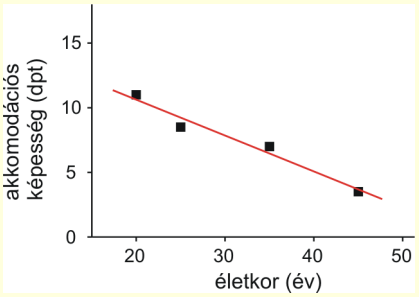
Hasonló módon járhatnánk el a (146/1) összefüggés esetében is, ha a négyzetre emelés elvégzése után az a-t tartalmazó tagokat vennénk figyelembe, de ott a számolás kicsit bonyolultabb, ezért annak bemutatásától eltekintünk.

Mintafeladat

A mellékelt táblázatban a normális szem átlagos akkomodációs képességét tüntettük fel négy különböző életkorban. Lineáris regresszió segítségével becsüljük meg az átlagos akkomodációs képességet 40 éves korban!

életkor (év)	20	25	35	45
akkomodációs	11	8,5	7	3,5

Megoldás: A regressziós egyenes becsült paraméterei a (146) összefüggés alapján: $a = -0{,}28$; $b = 16{,}2$. Az ábrázolás elvégzése után láthatjuk, hogy a lineáris összefüggés jól illeszkedik a megadott pontokra. Mivel nem ismerünk olyan matematikai modellt, törvényszerűséget, amely a két mennyiséget összekapcsolja, ezért a paramétereket csak interpolációs, azaz csak az ábrázolt intervallumon belüli becslésre használhatjuk. Az akkomodációs képesség becsült értékét 40 éves korban tehát $-0{,}28 \cdot 40 + 16{,}2 \approx 5$ (dpt).



Az, hogy a valódi összefüggés nem lehet lineáris, abból is kiderül, hogy biztosak lehetünk abban, hogy még 70 év felett sem kaphatunk negatív értéket, amit egyébként az összefüggés jósolna.

A regresszió során elkövetett hiba (H) jellemzésére a reziduális szórást használjuk. Ennek az alapsokaságra vonatkozó korrigált változata (vö. (145)):

$$s_h^* = \sqrt{\frac{\sum_{i=1}^{n} h_i}{n-2}} \,. \qquad (147)$$

A becsült paraméterek szórása is erre vezethető vissza:

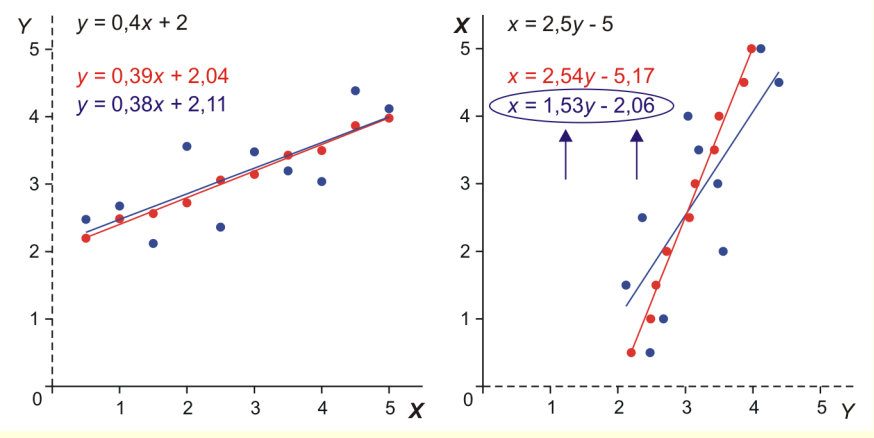
$$s_a^* = s_h^* \sqrt{\frac{1}{n s_x^2}} \,, \qquad s_b^* = s_h^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}} \,. \qquad (148)$$

Mintafeladat

Egyszerű példán keresztül mutassuk meg, hogy lineáris regresszió esetén a két változó nincs szimmetrikus viszonyban egymással, tehát egyáltalán nem mindegy, hogy melyik a független és melyik a függő változó.

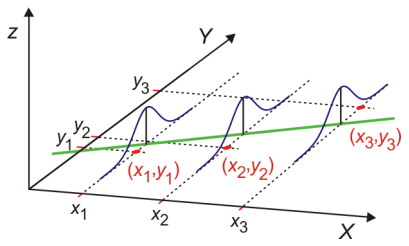
Megoldás: Tegyük fel, hogy az „eredeti” ($X \to Y$) összefüggés $y = 0{,}4x + 2$ alakban írható fel. Ennek felhasználásával számítsuk ki néhány x értékhez tartozó y értékét, majd ezekhez különböző mértékű (1×, 10×) véletlen hiba hozzáadásával ($y + h$) generáljunk két pontdiagramot (baloldali ábra: piros pontok (1×); kék pontok (10×)). Ezután a legkisebb négyzetek módszerének alkalmazásával becsüljük meg a legjobban illeszkedő egyenesek paramétereit és rajzoljuk be az illesztett egyeneseket. Láthatjuk, hogy mindkét egyenes igen jól illeszkedik a pontokra és a becsült paraméterekben is alig van eltérés, mindegyik jól közelíti a „valódi” értékeket.

Ezt követően cseréljük fel a tengelyeket a pontdiagramokkal együtt (jobboldali ábra) és végezzük el újra az illesztéseket. Ennek eredményeként az „eredetiből” kifejezett „fordított” ($Y \to X$) $x = 2{,}5y - 5$ lineáris összefüggéshez kellene jutnunk. Ha a hiba nem túl nagy (1×) az illeszkedés egészen jó és a becsült paraméterek jól közelítik a „valódi” értékeket. Ha azonban a hiba nő a helyzet drasztikusan megváltozik: az illeszkedés egyre rosszabb és a becsült paraméterek (kék nyilakkal jelölve) egyáltalán nem közelítik a „valódi” értékeket.



16.5. A lineáris regresszióval kapcsolatos hipotézisvizsgálatok

Az alapsokaságra vonatkozóan igen fontos kérdés az, hogy Y valóban függ-e X-től, amit például a regressziós egyenes meredekségének vizsgálatával dönthetünk el. Amennyiben $\alpha = 0$, és ezt az adatok nem cáfolják, akkor a regresszió nem szignifikáns. Azt is mondhatjuk, hogy a függést az adatok nem támasztják alá.



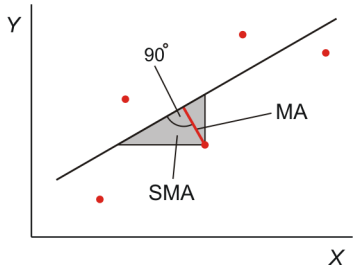
180. ábra
A lineáris regresszióra, illetve a regressziós egyenes meredekségére vonatkozó *t*-próba alkalmazhatósági feltételeinek szemléltetése.

181. megjegyzés
Ez a próba egyetlen független változó esetén ekvivalens az előzővel. Így igazi jelentősége akkor van, ha egyszerre több független változót vizsgálunk (többszörös lineáris regresszió), de erre itt részletekbe menően nem térünk ki.

182. megjegyzés
Többszörös lineáris regresszió esetén a (152) összefüggés módosul:

$$F = \frac{SS_R}{SS_H}\frac{n-k-1}{k},$$

ahol k a független változók száma. Ilyenkor a nulleloszlás a $\nu_1 = k$; $\nu_2 = (n-k-1)$ szabadságfokú F-eloszlás.



183. ábra
A lineáris regresszió olyan esetei, amikor nem a reziduumok négyzetösszegét minimalizáljuk. MA-regresszió esetén a pontoknak az egyenestől mért (merőleges) távolságainak négyzetösszegét, SMA-regresszió esetén pedig az ábrán látható szürke háromszögnek megfelelő hasonló területek összegét kell minimalizálni.

184. megjegyzés
Van azért olyan probléma is, amelynek megoldására egyik módszer sem igazán jó. Két hasonló mérési eljárás közötti **egyezés** vizsgálatára például sem a korreláció, sem a regresszió számítás nem alkalmas. Ezek használata helyett legjobb, ha a különbségek varianciáját elemezzük.

1. *t*-próba a meredekségre:

Nullhipotézis: $H_0 : \alpha = 0$.

Feltételek: X és Y között lineáris a kapcsolat. X a független változó elhanyagolható hibával (nem feltétlenül valószínűségi változó). A H hibatag mint valószínűségi változó (az Y értékek hibája) X-től független, normális eloszlású és állandó szórású. (A 180. ábra ezeket a feltételeket illusztrálja.)

Statisztika:
$$t^* = \frac{a}{s_a^*}. \qquad (149)$$

Nulleloszlás: $\nu = (n-2)$ szabadságfokú t-eloszlás.

2. *Y X*-től való függésének elemzése (181. megjegyzés), amely azon alapul, hogy Y varianciáját (pontosabban eltérés-négyzetösszegét) két részre bonthatjuk: az X-től eredő determinisztikus részre, és a H hibatagból származó statisztikusra. Mivel feltesszük, hogy ezek függetlenek egymástól, ezért a felbontás a (75) összefüggés alapján egyszerű összegzés (vö. (75) és a 14.1. rész mintafeladata):

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \qquad (150)$$

ahol a korábbi jelölésnek megfelelően $\hat{y}_i = f(x_i)$. Rövidebben írva a szokásos jelölésekkel:

$$SS_T = SS_R + SS_H, \qquad (151)$$

ahol az SS (sum of squares) a négyzetösszeget jelenti, az alsó indexek pedig rendre a teljesre (T), a regresszióra (R) és a hibatagra (H) utalnak.

Nullhipotézis: $H_0 :$ Y nem függ X-től.

Feltételek: A hiba itt is X-től független, normális eloszlású és állandó szórású.

Statisztika:
$$F = \frac{SS_R}{SS_H}(n-2). \qquad (152)$$

Nulleloszlás: $\nu_1 = 1$; $\nu_2 = (n-2)$ szabadságfokú F-eloszlás (182. megjegyzés).

Itt említjük meg, hogy a (150) illetve (151) összefüggés azt is megmutatja, hogy Y teljes varianciája hogyan oszlik meg a regresszió és a hiba között. Ennek alapján azt mondhatjuk, hogy az SS_R/SS_T arány azt adja meg, hogy a teljes varianciának hányad része tulajdonítható a regressziónak, illetve, hogy X mint az összefüggés magyarázó változója milyen mértékben magyarázza Y változásainak eredetét. Ez a hányados a determinációs együttható, amelyről belátható, hogy épp a korrelációs együttható négyzetével egyenlő:

$$\frac{SS_R}{SS_T} = 1 - \frac{SS_H}{SS_T} = r^2. \qquad (153)$$

16.6. A lineáris regresszió bonyolultabb esetei

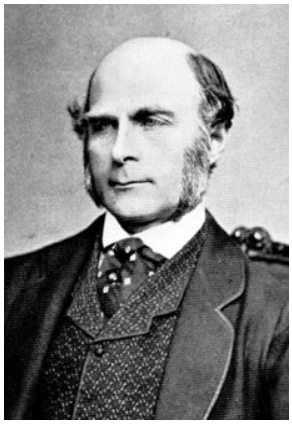
A 16.3. részben a (141) összefüggés felírásakor abból indultunk ki, hogy az X nem feltétlenül valószínűségi változó, illetve, ha az, akkor elhanyagolhatóan kis hibával rendelkezik. Vannak azonban olyan esetek is, amikor X igazi valószínűségi változó és hibája összemérhető Y-éval. Ilyenkor a legkisebb négyzetek módszerét nem a reziduumokra alkalmazzuk, hanem a pontdiagram pontjainak az illesztendő egyenestől mért (merőleges) távolságaira (183. ábra) (Ez az MA-regresszió (major axis)).

Egy másfajta illesztési módszert használunk akkor, ha mindkét változó (X és Y) szórása nagyjából a neki megfelelő hiba szórásával egyezik meg. Ekkor a pontdiagram pontjainak, az illesztendő egyeneshez, a tengelyekkel párhuzamosan húzott egyenes szakaszai és maga az egyenes által határolt derékszögű háromszögek területét kell minimalizálni (183. ábra) (Ez az SMA-regresszió (standard major axis)). Ezeket az esetek azonban részletesen nem tárgyaljuk.

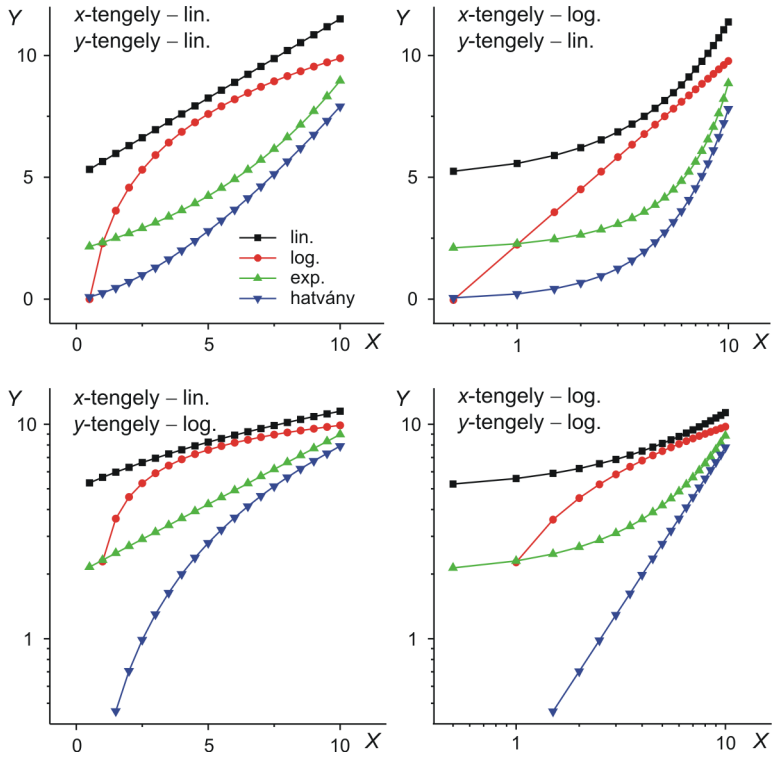
Az előzőkből úgy tűnhet, hogy a lineáris regresszió és a korreláció mégsem annyira különböző dolog. Valóban, adott kérdések megválaszolásához sokszor mindkét út járható. Azt kell csak eldöntenünk, hogy melyik típusú megfogalmazás tükrözi jobban céljainkat (184. megjegyzés).

16.7. Lineárisra visszavezethető nem lineáris regressziók

Néhány elemi függvény megfelelő logaritmikus transzformáció alkalmazásával lineáris függvénnyé alakítható. Ezek a függvények igen gyakran fordulnak elő a modellekben is (185. ábra).



Francis Galton (1822-1911) angol polihisztor, aki jelentősen hozzájárult a statisztika fejlődéséhez is, elsősorban a regresszió számítás témakörében.



185. ábra
Négy elemi függvény (lineáris-, logaritmus-, exponenciális- és hatványfüggvény) ábrázolása különböző skálabeosztású koordináta rendszerekben. Látható, hogy a megfelelő logaritmikus transzformáció ezeket a nemlineáris függvényeket „kiegyenesíti”.

186. megjegyzés
Ha a H hiba mint valószínűségi változó például normális eloszlású, akkor az e^H ún. log-normális eloszlású lesz.

A transzformáció a regressziós függvény mellett érintheti a hibát is, amennyiben a függő változót is transzformáljuk. Az eddigiekben a hibáról feltettük, hogy additív (vö. (141)), tehát a függő változóhoz hozzáadódik. Bizonyos esetekben a hiba szorzótényezőként jelenik meg, ilyenkor beszélünk multiplikatív hibáról. Lássuk ezek után a három legegyszerűbb esetet. (A transzformált változót ’-vel jelöljük.)

1. Ha $Y = \alpha \log X + \beta + H$ (logaritmus függvény additív hibával),
akkor ($X' \to Y$), ahol $X' = \log X$, tehát

$$Y = \alpha X' + \beta + H\,. \qquad (154)$$

2. Ha $Y = e^{\alpha X} e^{\beta} e^{H}$ (exponenciális függvény multiplikatív hibával),
akkor ($X \to Y'$), ahol $Y' = \ln Y$, azaz $\ln Y = \alpha X + \beta + H$, tehát

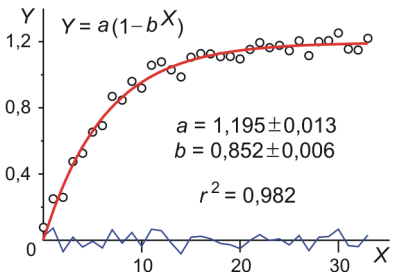
$$Y' = \alpha X + \beta + H\,. \qquad (155)$$

3. Ha $Y = X^{\alpha} e^{\beta} e^{H}$ (hatványfüggvény multiplikatív hibával),
akkor ($X' \to Y'$), ahol $X' = \ln X$, $Y' = \ln Y$, azaz $\ln Y = \alpha \ln X + \beta + H$, tehát

$$Y' = \alpha X' + \beta + H\,. \qquad (156)$$

A megfelelő logaritmikus transzformáció megválasztásával mindhárom esetben lineáris függvényhez jutottunk. Ilyenkor a regressziós eredményeket (próbák, r^2 stb.) ugyanúgy kell értelmezni, mint a lineáris esetben, de természetesen a transzformált változókra vonatkoztatva. Az eredményeket vissza szoktuk transzformálni az eredeti változókra, aminek az a következménye, hogy ezáltal a változók eloszlástípusa is megváltozik (186. megjegyzés).

A 187. ábrán azt mutatjuk be, hogy az $(y_i - f(x_i))$ reziduumok négyzetösszegének minimalizálása akkor is célravezető, ha az illesztendő függvény nem transzformálható lineáris függvénnyé.



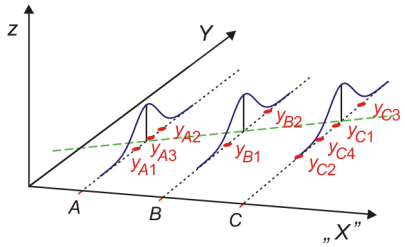
187. ábra
Az $Y = \alpha(1-\beta^X)$ függvény illesztése (piros görbe), a reziduumokra (kék görbe) vonatkozó legkisebb négyzetek módszerével. Bár az illesztendő függvény paramétereit – úgy mint a lineáris regressziónál – összefüggések formájában nem tudjuk megadni, számítógép segítségével azért becslést adhatunk rájuk, és a determinációs együttható is meghatározható.

17.0. Varianciaelemzés (varianciaanalízis)

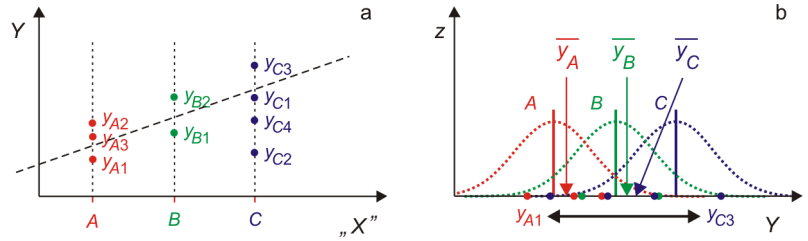
188. megjegyzés
A varianciaelemzéssel megoldható problémák megfogalmazásakor a szóhasználat eltérhet az eddigiektől. Első lépésként kiválasztjuk a megfigyelési egységeket, majd ezeket a kategoriális független változóknak megfelelően csoportokba osztjuk. A csoportok például abban különbözhetnek egymástól, hogy csoportonként a megfigyelési egységeket más-más kezelésnek vetjük alá. Így az adott kezelés típusa lehet az a tényező, „független változó", amely megkülönbözteti a csoportokat. Miután a függő változó konkrét értékeit megállapítottuk az egyes megfigyelési egységeken, a csoportok már mintaként kezelhetők (vö. 13.0. rész).

Elöljáróban a módszerrel kapcsolatosan két félreértést kell tisztázni. Az egyik az, hogy a neve alapján (angolul: ANalysis Of VAriance = ANOVA) azt gondolhatná az ember, hogy ez valami igen bonyolult eljárás. A következőkben látni fogjuk, hogy a varianciaelemzés legegyszerűbb változata nem több mint egy *F*-próba (vö. 15.7. rész). A másik félrevezető dolog az, hogy bár a módszer nevében a variancia szerepel, valójában várhatóértékek összehasonlítására használható.

Az egyik megfogalmazás szerint azt mondhatjuk, hogy a varianciaelemzés a kétmintás *t*-próbával ellenőrizhető kérdéstípus több mintára is kiterjeszthető megválaszolására alkalmas (vö. 15.6. rész). Más megfogalmazás szerint a varianciaelemzés a regresszió számítással van rokonságban. Ezt úgy képzelhetjük el, hogy a független változók itt kategoriálisak, és nem változnak megfigyelési egységenként, csak mintánként. Ezek a „független változók" lehetnek hatással a megfigyelési egységenként is különböző értékű függő változó várhatóértékére, ami a mintaátlagok különbözőségét eredményezheti (188. megjegyzés). Az elmondottakat a 189. és 190. ábrán szemléltetjük.

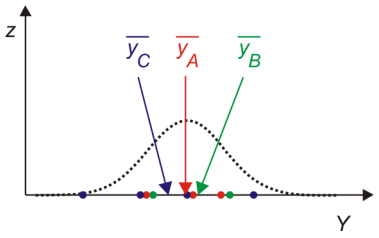


189. ábra
Első lépésként a 180. ábrát úgy módosítottuk, hogy a független változó kategoriális volta legyen szembetűnő (*A*, *B*, *C* kategóriák). Mivel a legtöbb esetben a kategóriákra való áttérés után a sorrendnek már nincs igazán jelentősége, *X* elvesztette eredeti jelentését, amit az ábrán az idézőjel érzékeltet („*X*").



190. ábra
a) A 189. ábra *z* irányú vetülete, ahol jobban látható, hogy egy kategórián belül több megfigyelési egységen is meghatározzuk a függő változó értékét. Itt már az azonos mintához tatozó elemeket azonos színnel jelöltük.

b) A 189. ábra „*X*" irányú vetülete, amelyen az egyes kategóriákhoz tartozó alapsokaság ismeretlen sűrűségfüggvénye (pontozott vonal), illetve a megfelelő mintaelemek, valamint azok átlaga van feltüntetve. Itt látható a módszer alkalmazhatósági feltételei közül kettő, nevezetesen minden kategóriához tartozó alapsokaság legyen normális eloszlású, és legyen azonos a szórásuk is. Eltérés csak a várható értékükben lehet, amelyek becsült értékét a mintaátlagok jelenítik meg.



191. ábra
A 190b. ábrával ellentétben itt egyetlen közös alapsokaságból származnak a minták. Ennek feltételezett sűrűségfüggvényét szemlélteti a pontozott vonal.

A módszerrel tehát a várhatóértékek közötti különbséget tudjuk vizsgálni. Ennek lényegét egyszerűen úgy fogalmazhatjuk meg, hogy **amennyiben a mintaátlagok szóródása nagyobb, mint az egyes mintákon belül a mintaelemek szóródása, akkor azt mondhatjuk, hogy a minták különböznek egymástól** (különböző várhatóértékű alapsokaságból származnak), vagy legalábbis nem mind származnak ugyanabból az alapsokaságból (190b. ábra). Megfordítva, ha a mintaátlagok szóródása kisebb, mint az egyes mintákon belül a mintaelemek szóródása, akkor nem mondhatjuk azt, hogy a minták különböznek egymástól, tehát akár azonos alapsokaságból is származhatnak, így várhatóértékük is megegyezik (191. ábra).

Ha innen gondolatban visszaugrunk a 189. ábrának megfelelő ábrázolásra akkor ez utóbbi megállapítás a lineáris regresszió „nyelvén" azt jelenti, hogy $\alpha = 0$, tehát *Y* nem függ „*X*"-től. Amennyiben a vizsgálat kezelésekre vonatkozott, akkor azt mondhatjuk, hogy nincs különbség közöttük.

A fenti egyszerűsített megfogalmazás a következőképpen pontosítható. Az eljárás során a függő változó változásait két részre bontjuk. Az egyik rész a determinisztikus hatásnak (például a kezelések hatásának), a másik a véletlennek tulajdonítható. Mivel feltesszük, hogy ezek függetlenek egymástól ezért varianciájuk (így eltérés-négyzetösszegük is) összegződik (vö. (150)):

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2, \qquad (157)$$

ahol *k* a minták száma, n_i az *i*-edik minta elemszáma, y_{ij} az *i*-edik minta *j*-edik eleme, $\bar{y}$ az összes elemből számolt átlag, míg $\bar{y}_i$ az *i*-edik minta átlaga. Az első összegzést (*j*-re) a mintán belül, a másodikat (*i*-re) a minták között kell elvégezni. A (150) összefüggéshez képest annyi a különbség, hogy a regresszió számítás becsült értékének ($\hat{y}$) itt a mintaátlag felel meg ($\bar{y}_i$).

192. megjegyzés
Amennyiben a csoportok abban különböznek, hogy különböző kezelésnek lettek alávetve, akkor a klasszikus megoldás a kiegyensúlyozott kísérleti elrendezés, amikor a kezeléseket ugyanannyi megfigyelési egységen alkalmazzuk, tehát minden csoport azonos elemszámú (n). Ilyenkor $N = kn$.

Ha az előbbi eltérés-négyzetösszeg felbontást a (151) összefüggéshez hasonló jelölésekkel írjuk fel, akkor:

$$SS_T = SS_K + SS_H, \qquad (158)$$

ahol az SS (sum of squares) itt is a négyzetösszeget jelenti, az alsó indexek pedig rendre a teljesre (T), a minták közöttire (K) és a hibatagra (H) utalnak.

Általánosan igaz, hogy : $\sum_{i=1}^{k} n_i = N$ (192. megjegyzés).

Az eltérések négyzetösszegéből úgy kapunk varianciát, hogy elosztjuk őket a szabadságfokukkal. A négyzetösszeg szabadságfoka független tagok esetén a tagok száma, egyébként pedig a tagok száma mínusz a becsült paraméterek száma (vö. 14.2. rész). Ennek alapján elkészíthetjük a varianciatáblát a szokásos jelölésekkel:

A variancia eredete	Eltérés-négyzetösszeg	Szabadságfok	Variancia
Teljes	SS_T	$N-1$	$MS_T = \frac{SS_T}{N-1}$
Minták közötti	SS_K	$k-1$	$MS_K = \frac{SS_K}{k-1}$
Mintán belüli (Hiba)	SS_H	$N-\mathrm{k}$	$MS_H = \frac{SS_H}{N-k}$

193. táblázat
Összesítés a négyzetösszegek felbontásáról és a varianciákról (ANOVA-tábla).

A „determinisztikus” (Minták közötti) és a statisztikus (Mintán belüli) komponenst F-próbával hasonlítjuk össze. Ha a determinisztikus rész szignifikánsan nagyobbnak bizonyul, mint a véletlennek tulajdonítható rész, akkor azt gondolhatjuk, hogy a vizsgált tényezőnek van hatása, (a minták nem mind származnak ugyanabból az alapsokaságból).

194. megjegyzés
1. Ha $k = 2$, a varianciaelemzés ekvivalens a kétmintás t-próbával.
2. Az alkalmazhatósági feltételek ellenőrzésére inkább a grafikus módszereket használjuk, mert kis minta-elemszámok esetén a megfelelő próbák kevéssé érzékenyek.
3. Amennyiben a feltételek fennállását mégis próbákkal szeretnénk eldönteni, akkor a normális eloszlás ellenőrzésére jól használható a Shapiro–Wilk-próba, több variancia azonosságának ellenőrzésére pedig a Levene-próba vagy a Bartlett-próba, amelyeket itt nem részletezünk. Nem árt tudni, hogy míg a Bartlett-próba csak normális eloszlású változók esetében használható, a Levene-próbára vonatkozóan ez nincs előírva. A feltétel teljesülése esetén viszont a Bartlett-próba a hatékonyabb.

Varianciaelemzés, ahogy a próbáknál már megszokhattuk (194. megjegyzés).

Nullhipotézis: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$.

Feltételek: A vizsgált változó mind a k csoportban legyen azonos varianciájú normális eloszlású és az egyes megfigyelések legyenek egymástól függetlenek.

Statisztika: $$F = \frac{MS_K}{MS_H} = \frac{SS_K}{SS_H}\frac{N-k}{k-1}, \qquad (159)$$

ahol MS_K, MS_H, SS_K, SS_H jelentése a 193. táblázatból kiolvasható, továbbá értékük a (157) és a (158) összefüggések segítségével kiszámítható (vö. (152) és a 182. megjegyzés).

Nulleloszlás: $\nu_1 = (k-1)$; $\nu_2 = (N-k)$ szabadságfokú F-eloszlás.

Mintafeladat

A mellékelt táblázatban 3 különböző régióban élő férfiak testmagasság adatait (cm) tüntettük fel. Van-e különbség e régiókban a testmagasságok várható értékei között?

1. régió	173	175	168	169	
2. régió	170	163	165		
3. régió	175	174	171	172	172

Megoldás: Használjuk a varianciaelemzést. (A feltételek teljesülését ellenőrizni kell.)

$\boldsymbol{H_0}$: A régiók között nincs különbség, a minták azonos alapsokaságból származnak, tehát $\mu_1 = \mu_2 = \mu_3$.

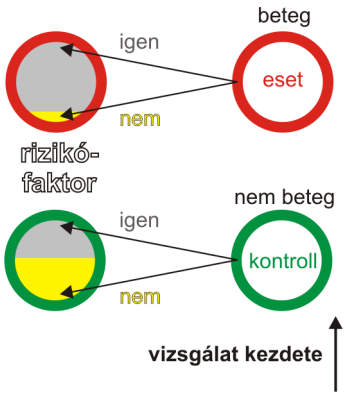
Először meghatározzuk a feladatnak megfelelő ANOVA-tábla (vö. 193. táblázat) számításhoz szükséges sorait a megadott összefüggések segítségével. Ezekből $F^* = 5{,}78$; a hozzátartozó $p^* = 0{,}024$; amelyet a megfelelő számítógépes program segítségével kaphatunk meg.

	SS	ν	Var
Minták közötti	89,37	2	44,68
Mintán belüli	69,55	9	7,73

5%-os szignifikancia szinten tehát a nullhipotézist elvetjük. (Amennyiben az F táblázatot használjuk, az 5%-hoz és a megfelelő szabadságfokokhoz (2;9) tartozó kritikus érték $F = 4{,}26$.) A különbség szignifikáns ezen a szinten.

rizikófaktor jelen van	beteg	
	igen	nem
igen	*a*	*b*
nem	*c*	*d*

195. táblázat
Általános, 2·2-es táblázat egy rizikófaktor szerepének vizsgálatára.



196. ábra
A visszatekintő vizsgálat sémája.

18.0. Néhány módszer az orvosi statisztika köréből

Gyakran merül fel az a kérdés, hogy egyes **rizikó faktorok** (például elhízás, dohányzás, alkohol fogyasztás, védőoltás elmulasztása) milyen szerepet játszanak a betegségek kialakulásában, előfordulásában. A vizsgálatok során gyakorisági adatokat gyűjtünk két mintához és az adatokat egy 2·2-es táblázatban rendezzük el (195. táblázat). Az *a*, *b*, *c*, *d* betűk a megfelelő abszolút gyakorisági adatokat jelölik. Azt várjuk, hogy amennyiben a gyakorisági értékek arányai a beteg és nem beteg csoportban azonosak, akkor a betegség kialakulása független a rizikófaktortól. A körülményektől függően két alapvető vizsgálati módszert végezhetünk.

18.1. Visszatekintő (eset-kontroll) vizsgálat

Ez a gyakoribb és egyszerűbb vizsgálati módszer, amelynek elvégzése például korábbi kórlapok adatai, továbbá kérdőíveken feltett kérdések válaszai, valamint már elvégzett orvosi vizsgálati eredmények alapján is történhet.

Ilyen vizsgálatnál a betegek és nem betegek csoportjában visszamenőleg nézzük meg annak esélyét, hogy egy rizikófaktor hatással lehetett-e a betegség kialakulására (196. ábra).

Példaként vegyünk egy lehetséges konkrét vizsgálatot, amelynek célja ***az influenza elleni védőoltások eredményességének meghatározása***. Első lépésként kiválasztjuk az **„eseteket”**, azaz meghatározzuk azoknak a személyeknek a csoportját, akik influenzában betegedtek meg az adott szezonban. A kiválasztás történhet kérdőív alapján, de ezt laboratóriumi vizsgálattal is illik megerősíteni. Második lépésként kiválasztjuk a **„kontroll”** csoportot, olyan személyeket, akiknél az influenza kimutatására irányuló laboratóriumi vizsgálat negatív eredményt adott. Ezt követően mindkét csoportban meghatározzuk az influenza elleni védőoltásban részesültek számát, majd ezeket a gyakorisági adatokat a 195. táblázat szerint összegezzük.

A megválaszolandó általános kérdés tehát az, hogy ***lehet-e a betegség kialakulásában szerepet játszó tényező az adott rizikófaktor***, nevezetesen az oltás elmulasztása?

Nullhipotézis: H_0: nincs összefüggés a rizikófaktor jelenléte (az oltás elmulasztása) és a betegség kialakulása között.

Hangsúlyozzuk tehát, hogy ebben az esetben ***beteg és nem beteg embereket*** választanak ki és őket osztják további két csoportra, a ***rizikófaktor megléte vagy hiánya*** alapján (197. táblázat).

	beteg		
rizikófaktor (védőoltás elmulasztása)	igen (eset)	nem (kontroll)	összesen
igen (nem oltott)	*a*	*b*	*a*+*b*
nem (oltott)	*c*	*d*	*c*+*d*
összesen	*a*+*c*	*b*+*d*	*n*=*a*+*b*+*c*+*d*

197. táblázat
Az influenza előfordulásának gyakorisági adatai a nem oltottak és az oltottak között.

Ezután meghatározzuk a **betegség esélyét** mindkét csoportban (vö. 3.4. rész):

A betegség esélye a rizikófaktor ***megléte*** esetében: a/b.
A betegség esélye a rizikófaktor ***hiánya*** esetében: c/d.

Ezekből kiszámoljuk az esélyhányadost (*OR*, odds ratio):

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}, \qquad (160)$$

amely megadja, hogy ***hányszor nagyobb a betegség kialakulásának esélye*** a rizikófaktor fennállása, (azaz a védőoltás elmulasztása esetén), mint annak hiányában, (azaz, ha a védőoltás beadásra került). Ezek szerint amennyiben H_0 igaz, akkor $OR = 1$.

Következő lépésként meghatározzuk *OR* konfidencia intervallumát. A számítás bonyolultsága miatt (amit itt nem részletezünk) először kiszámítjuk az ln(*OR*)

standard hibáját az alábbi összefüggés szerint:

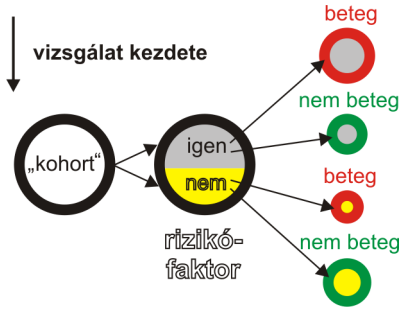
$$SE(\ln OR) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,. \qquad (161)$$

5%-os szignifikancia szint választása esetén ln*OR* konfidencia intervallumát a standard hibája ±1,96-szorosával jelölhetjük ki: $\ln OR \pm 1{,}96\,SE(\ln OR)$ (vö. 115. megjegyzés). (A visszatranszformálás elvégzése után az *OR* konfidencia intervalluma *OR*-re nézve már nem szimmetrikus.) Ha ez az intervallum magába foglalja az 1 értéket, akkor megtartjuk a nullhipotézist, ilyenkor azt mondhatjuk, hogy a betegség esélyét nem növeli az adott rizikófaktor; ellenkező esetben viszont elvetjük a nullhipotézist (az adott szignifikancia szinten).

18.2. Előretekintő, követéses (kohort) vizsgálat

198. megjegyzés
Kohortoknak nevezzük az olyan embercsoportokat (a „cohors" római hadszervezeti egység után), amelyek valamilyen tulajdonságukban közösek, és amelyeket hosszabb időn keresztül szándékozunk megfigyelni, adataikat nyomon követni.

Ebben az esetben például 18 éves egészségesnek minősített személyek csoportját (198. megjegyzés) további két csoportra osztjuk a **rizikófaktor megléte**, illetve **hiánya** alapján. A két csoportot hosszabb időn keresztül (akár több éven át) megfigyeljük, majd a vizsgálat végén a két minta gyakorisági adatait a 195. táblázat szerint összegezzük. Ilyenkor tehát a visszatekintő vizsgálattal ellentétben, a vizsgálat kezdetén még nem tudjuk, hogy ki beteg és ki nem (199. ábra).



199. ábra
Az előretekintő, követéses vizsgálat sémája.

Példaként ismét vegyünk egy lehetséges konkrét kérdést: ***Mekkora kockázatot jelent a dohányzás*** (mint rizikófaktor), ***a szívinfarktus*** (mint betegség) ***kialakulása szempontjából?***

Nullhipotézis: H_0: nincs összefüggés a dohányzás és a szívinfarktus kialakulása között.

A vizsgálandó személyek két csoportra osztása most a ***dohányzás***, ***nem dohányzás*** szerint történik. Adott idő elteltével, mondjuk 10 év múlva meghatározzuk az infarktus előfordulási gyakoriságát a két mintában, majd ezeket az adatokat beírjuk a táblázatba (200. táblázat).

	beteg (infarktus)		
rizikófaktor (dohányzik)	igen	nem	összesen
igen	*a*	*b*	*a*+*b*
nem	*c*	*d*	*c*+*d*
összesen	*a*+*c*	*b*+*d*	*n*=*a*+*b*+*c*+*d*

200. táblázat
Az infarktus előfordulási gyakorisága a dohányzók és nem dohányzók között (tíz év elteltével).

Ezután meghatározzuk a kockázat mértékét mindkét mintában:

A kockázat mértéke a ***dohányosok*** között: $a/(a+b)$.
A kockázat mértéke a ***nem dohányzók*** között: $c/(c+d)$.

Ezekből kiszámítjuk a relatív kockázatot (*RR*, relative risk):

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}\,, \qquad (162)$$

amely megadja, hogy ***hányszor gyakoribb a szívinfarktus*** (általánosan az adott betegség) ***kockázata a dohányzók körében*** (vagy általában a vizsgált rizikófaktor fennállása esetén), mint a nemdohányzóknál (vagy a rizikófaktor hiányában). Amennyiben H_0 igaz, akkor $RR = 1$.

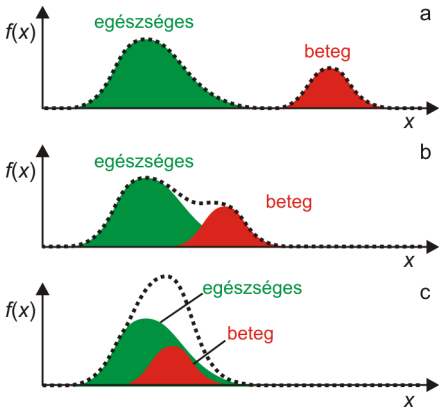
Az előzőkhöz hasonlóan kiszámítjuk az ln(*RR*) érték standard hibáját:

$$SE(\ln RR) = \sqrt{\frac{1-a/(a+b)}{a}+\frac{1-c/(c+d)}{c}}\,, \qquad (163)$$

majd szintén meghatározzuk a megfelelő konfidencia intervallumot. Ha ez tartalmazza az 1 értéket, akkor megtartjuk a nullhipotézist, ami azt jelenti, hogy nincs összefüggés a dohányzás és a szívinfarktus kialakulása között (az adott szignifikancia szinten). Ellenkező esetben pedig elvetjük a nullhipotézist (201. megjegyzés).

201. megjegyzés
a) A követéses vizsgálat, mint említettük, évekig is eltarthat, ezért igen költséges. A hosszú idő miatt általában nagyobb elemszámmal kell dolgozni az esetleges kiesések kompenzálására. Viszont abban az esetben, amikor a rizikófaktor előfordulás ritka, gyakorlatilag csak ez a módszer alkalmazható, mert visszatekintéses esetben nehezebb biztosítani a megfelelő nagyságú gyakorisági értékeket.

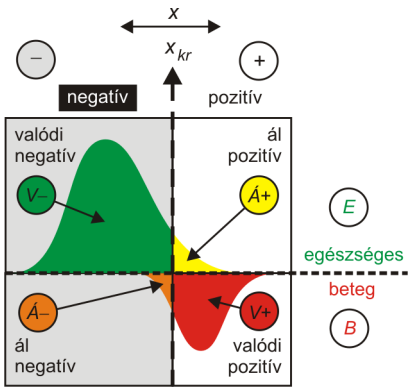
b) Az ilyen típusú gyakorisági adatok természetesen χ^2-próbával is feldolgozhatóak. Ezért felmerül a kérdés, hogy akkor miért használják mégis ezeket a módszereket? Az itt bemutatott eljárásoknak az az előnye, hogy a χ^2-próbával szemben, ahol valójában csak a *p* érték érdekel bennünket, itt a döntéshez egy-egy jellemző számszerű érték is rendelhető (*OR*, *RR* a hozzájuk tartozó hibával együtt), amely a kockázat mértékéről is felvilágosítást ad.



202. ábra
A diagnosztikai teszt alapjául szolgáló mennyiség (x) tipikus sűrűségfüggvényei, $f(x)$ (pontozott vonal).
a) Az egészségesekre és betegekre jellemző értékek jól elkülönülnek.
b) Az elkülönülés csak részleges.
c) A két csoport teljes átfedésben van.

203. megjegyzés
A 202. ábra azt az esetet mutatja, amikor az x mennyiség „középértéke” a betegekre nézve nagyobb, mint az egészségesekre. Mivel a lényeg az eltérésen van, ezért lehetne kisebb is, de ez a lehetőség a továbbiakat tartalmilag nem érinti.

204. megjegyzés
A többféle lehetőség közül a félreértések elkerülése végett a „valódi pozitív” (V+), „valódi negatív” (V–), „ál pozitív” ($Á$+), „ál negatív” ($Á$–) elnevezéseket használjuk. (A „helyes”, a „helytelen”, a „hamis” szavak ugyanazzal a betűvel kezdődnek, így használatuk rövidítésben nem lenne szerencsés.)



205. ábra
A diagnosztikai tesztek jellemzésére használható alap valószínűségek (a sűrűségfüggvény alatti területek). A sűrűségfüggvény definíciója alapján itt (V–)+($Á$+)+($Á$–)+(V+) = 1.

18.3. Diagnosztikai tesztek jellemzésére szolgáló statisztikai módszerek

Az orvosi statisztikának egy másik fontos területe a diagnosztikai tesztek vizsgálata, illetve használhatóságuk jellemzése. Gondoljunk egy olyan módszerre, amely azon alapul, hogy mondjuk egy folytonos skálán mérhető mennyiség, például valamilyen anyag koncentrációja eltérő az adott betegségben szenvedők (***betegek***) és ebben a betegségben nem szenvedők (***nem betegek, egészségesek***) csoportjában. Ilyen esetben az egészségeseket és a betegeket is tartalmazó alapsokaságra nézve az adott x mennyiségnek, mint valószínűségi változónak az $f(x)$ sűrűségfüggvénye jól szemlélteti az adott betegségnek ezt a jellemzőjét. A 202. ábra a két csoport elkülönülésének három alaptípusát mutatja be: a „teljes”, és a részleges elkülönülés mellett azt az esetet, amikor elkülönülésről már nem beszélhetünk (teljes átfedés) (203. megjegyzés).

A „teljes” elkülönülés (202a. ábra) csak igen ritkán fordul elő, ilyenkor a teszt gyakorlatilag egyértelmű eredményre vezet, további vizsgálatot nem is igényel. Teljes átfedés esetén (202c. ábra) a teszt használhatatlan, emiatt ezzel nem is foglakozunk.

Leggyakrabban részleges elkülönüléssel (202b. ábra) találkozunk, ami a teszt szempontjából további kérdéseket vet fel. A legfontosabb kérdés az, hogy **hol húzzuk meg a határt**, azaz mekkora az a kritikus érték (x_{kr}), amely fölött (vagy alatt) a tesztet pozitívnak (vagy negatívnak) tekintjük.

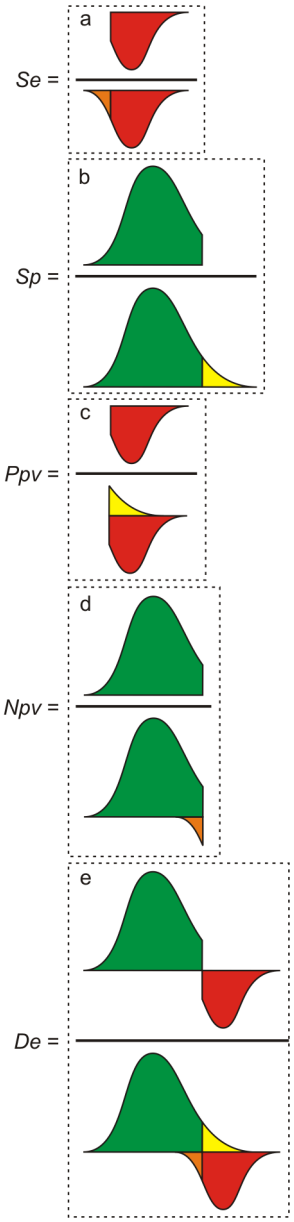
A probléma hasonló a H_0 nullhipotézis és a H_1 alternatív hipotézis elkülönítéséhez (vö. 137. ábra). Ugyanúgy, ahogy a nullhipotézis elvetésekor mindig számolnunk kell az elsőfajú hibával, elfogadásakor pedig a másodfajúval, a diagnosztikai tesztek esetében is függetlenül attól, hogy a teszt eredménye **pozitív** vagy **negatív**, a diagnózis lehet **helyes** (igaz, valódi), de **helytelen** (téves, hamis, ál) is (204. megjegyzés). Egy téves teszteredmény alapján végzett terápia beláthatatlan következményekkel járhat. Többek között ezért végez az orvos többféle vizsgálatot is ugyanannak a kóros elváltozásnak a felderítésére. Ennek ellenére igen fontos, hogy munkája során mindig tisztában legyen azzal, hogy egy adott teszteredmény valójában mire ad választ, és az mennyire jelentős az adott betegség kimutatásában, illetve mennyire bízhat meg a kapott eredményben.

A jobb szemléltetés kedvéért a 205. ábra segítségével új ábrázolásmódra térünk át. A vízszintes tengely fölött csak az egészségesekre vonatkozó sűrűségfüggvény, alatta pedig (tükrözve) csak a betegekre vonatkozó sűrűségfüggvény látható. A függőleges tengely a diagnosztikai teszt alapjául szolgáló mennyiség kritikus értékén (x_{kr}) halad keresztül. Ez a tengely választja ketté a teszteredményeket negatívra és pozitívra. Amennyiben a kritikus értéket megváltoztatjuk, ezzel együtt a színes területek nagysága is változik.

Az elméleti háttér után térjünk rá a gyakorlati kérdésekre. Nyilvánvaló, hogy mindegyik tesztmódszert még a széleskörű bevezetése előtt össze kell hasonlítani a „valósággal”. No, de honnan tudjuk a „valódi” eredményeket? Erre azt mondhatjuk, hogy vannak elég egyértelmű esetek, amikor ismerjük a válaszokat (ezek néha csak a boncolás eredménye alapján állapíthatók meg), továbbá vannak olyan, legtöbbször igen költséges, de nagyon nagy megbízhatóságú tesztek, amelyek segítségével hozzájuthatunk a szükséges adatokhoz. Mindezek alapján a tesztet elég nagy mintán elvégezve el tudjuk készíteni az alábbi 206. táblázatot, amely a lehetséges kimenetelek gyakoriságát mutatja be:

Diagnózis	A teszt eredménye		
A valóságban	negatív	pozitív	összesen
egészséges (nem beteg)	(V–) (valódi negatív)	($Á$+) (ál pozitív)	E=(V–)+($Á$+)
beteg	($Á$–) (ál negatív)	(V+) (valódi pozitív)	B=($Á$–)+(V+)
összesen	(–)=(V–)+($Á$–)	(+)=($Á$+)+(V+)	$Ö$=(V–)+($Á$+)+($Á$–)+(V+)

206. táblázat
A teszt lehetséges kimeneteleinek gyakorisága, vagyis a (V–), ($Á$+), ($Á$–), (V+) mennyiségek valamint a közöttük lévő egyszerű összefüggések.



207. ábra
A diagnosztikai paraméterek szemléletes jelentése a 205. ábra alapján.
a) Diagnosztikai szenzitivitás;
b) Diagnosztikai specificitás;
c) Diagnosztikai relevancia;
d) Diagnosztikai szegregancia;
e) Diagnosztikai effektivitás.

208. megjegyzés
További, ritkábban használt diagnosztikai paramétereket képezhetünk a már megadottak alapján:

$(1 - Se) = (\acute{A}-)/B$: álnegatív arány; $(P(-|B))$, (olyan, mint a **másodfajú hiba**);

$(1 - Sp) = (\acute{A}+)/E$: álpozitív arány; $(P(+|E))$, (olyan, mint az **elsőfajú hiba**);

$(1 - Ppv) = (\acute{A}+)/(+)$: téves figyelemfelkeltő arány; $(P(E|+))$,

$(1 - Npv) = (\acute{A}-)/(-)$: téves megnyugtatási arány; $(P(B|-))$.

Kérdés: Mennyire megbízható a teszt eredménye?

Erre a kérdésre nem adható egyértelmű válasz, pontosabban többféle válasz is adható. A válaszokat a diagnosztikai paraméterek számszerűen is jellemzik. Ezek közül némelyek függetlenek attól, hogy a betegségnek milyen az elterjedtsége, mennyire gyakori az adott populációban, de vannak olyanok, amelyek ettől erősen függenek. Ezért először erről ejtünk néhány szót.

Az elterjedtség, idegen szóval prevalencia (Pr) gyakran használt fogalom, amely a teszttől függetlenül az adott betegség előfordulásának valószínűségét adja meg (a 202. ábrán a piros terület nagysága). Amennyiben csak gyakorisági adatok állnak rendelkezésünkre, akkor ezt az értéket a relatív gyakorisággal közelíthetjük (a 206. táblázat jelöléseinek megfelelően):

$$Pr = B/(E+B) = B/\ddot{O}. \qquad (164)$$

A teszt megbízhatóságára jellemző paraméterek sorát a diagnosztikai szenzitivitással (vagy érzékenységgel) kezdjük (207. ábra). Ez a paraméter valójában egy feltételes valószínűség ($P(+|B)$), azt adja meg, hogy amennyiben csak a **betegek csoportjában** vizsgálódnánk, **mekkora a valószínűsége a pozitív teszteredménynek**. Más szavakkal annak a valószínűsége, hogy a teszt egy **beteget pozitívnak** is diagnosztizál. Gyakoriságokkal kifejezve (a valódi pozitívok aránya a betegek között):

$$Se = (\acute{A}+)/B. \qquad (165)$$

A nagy szenzitivitású tesztek a korai diagnózis során kívánatosak, ilyenkor ugyanis az a cél hogy a lehető legtöbb esetben derüljön ki a betegség. Ez a paraméter az egészségeseket egyáltalán nem veszi figyelembe (**a prevalenciától független**), emiatt a nagy szenzitivitású kritikus érték mellett a legtöbb esetben igen sok az ál pozitív ($\acute{A}+$) teszteredmény is.

A diagnosztikai specificitás a szenzitivitás fordítottja, szintén egy feltételes valószínűség ($P(-|E)$), azt adja meg, hogy amennyiben csak az **egészségesek csoportjában** vizsgálódnánk, **mekkora a valószínűsége a negatív teszteredménynek**. Vagy annak a valószínűsége, hogy a teszt egy **nem beteget** (egészségeset) **negatívnak** is diagnosztizál. Gyakoriságokkal kifejezve (a valódi negatívok aránya a nem betegek között):

$$Sp = (\acute{A}-)/E. \qquad (166)$$

A magas specificitású tesztek akkor fontosak, ha a pozitív eredmény súlyos következménnyel jár. Például kockázatos terápiát vagy műtétet von maga után. Mivel a **prevalenciától ez a paraméter is független** és a betegekkel nem is számol, a magas specificitással legtöbbször együtt jár a sok ál negatív ($\acute{A}-$) teszteredmény is.

Bár a szenzitivitás és a specificitás is fontos jellemzői egy tesztnek, talán még fontosabb az, hogy egy pozitív teszteredmény alapján milyen mértékben valószínűsíthető a betegség megléte, illetve negatív eredmény alapján annak hiánya.

Ezekről a következő paraméterek adnak számot:

A diagnosztikai relevancia (más néven korrekt pozitivitás), angol neve alapján rövidítve Ppv (positive predictive value) a valódi pozitív eredmény (vagyis a betegség meglétének) valószínűsége azzal a feltétellel, hogy a teszt pozitív ($P(B|+)$). Gyakoriságokkal kifejezve (a valóban betegek aránya a pozitívok között):

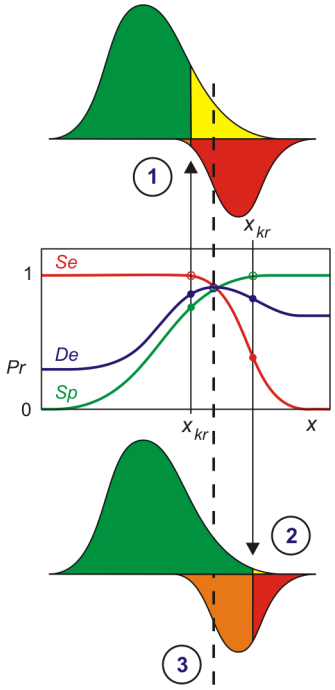
$$Ppv = (\acute{A}+)/(+). \qquad (167)$$

A diagnosztikai szegregancia (más néven korrekt negativitás), angol neve alapján rövidítve Npv (negatív predictive value) a valódi negatív eredmény (vagyis a betegség hiányának) valószínűsége azzal a feltétellel, hogy a teszt negatív ($P(E|-)$). Gyakoriságokkal kifejezve (a valóban egészségesek aránya a negatívok között):

$$Npv = (\acute{A}-)/(-), \qquad \text{(208. megjegyzés)} \quad (168)$$

A diagnosztikai effektivitás (más néven korrekt klasszifikáció) annak a valószínűsége, hogy a teszt helyesen diagnosztizál (a betegeket pozitívnak, az egészségeseket negatívnak). Gyakoriságokkal kifejezve (a valódi negatívok és a valódi pozitívok együttesének aránya a teljes populációban):

$$De = [(\acute{A}+) + (\acute{A}-)]/\ddot{O}. \qquad (169)$$



209. ábra
A kritikus érték (x_{kr}) megválasztásának lehetséges szempontjai:
1) $Se \approx 1$;
2) $Sp \approx 1$;
3) De = maximum.

Ez utóbbi három paraméter (*Ppv*, *Npv*, *De*) már **függ a prevalenciától** és ki is fejezhetők az előzőekkel:

$$Npv = \frac{Sp \cdot (1 - Pr)}{Sp \cdot (1 - Pr) + (1 - Se) \cdot Pr}, \tag{170}$$

$$Ppv = \frac{Se \cdot Pr}{Se \cdot Pr + (1 - Sp) \cdot (1 - Pr)}, \tag{171}$$

$$De = Se \cdot Pr + Sp \cdot (1 - Pr). \tag{172}$$

Az itt megismertetett 10 paraméter közül csak 3 független egymástól. (Néhányat csak kifejező neve miatt vezettünk be: például „téves figyelemfelkeltő arány”.) Azt várjuk azonban tőlük, hogy segítenek abban, hogy a teszt céljainak leginkább megfelelő kritikus értéket (x_{kr}) megtaláljuk.

Láthattuk, hogy a szenzitivitás (*Se*) és a specificitás (*Sp*) x_{kr} megváltoztatására éppen ellentétesen változik, ha nő a szenzitivitás akkor csökken a specificitás és fordítva. Emiatt három lehetőség közül választhatunk (209. ábra):

1. *Se* legyen közel 1, de *Sp* se legyen azért túlságosan kicsi. Használata elsősorban a szűrővizsgálatok esetében kívánatos, ugyanis ilyenkor kevés beteg marad felismerés nélkül, azaz praktikusan minden gyanús esetet kiszűrünk. A pozitívokat még tovább vizsgáljuk a terápia megkezdése előtt.

2. *Sp* legyen közel 1, de *Se* is maradjon számottevő. Használata elsősorban a komolyabb veszéllyel járó beavatkozások előtt fontos. Ilyenkor a cél ugyanis az, hogy ne tegyük ki felesleges kockázatnak a nem beteg embereket.

3. Mindkét paraméter (*Se* és *Sp*) egyidejű optimumát a diagnosztikai effektivitás (*De*) méri. Ha mindkét szempont egyformán fontos (legyen elég érzékeny a teszt, de lehetőség szerint nem betegeket ne minősítsen pozitívnak), akkor ezt a paramétert kell maximálisra állítanunk x_{kr} megválasztásával.

A paraméterek definíciójából adódóan a relevancia (*Ppv*) és a szegregancia (*Npv*) is ellentétesen változik, ha x_{kr}-t változtatjuk. Ilyen értelemben akár azt is mondhatnánk, hogy éppen úgy, mint a szenzitivitás (*Se*) és a specificitás (*Sp*). Megfigyelhettük azonban, hogy egyáltalán nem mindegy, hogy például a pozitív teszteredményt produkálók között keressük a betegeket, vagy a betegek között a pozitív teszteredményt produkálókat. Mint már említettük, vannak ugyanis a betegség elterjedtségétől (prevalenciától, *Pr*) független, illetve attól függő paraméterek. Míg *Se*, *Sp* *Pr*-től független, *Ppv*, *Npv* *Pr*-től függő paraméter.

Mintafeladat

Mutassuk meg, hogy a relevancia (*Ppv*) és a szegregancia (*Npv*) hogyan változik a prevalencia (*Pr*) megváltozásával!

Megoldás: Tegyük fel, hogy a populációra jellemző „mért” gyakoriságok (a 206. táblázatnak megfelelően) a következők:

90	10	E=100
10	90	B=100
(–)=100	(+)=100	$Ö$=200

Ez azért célszerű választás, mert ilyenkor $Pr = 0{,}5$; (a betegek és nem betegek aránya megegyezik), az $Se = Sp = Ppv = Npv = De$ diagnosztikai paraméterek pedig mind azonosak (0,9), tehát $Ppv = Npv = 0{,}9$ is (vö. 207. ábra).

Ha olyan populációval van dolgunk, ahol $Pr = 0{,}1$ (tehát csak 10% a beteg), akkor ezt például a következő gyakoriságok idézhetik elő, amelyek meghatározásakor ügyeltünk arra, hogy a prevalenciától független paraméterek nem változzanak meg:

810	90	E=900
10	90	B=100
(–)=820	(+)=180	$Ö$=1000

Láthatjuk, hogy ilyenkor az $Se = Sp = De$ paraméterek továbbra is azonos értéken maradnak (0,9), de a relevacia és a szegregancia megváltozik $Ppv = 0{,}5$; $Npv = 0{,}99$.

Az **alacsony prevalencia** tehát **lényegesen alacsonyabb relevanciát**, továbbá magasabb szegreganciát **eredményez**.

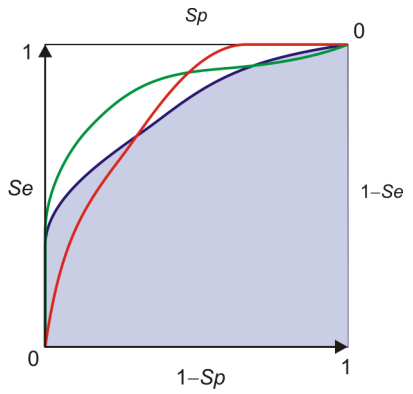
A mintafeladat éppen arra világít rá, hogy amennyiben a prevalencia (*Pr*) nagyon alacsony (tehát ritka betegségek esetén) hiába érhető el magas szenzitivitás (*Se*) és magas specificitás (*Sp*) egyszerre, ettől a relevancia (*Ppv*) még alacsony marad, ami azt jelenti, hogy a betegség kimutatására egy ilyen teszt elvégzésének nincs sok értelme. A szegreganciára (*Npv*) vonatkozóan a megállapítás éppen fordítva igaz, ott a túl nagy prevalencia teszi értelmetlenné a teszt elvégzését, ugyanis ilyenkor kapunk nagyon alacsony szegreganciát.

18.4. A diagnosztikai tesztek összehasonlításának szempontjai, a hatékonyság jellemzése, ROC elemzés

Amíg csak egy diagnosztikai tesztet vizsgálunk, addig az előző részben ismertetett diagnosztikai paraméterek is jól jellemezik a teszt teljesítőképességét, azaz, hogy mikor és mennyire használható az adott betegség kimutatására. Ilyenkor a ROC „görbét” csak a teszteredmények szemléletes bemutatására használjuk (210. megjegyzés). A ROC elemzés azonban elsősorban több diagnosztikai teszt összehasonlítására való, amikor ugyanazon betegség diagnosztizálására többfajta teszt eredményei ismertek, és azt kell megmondanunk, hogy melyik a hatékonyabb módszer, illetve melyiknek mikor javasoljuk a használatát.

210. megjegyzés
A ROC a Receiver Operating Characteristics rövidítése (jelentése kb. a vevőegység működését jellemző görbe). A fogalom az 1950-es években a radarjelek feldolgozásával kapcsolatosan került a köztudatba (a radarvevők érzékenységének és szelektivitásának beállításához használták). A módszert később átültették osztályozási, rangsorolási problémák kiértékelésére. Nagyjából az 1970-es évektől alkalmazzák széles körben orvosi diagnosztikai tesztek elemzésére.

A ROC „görbe” minden pontja egy (szenzitivitás; specificitás) (*Se*; *Sp*) pontpár, amely adott x_{kr}-hez tartozik. A görbét az egységnégyzetben szokás ábrázolni (211. ábra). A koordinátarendszer vízszintes tengelyén az $(1-Sp)$-t, a függőleges tengelyén az *Se*-t jelenítik meg. (A négyzet jobb oldalán feltüntethetők az $(1-Se)$ értékek, a felső oldalán pedig az *Sp* értékek is.) A különböző tesztek az adott görbék egymáshoz való viszonyítása és az egységnégyzetbeli elhelyezkedésük alapján hasonlíthatók össze.

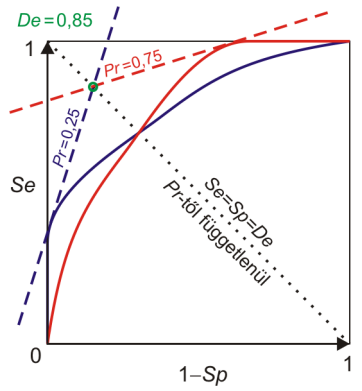


211. ábra
A ROC „görbék” (néhány lehetséges változat). A kék és a piros görbe alatti terület megegyezik.

A teszt diagnosztikai **hatékonyságát a ROC görbe alatti terület nagyságával** szokás kifejezni. Ez az érték csak akkor maximális (=1), amikor a teszt a betegeket és az egészségeseket „teljes” mértékben elkülöníti (vö. 202a. ábra). A diagnosztikai hatékonyság megítélésénél azonban figyelembe kell venni a ROC „görbék” alakját is, mivel két görbe alatti terület akkor is lehet azonos, ha maguk a görbék lényegesen eltérnek egymástól (vö. 211. ábra). A kérdés az, hogy ilyenkor milyen elvek szerint járjunk el, melyik ROC „görbével” jellemzett tesztet válasszuk.

Mint korábban említettük, de a (172) összefüggésből is kiderül, hogy a diagnosztikai effektivitás (*De*) általában a prevalenciától (*Pr*) is függ. Azonban azokra a *De* értékekre, amelyekre az $Se = Sp$ egyenlőség teljesül:

$$Se = Sp = De, \qquad Pr\text{-től függetlenül.} \tag{173}$$



212. ábra
A 211. ábra két azonos területű ROC „görbéje”. Az ábrán szaggatott vonallal jelöltük a prevalenciától (*Pr*) független különböző diagnosztikai effektivitású (*De*) pontokat (egyenes).
Az ábrán kék illetve piros pontozott vonallal feltüntettük a két különböző prevalenciájú ($Pr = 0{,}25$; $Pr = 0{,}75$), de azonos diagnosztikai effektivitású ($De = 0{,}85$) egyenest is.

Ezeket a pontokat a 212. ábrán a szaggatott vonal (egyenes) szemlélteti. Ugyancsak a (172) összefüggésből meghatározhatjuk az azonos diagnosztikai effektivitású (*De*) pontok által meghatározott görbéket is:

$$Se = \underbrace{\frac{1-Pr}{Pr}}_{a}\underbrace{(1-Sp)}_{x} + \underbrace{\frac{1}{Pr}De - \frac{1-Pr}{Pr}}_{b} \qquad y = ax + b \tag{174}$$

Amint látható ezek a „görbék” is egyenesek és meredekségük (*a*) csak a prevalenciától függ. Ezek szerint adott prevalencia esetén az azonos diagnosztikai effektivitású pontok mindig egy olyan egyenesen helyezkednek el, amelyik keresztül halad a (173) összefüggéssel megadott egyenes adott *De*-vel jellemzett pontján is.

Mindezek ismeretében az azonos ROC „görbe” alatti területtel jellemzett tesztek közül azt célszerű kiválasztanunk, amelyikhez a legnagyobb diagnosztikai effektivitás (*De*) tartozik. A 212. ábrán azt szemléltettük, hogy alacsony prevalencia esetén a kék, míg magas prevalencia esetén a piros görbével jellemzett teszt jelenti a jó választást.

19.0. Néhány összegző megjegyzés

A statisztika rész végén összefoglalásképpen két táblázatot mutatunk be. Az egyikben a ***visszatekintő*** és az ***előretekintő*** vizsgálatokat hasonlítjuk össze, a másikban a korábbiakban már ismertetett ***statisztikai próbákat szemléltetjük***. Mivel a próbák igen sokféle szempont szerint csoportosíthatók, nem tudunk általánosan elfogadott, átfogó képet adni róluk, így a bemutatott verzió csak egy a lehetséges táblázatok közül. A félreértések elkerülése végett fontos hangsúlyoznunk, hogy ez utóbbi táblázatban változón a sokaság definíciójában szereplő függő változót értjük (vö. 13. rész).

	Visszatekintő (retrospektív) (eset-kontroll) vizsgálat	**Előretekintő** (prospektív) követéses (kohort) vizsgálat
A kiválasztott egyedek megkülönböztetésének alapja (de egyébként hasonló összetételű csoportok)	**beteg – nem beteg** (optimálisan 50-50%)	**a rizikófaktor megléte – hiánya** (a kohorton belül) (optimálisan 50-50%)
A vizsgálat tulajdonságai	egyszerű, olcsó, azonnali; ritka betegségek esetén is használható	drága, hosszadalmas; ritka betegségek esetén nem, de ritkán előforduló rizikófaktorok esetén használható
Jellemző számszerű érték	esélyhányados (*OR*, odds ratio)	relatív kockázat (*RR*, relative risk), (ritka betegségeknél $RR \approx OR$, így *RR* eredeti jelentése elvész, ezért a módszer ilyenkor nem használható)

213. táblázat
A visszatekintő és az előretekintő vizsgálatokat összevetése.

<table>
<tr><td colspan="2" rowspan="3"></td><td colspan="3">folytonos</td><td rowspan="2">nem folytonos</td></tr>
<tr><td>normális</td><td colspan="2">nem normális</td></tr>
<tr><td></td><td>szimmetrikus</td><td>nem szimmetrikus</td><td></td></tr>
<tr><td rowspan="7">egy vagy több változó</td><td rowspan="3">várható érték</td><td>t</td><td rowspan="3">Wilcoxon</td><td rowspan="3">Mann-Whitney</td><td rowspan="3">Kruskal-Wallis</td><td></td></tr>
<tr><td>t</td><td></td></tr>
<tr><td>ANOVA</td><td></td></tr>
<tr><td>medián</td><td></td><td colspan="2">előjel</td><td></td></tr>
<tr><td>variancia</td><td>χ^2
F</td><td></td><td></td><td></td><td></td></tr>
<tr><td>eloszlás illeszkedés</td><td></td><td></td><td></td><td></td><td>χ^2</td></tr>
<tr><td>eloszlás homogenitás</td><td></td><td></td><td></td><td></td><td>χ^2</td></tr>
<tr><td rowspan="3">legalább két változó</td><td>korrelációs együttható</td><td>t</td><td></td><td></td><td></td><td></td></tr>
<tr><td>regressziós meredekség</td><td>t</td><td></td><td></td><td></td><td></td></tr>
<tr><td>függetlenség</td><td>F</td><td></td><td></td><td></td><td>χ^2</td></tr>
</table>

egymintás	
kétmintás	
kettőnél több mintás	
rang	

214. táblázat
A statisztikai próbák áttekintése.

215. megjegyzés
a) Az „információ” szó a latin „informo” igéből származtatható. Főnévi igenévként használva „informare”, amelynek jelentése: formálni az anyagot és az értelmet filozófiai, morális és pedagógiai értelemben (vagyis oktatni, nevelni). Az „informatio” az a főnév, amely erre a folyamatra utal és utal az ehhez kapcsolódó fogalmakra is, főként a „forma” fogalmára, amely úgy szerepel, mint a lehetőség a tudásra.

b) Az informatikusok szakmai közösségében a két alapfogalom, az adat és az információ viszonya szerzőnként eltérő. Az egyik megfogalmazás szerint az információ az adat részhalmaza, a másik szerint a rész-egész viszony pont fordított. Mindkét megközelítésben a közös vonás az, hogy az információ olyasvalami, amely a befogadó számára jelentéssel bír, továbbá fontos, illetve értékes lehet.

216. megjegyzés
A számítástechnika az informatika egyik részterülete, amely egy konkrét eszközrendszerre vonatkozóan végzi ugyanezeket a feladatokat.

217. példa

Képzeljük el azt az esetet, hogy lóversenyen szeretnénk fogadni TÉT-re, azaz azt kell megmondanunk, hogy az adott versenyben induló lovak közül a futam végén melyik ló lesz az első.

Amennyiben a lovakról semmiféle ismerettel nem rendelkezünk, akkor csak azt tehetjük, hogy az egyiket véletlenül kiválasztjuk és arra fogadunk. Ilyenkor ha mondjuk 9 induló van a futamban, akkor $P = 1/9$ annak a valószínűsége, hogy nyerünk. Ha azonban a futam résztvevőiről bizonyos információval rendelkezünk, mondjuk ismerjük a futam aktuális favoritjait, vagyis azokat az indulókat, amelyek jó eséllyel végeznek az első helyen, akkor ezt az információt felhasználva már csak azok közül kell véletlenül választanunk. Tételezzük fel, hogy az indulók között 3 favorit van, amelyek közül bármelyik egyformán esélyes az első helyre. Ilyenkor, ha az egyik favoritra fogadunk, nyerési esélyünk megnő és $P = 1/3$-ra módosul.

218. megjegyzés
A (177) összefüggést a fizikában használt, Boltzmann által bevezetett entrópiával való hasonlósága miatt a kísérlet entrópiájának is szokás nevezni.

20.0. Informatikai alapfogalmak

Az 1.0. bevezető részben alapfogalomként az adatokat és a jeleket említettük és szándékosan kerültük az egyértelműnek tűnő definíciókat, mivel jó meghatározások nem is léteznek. Hasonlóan járunk el most is és csak annyit mondunk, hogy akinek adatok vannak a birtokában, az információval rendelkezik (215. megjegyzés). Az információt valamilyen módon elő kell állítani, adott esetben meg kell őrizni, tárolni kell, ismerni kell az értelmezését stb. Azoknak az eszközöknek és módszereknek az összességét, amelyek minderre hivatottak, információtechnológiának nevezzük. Azt a tudományágat pedig, amely a fentieknek megfelelő rendszerek fejlesztési, üzemeltetési, elemzési kérdéseivel foglalkozik, informatikának hívjuk (216. megjegyzés).

Az információ sok esetben nem a keletkezési helyén kerül feldolgozásra, így azt általában el kell juttatni az egyik rendszerből a másikba. Az ezzel kapcsolatos eljárásokat nevezzük összefoglaló néven kommunikációnak. A kommunikációs folyamatban legalább két fél vesz részt: az információt közlő (adó) és az információt fogadó (vevő). A kommunikáció csak akkor lehet sikeres, ha az adó által közölt információ ugyanazzal a jelentéstartalommal jelenik meg a vevőnél. Itt ismét utalhatunk a bevezetőben mondottakra: a közvetítők a jelek, tehát a kommunikáció is jelekkel történik. Ha beszélünk vagy leírunk valamit, jelsorozatokat használunk. Ezeket meghallgatva, elolvasva **bővülhet** a már birtokunkban lévő információ. Ezt a bővülést, növekedést használhatjuk fel az információ mennyiségének meghatározásához (217. példa).

A példából kiderül, hogy valamilyen **új ismeret** csökkenheti egy döntés bizonytalanságát, növelheti a cél elérésének az esélyét. A birtokunkba jutó információ mennyiségét éppen annak a bizonytalanságnak, határozatlanságnak a mennyiségével fogjuk mérni, amelyet az új ismeret eloszlat, megszüntet (vö. 20.2. rész).

20.1. A megfigyelések, „kísérletek” határozatlansága

Az egyszerűség kedvéért először olyan megfigyeléseket tekintsünk, ahol sokféle eredményt kaphatunk, de a kísérletek lehetséges kimenetelei egyformán valószínűek. Ilyen esetekben elég egyértelmű, hogy a kísérlet kimenetele annál határozatlanabb, minél nagyobb a kimenetelek száma (k) (9 ló közül nehezebb eltalálni a nyerőt, mint 3 közül). Amennyiben a H határozatlanságot nem a kimenetelek számával, hanem annak logaritmusával jellemezzük:

$$H = \log_a k, \tag{175}$$

akkor ez többek között azzal az előnnyel is jár, hogy egyetlen kimenetel esetén a határozatlanság 0 ($\log_a 1 = 0$), amit értelemszerűen várunk.

Ezek után térjünk rá arra az esetre, amikor a kísérletek lehetséges kimenetelei nem egyformán valószínűek. Jelölje A_i ($i = 1, 2, 3, \ldots, k$) a kísérlet kimeneteleit és $P(A_i)$ az egyes kimenetelek bekövetkezéséhez mint eseményekhez tartozó valószínűségeket. Ezzel a jelöléssel a (175) összefüggés akkor írható fel, ha $P(A_i) = 1/k$ (minden i-re). Az általánosításra a következő lépéseken keresztül juthatunk el:

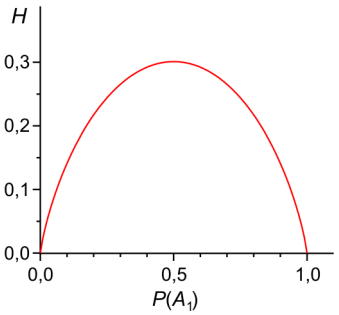
$$H = \log_a k = -\log_a \frac{1}{k} = -k\frac{1}{k}\log_a \frac{1}{k} = \underbrace{-\left(\frac{1}{k}\log_a \frac{1}{k}\right)-\left(\frac{1}{k}\log_a \frac{1}{k}\right)-\ldots}_{k \text{ tagú összeg}}. \tag{176}$$

$1/k$ helyére $P(A_i)$-t írva:

$$H = -\sum_{i=1}^{k} P(A_i)\log_a P(A_i), \tag{177}$$

amely egy megfigyelés határozatlanságát adja meg most már általános formában. Látható, hogy ez az összefüggés valójában egy várható érték, a $\log_a P(A_i)$ mennyiségek súlyozott átlaga (vö. (33), 218. megjegyzés).

A határozatlanság akkor maximális, ha az összes $P(A_i)$ valószínűség megegyezik. Bár ez a megállapítás általános érvényű, kétféle kimenetel esetén könnyen szemléltethető is. Ilyenkor ugyanis $P(A_2) = 1\text{-}P(A_1)$ és H minden P-re egyszerűen kiszámítható. Az eredményt a 219. ábra mutatja be, ahol a számolást 10-es alapú logaritmus használatával végeztük el ($a = 10$). Jól megfigyelhető, hogy az ábrán a görbe a $P(A_1) = P(A_2) = 0{,}5$-nél éri el a maximumát, ahol $H_{max} \approx 0{,}3$.



219. ábra
A H határozatlanság két lehetséges kimenetelű kísérletekben, az egyik kimenetel bekövetkezése mint esemény (A_1) valószínűségének ($p(A_1)$) függvényében. A függvény 0,5-nél éri el a maximumát, ahol $H_{max} = -\lg(1/2) \approx 0{,}3$ (vö. (177) összefüggés).

20.2. Információmennyiség

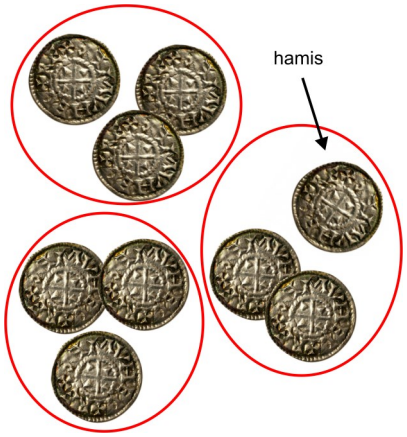
Az információmennyiség meghatározásának igénye először az **üzenetek** továbbításával kapcsolatosan merült fel. Az információ ugyanis szubjektív fogalom, értelmezése és fontossága függ annak befogadójától. Például az „esik az eső” kijelentést másként értékeli az, aki éppen sétálni indul, mint aki a kertjét akarta meglocsolni. Ráadásul az adott információ fontossága akár ugyanazon személy számára is változhat az idővel. Ezért az információ „eredeti” jelentése helyett kell egy absztraktabb meghatározás annak érdekében, hogy az információmennyiséget egyértelműen tudjuk megadni. Ehhez **az információnak azt az általános tulajdonságát használjuk fel, hogy bizonytalanságot, határozatlanságot szüntet meg**. A határozatlanság folyamatos csökkenésére jó példa a barkochba játék, ahol az a cél, hogy kitaláljuk, éppen mire gondolt játszótársunk. Kezdetben fogalmunk sincs a kigondolt valamiről, de jó kérdések és az igen-nem válaszok eredményeként eljuthatunk a megoldáshoz. Egy ilyen folyamatban ahogy csökken a határozatlanság, úgy nő a birtokunkba jutó információ.

Nem mindegy azonban, hogy egy üzenet milyen feltételek mellett jut tudomásunkra. Például az a nem túl bőbeszédű távirat, hogy „a holnapi vonattal érkezem...”, sokkal több bizonytalanságot szüntet meg akkor, ha naponta csak két vonat megy az adott településre, mint, ha mondjuk tíz. A bizonytalanság akkor szűnne meg teljesen, ha egyértelműen azonosítani tudnánk a vonatot. Ennek alapján a következőkben **információn általában valamely véges számú és előre ismert lehetőség valamelyikének a megnevezését értjük**, ami nem más, mint egy esemény bekövetkezése.

Mintafeladat

Példaként tekintsük azt a „hamis pénz” problémát, ahol 9 külsőre megegyező pénzérme közül kell kiválasztani a hamisat, amelyik könnyebb a többinél, de csak egy olyan kétserpenyős mérleg áll rendelkezésünkre, amely a súlyok kisebb-nagyobb voltát, illetve egyenlőségét tudja jelezni.

Megoldás: Első lépésben 3 darab hármas csoportból kettőt felteszünk a mérlegre. Ha az egyik serpenyő lebillen, akkor a másikban van a hamis pénz, ha egyensúlyban marad, akkor a harmadikban, tehát abban a hármas csoportban van, amelyiket nem tettük fel a mérlegre. A következő lépésben az előző körben kiválasztott hármas csoportból teszünk egy-egy érmét a két serpenyőbe és ismét alkalmazzuk az előző szabályt. Ily módon **két mérlegeléssel** a hamis pénzt sikerült azonosítanunk és ez az eljárás mindig eredményre vezet (220. ábra; 221. megjegyzés).



220. ábra
Aba Sámuel 9 darab ezüst pénze között van egy hamis. A kérdés az, hogy mennyi információ szükséges a hamis pénz megtalálásához, ha feltesszük, hogy a hamis pénz mérhetően könnyebb a többinél, (amelyek nagyjából egyforma súlyúak).

221. megjegyzés
Természetesen van olyan eset is, hogy akár egy méréssel is sikerül a kiválasztás. Ha például 4-4 érmét teszünk a serpenyőkbe és az éppen egyensúlyban van akkor nyilvánvalóan a kimaradt érme a hamis. Ha azonban a 4-4 érme nincs egyensúlyban, akkor akár további két mérésre is szükségünk lehet az azonosításhoz.

Matematikai értelemben a mintafeladat megoldásakor először a kilencelemű alaphalmazt előállítottuk 3 darab háromelemű diszjunkt részhalmaz uniójaként (vö. 2.2. rész; (2), (3)). Amikor az első méréssel azonosítottuk a háromelemű részhalmazok egyikét, akkor bekövetkezett az „ebben a hármas csoportban van” esemény (A_1). Amikor a második méréssel ezen a részhalmazon belül azonosítottuk a hamis pénzérmét, akkor bekövetkezett a „hármas csoporton belül ez a hamis pénz” esemény (A_2). Természetesen, ha valaki egyből rábök a hamis pénzre, akkor bekövetkezik az „ez a hamis pénz” esemény (A). Ezek szerint $A = A_1 \cap A_2$. A bevezetendő I információmennyiségtől elvárjuk, hogy nagysága ne függjön attól, hogy milyen módon (egy vagy két lépésben) határoztuk meg, tehát ne legyen különbség aközött, hogy egyből rábökünk a hamis pénzre, vagy a megoldásban leírtak szerint határozzuk meg. Ebben az értelemben tehát I **legyen additív**:

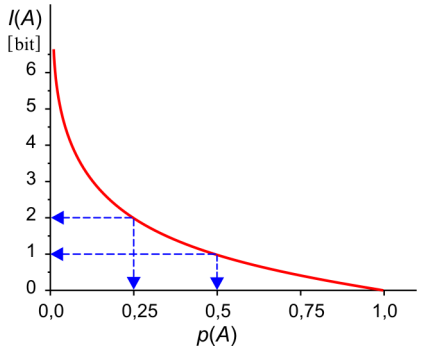
$$I(A) = I(A_1) + I(A_2), \tag{178}$$

továbbá kötődjön az események valószínűségéhez. Azt már a lóversennyel kapcsolatban is láthattuk, hogy nagyobb bizonytalanságot szüntetünk meg akkor, ha mind a kilenc induló ló közül találjuk el a győztest, mintha csak a három favorit közül kellett volna ugyanezt tennünk. Ezek szerint egy **valószínűbb eseményhez kisebb információmennyiségnek** kell tartoznia, mint egy kevésbé valószínűhöz. Amennyiben egy A esemény információmennyiségét az alábbi összefüggéssel adjuk meg:

$$I(A) = \log_a \frac{1}{p(A)} = -\log_a p(A), \tag{179}$$

ahol $p(A)$ az A esemény bekövetkezésének valószínűsége, akkor az előírt feltételek teljesülnek. Ezek közül az $1/P(A)$ alkalmazása nem szorul külön bizonyításra, a (178) feltétel teljesülését pedig a „hamis pénz” probléma kapcsán ellenőrizhetjük.

222. megjegyzés
a) Az információmennyiség további egységei is használatosak, bár kevésbé elterjedtek:
1 nat = $\log_2 e$ bit ≈ 1,44 bit (e-ed részére csökken a bizonytalanság)
(natural unit = természetes egység).
1 hartley = $\log_2 10$ bit ≈ 3,32 bit (tizedére csökken a bizonytalanság)
(Ralph Vinton Lyon Hartley (1888 - 1970) amerikai elektromérnök és informatikus neve után);
b) A „bit” szót kicsit más jelentéstartalommal az adatmennyiség megnevezésére is használják (vö. 20.6. rész). Ez a „binary digit” (bináris számjegy) kifejezésből származik, annak rövidítése (vö. 95. megjegyzés).



223. ábra
Az A esemény egyedi információmennyiségének bitekben kifejezett értéke ($I(A)$) mint a $p(A)$ valószínűség függvénye.

Tudjuk, hogy a hamis pénz megtalálásának valószínűsége 1/9. A mintafeladat megoldása szerint, mivel egy mérlegelésnek 3 kimenetele lehet, ezért 1/3 a valószínűsége annak, hogy az első lépésben megtaláljuk a hamis pénzt is tartalmazó hármas csoportot, majd szintén 1/3 a valószínűsége annak is, hogy ezen belül rátaláljunk a hamis pénzre (feltételes valószínűség). Ismerve a valószínűségek szorzási szabályát (vö. 4.1. rész; (14)), az információmennyiségekre igaz, hogy

$$-\log_3 \frac{1}{9} = 2 = -\log_3\left(\frac{1}{3}\cdot\frac{1}{3}\right) = -\log_3\left(\frac{1}{3}\right) - \log_3\left(\frac{1}{3}\right) = 1+1 = 2\,, \qquad (180)$$

ami megfelel a (178) összefüggésben szereplő feltételnek.

Azt is láthatjuk, hogy az eredmény éppen a mérlegelések számával egyenlő. Általánosítva azt mondhatjuk, hogy ha 9 pénzdarabból 2 méréssel tudtuk kiválasztani a hamisat, akkor általában $n = 3^k$ pénzdarabból k méréssel, tehát ebben az esetben $\log_3 3^k = k$ adja meg az információmennyiséget. Itt praktikus okokból a 3-as alapú logaritmust használtuk, amit megtehettünk, hiszen a (179) összefüggésben ez még nincs rögzítve, és általánosan csak a-ként szerepel. Megállapodás szerint azonban az információmennyiség skáláját definíció szerint úgy választották meg, hogy akkor legyen **egységnyi az információmennyiség, ha felére csökken a bizonytalanság**, tehát $I(A) = 1$, ha $p(A) = 1/2$. Ez viszont csak akkor teljesül, ha $a = 2$, azaz 2-es alapú logaritmust használunk, hiszen ilyenkor $-\log_2(1/2) = 1$.

Az egység neve a bit, amely a „binary unit” (kettes egység) rövidítése (222. megjegyzés). Például egy eldöntendő kérdésre adott válasz információtartalma akkor 1 bit, ha mindkét válasz egyformán valószínű. Ezek után feltéve azt a konkrét kérdést, hogy hány bit információra van szükség egy adott véges halmaz valamely tetszőleges elemének azonosításához, akkor erre a következő összefüggéssel válaszolhatunk:

$$I(A) = -\log_2 p(A)\,[\text{bit}]\,, \qquad (181)$$

ahol A jelenti azt az eseményt, hogy az adott elemet azonosítottuk és $p(A)$ az esemény bekövetkezésének valószínűsége (223. ábra).

Mintafeladat

Hány bit információra van szükség ahhoz, hogy egy 32 lapos magyarkártya-pakli egyik kiválasztott lapját barkochba játék keretében kitaláljuk. (A játékszabály ismert: csak olyan kérdést lehet feltenni, amire „igen-nem” válasz adható.)

Megoldás: Legyen a kitalálandó kártyalap a makk felső. Nyilvánvalóan eljárhatunk úgy is, hogy sorra vesszük az összes kártyát és akkor a 31. kérdés után (a 32. már nem kell feltenni) biztosan meg tudjuk nevezni a keresett lapot. Természetesen az is lehetséges, hogy egyből kitaláljuk, tehát az első kérdés után. Azt mondhatjuk, hogy átlagosan 16 kérdéssel eljuthatunk a megoldáshoz.

Hatékonyabb módszerhez folyamodhatunk azonban, ha először csak azt kérdezzük meg, hogy a lap „színe” mondjuk a „tök” vagy „piros” lapok közül való. Erre a kérdésre kapott „nem” válasz után ugyanis már csak 16 kártyalap között kell keresnünk a megfejtést. Ismét felezhetjük a szóba jövő kártyák számát (8 kártyalap), ha a következő kérdésünk az, hogy a lap „színe” „zöld”-e (nem). Ezután megkérdezhetjük, hogy található-e szám a lapon (nem), ezzel ismét feleztük a lehetséges kártyák számát (4 kártyalap)... Ezt a módszert folytatva az 5. kérdés után azonosítani tudjuk a kitalálandó kártyalapot. Így az előző módszerrel szemben csupán 5 kérdés elég volt az azonosításhoz (vö. előző mintafeladat: két mérés). Mivel minden kérdés után felére csökkent a bizonytalanság, ez éppen 5 bitnyi információt jelent. Vagy másképpen $p(A) = 1/32$, amelyet a (181) összefüggésbe helyettesítve szintén 5 bitet kapunk eredményül.

20.3. Üzenetek információmennyisége

Az eddigiekben az egyes eseményeknek csak az **egyedi információmennyiségét** tanulmányoztunk. Célunk azonban, amint azt már az előző rész elején említettük, a továbbított **üzenetek információmennyiségének** jellemzése. Ezért most azt vizsgáljuk meg, hogy hogyan áll elő egy üzenet.

224. megjegyzés
A jelek egymásutánja reprezentálja az információt. Az elrendezés lehet időbeli, ilyenek például a beszédhangok, de lehet térbeli is, ilyenek például az írás betűi.

Mindenekelőtt szükségünk van egy jelkészletre, amely N különböző jelet tartalmaz (x_j, $j = 1, 2, 3, \ldots, N$) és ebből választunk ki n darab jelet az adott üzenet számára (224. megjegyzés). Azt is mondhatjuk, hogy matematikai értelemben a jelkészlet egy ismert eloszlású diszkrét valószínűségi változó (ξ) összes lehetséges kimenetelét jelenti, ahol $p_j = P(\xi = x_j)$ annak a valószínűsége, hogy a jelkészlet j-edik elemét azonosítottuk. Amennyiben a p_j valószínűségek megegyeznek, tehát

mindegyik jel ugyanakkora valószínűséggel szerepelhet egy üzenetben ($p_j = 1/N$ minden j-re), akkor az n jelből alkotott lehetséges üzenetek száma N^n, ugyanis minden egyes jel, függetlenül a többitől N jel közül választható ki. Ha ennek a mennyiségnek vesszük a 2-es alapú logaritmusát érdekes eredményre jutunk:

$$\log_2 N^n = n\log_2 N = -n\log_2 \frac{1}{N} = -n\log_2 p_j \,. \qquad (182)$$

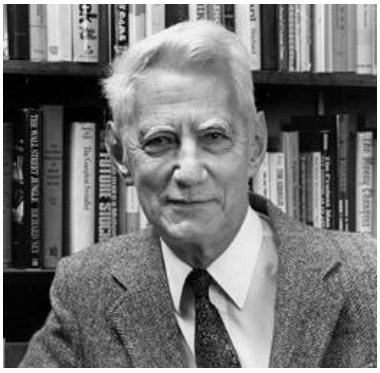
A (181) összefüggéssel való összevetés után láthatjuk, hogy a kapott mennyiség az egyetlen jel azonosításához szükséges egyedi információmennyiség n-szerese, bitekben kifejezve. Ilyen esetekben ezt tekintjük **az *n* jelből álló üzenet információmennyiségének** (I-nek) (225. példa). Az üzenet információmennyiségének ilyen módon történő megadása azért is praktikus, mert ez a mennyiség **az üzenet hosszával** (n-el) **arányos**.

225. példa

A (182) összefüggés segítségével számítsuk ki a bázishármasokra mint üzenetekre vonatkozó információmennyiséget. Mivel az mRNS-ben mind a négy bázis (A, U, G, C) előfordulása ugyanakkora valószínűségű, $p_j = 1/4$, továbbá $n = 3$, ezért:

$$I = -3\log_2 (1/4) = 6 \text{ bit}.$$

Felmerül a kérdés, hogy miként járjunk el abban az esetben, ha a p_j valószínűségek különböznek egymástól. A (182) összefüggést látva elsőre az egyedi információmennyiségek ($-\log_2 p_j$) egyszerű összegzése juthat eszünkbe. Ezzel a definícióval az lenne a probléma, hogy az előbb említett (n és I közötti) arányosságot elrontaná, hiszen a különböző valószínűségű jelek különböző gyakorisággal fordulnak elő az azonos hosszúságú üzenetekben. A megoldás Shannon nevéhez fűződik, aki a következő javaslattal állt elő.



Claude Elwood Shannon (1916 - 2001) amerikai matematikus és elektromérnök, az „információelmélet atyja”.

Meghagyva az üzenet információmennyiségének azt a tulajdonságát, hogy az egyes jelek által hordozott egyedi információmennyiség valamilyen összegződéséből adódik, először kiszámítjuk az egész jelkészletre vonatkozó jelenkénti közepes információmennyiséget, azaz a várható értéket, és a továbbiakban ezzel az átlaggal számolunk. Így egy üzenet soron következő jelének átlagos hozzájárulása az üzenet információmennyiségéhez mindig:

$$I(\xi) = -\sum_{j=1}^{N} p_j \log_2 p_j \,[\text{bit}], \qquad (183)$$

ahol ξ az adott jelkészlethez tartozó, a p_j valószínűségekkel megadott eloszlású valószínűségi változó. Ezt a mennyiséget egy megfigyelés határozatlanságával való formai hasonlósága miatt Shannon-féle entrópiának nevezzük (vö. (177); 218. megjegyzés). Mivel 2-es alapú logaritmus használata esetén $I = H$, a továbbiakban a szokásoknak megfelelően ezt a mennyiséget is ***H*-val jelöljük**.

Mintafeladat

Mennyi a Shannon-féle entrópiája az angol ábécének?

Megoldás: A feladat megoldásához szükségünk van arra az ismeretre, hogy az ábécében szereplő 26 betű milyen gyakran fordul elő az angol nyelvben. Az alábbi táblázatban ezt foglaltuk össze:

E / 12,31%	O / 7,94%	S / 6,59%	L / 4,03%	U / 3,1%	M / 2,25%	B / 1,62%	K / 0,52%	J / 0,1%
T / 9,59%	N / 7,19%	R / 6,03%	C / 3,2%	P / 2,29%	W / 2,03%	G / 1,61%	Q / 0,2%	Z / 0,09%
A / 8,05%	I / 7,18%	H / 5,14%	D / 3,65%	F / 2,28%	Y / 1,88%	V / 0,93%	X / 0,2%	

Ha ezeket az értékeket behelyettesítjük a (183) összefüggésbe, akkor 4,167 bitet kapunk eredményül.

Érdemes megjegyezni, hogy amennyiben minden betű azonos valószínűséggel fordulna elő ($p_j = 1/26$), akkor 4,7 bit lenne a betűnkénti információmennyiség (vö. 223. ábra).

226. példa

Ha egy A esemény valószínűsége $P(A) = p$, de a B esemény megfigyelése után $P(A|B) = q$-ra módosul, akkor ezzel

$$(-\log_2 p) - (-\log_2 q) = \log_2 \frac{q}{p} = \log_2 \frac{P(A|B)}{P(B)}$$

információmennyiséget nyerünk (vagy veszítünk). Azt is érdemes megfigyelni, hogy független események esetén ez a változás 0 (vö. (16)).

A jelenkénti átlagos információmennyiség (183) és egy megfigyelés határozatlansága (177) közötti hasonlóság nem meglepő, hiszen az információmennyiség bevezetésekor az információnak éppen azt az általános tulajdonságát használtuk fel, hogy határozatlanságot szüntet meg. A határozatlanságot tehát felfoghatjuk úgy is, mint információhiányt. A két fogalom viszonyát jól szemlélteti a 226. példa.

A 20.1. részben (kétféle kimenetel esetén) már bemutattuk, hogy a határozatlanság akkor maximális, ha a valószínűségek megegyeznek (vö. 219. ábra). Ugyanerre utal az angol ábécére vonatkozó mintafeladat megjegyzése is, hiszen $I = H$. Azt mondhatjuk tehát, hogy minden olyan jelkészletben, ahol az egyes jelek előfordulásához tartozó valószínűségek különbözőek, a jelenkénti átlagos információmennyiség nem érheti el a maximumát, azaz $H < H_{\max}$.

227. példa

Például, ha ismerjük valakinek a nevét, akkor legtöbbször azt is tudjuk, hogy az illető férfi, vagy nő. Mégis sok kérdőíven ki kell tölteni az erre szolgáló rovatot is, hiszen csupán a név alapján nem mindig lehet egyértelműen megállapítani a nemet.

228. példa

Egy másik példaként idézzük fel a 19.2. rész második mintafeladatát. A megoldásból az derül ki, hogy a barkochbajátékban minden kérdés-feleletpár maximálisan egy bit információt szolgáltathat. Ezt tudva kidolgozhatunk egy optimális stratégiát a játék lefolytatására. Ha tehát helyesen kérdezünk, minden kérdéssel egy bit információt nyerve, öt kérdésre lesz szükségünk a 32 lapos magyar kártya egyik lapjának kitalálásához.

Természetesen, ha rosszul kérdezünk, nem lesz elegendő az öt kérdés. Vagy azért, mert kérdésünkre egy bitnél kevesebb információt kapunk, vagy azért, mert a válasz egy része már benne volt egy előző kérdésre adott válaszban. A válasz ebben az esetben is redundáns, és végeredményben ilyenkor is kevesebb mint egy bit információt tartalmaz.

230. megjegyzés
Ilyen jelrendszer például a Morse-ábécé. Ez olyan kommunikációs kód, amely szöveges információ átvitelét teszi lehetővé valamilyen csatornán keresztül. A kommunikációs eszköz egy adott időpontban vagy kikapcsolt (szünet), vagy pedig bekapcsolt (adás) állapotban van. Maguk a Morse-jelek ezeknek az állapotoknak a rövidebb-hosszabb idejű változásai. Ha meg akarjuk jeleníteni őket, akkor ezt több módon is megtehetjük. Kimondva vagy leírva például a „ti-tá” szavakkal, esetleg lerajzolva pontokkal és vonalakkal (• —), de közvetlenül is hallhatóvá vagy láthatóvá tehetjük sípszóval, illetve egy villogtatott lámpával.

231. példa

Az egyik legismertebb hírközlő rendszer a rádió. Itt a forrás a bemondó, az ő hangja kerül megfelelő átalakítás után a rádióadó-berendezésbe. A kódolás magában foglalja mindazokat a folyamatokat, amelyek eredményeként a hangrezgések modulált rádióhullámokká alakulnak. A csatorna szerepét a tovaterjedő rádióhullámok játsszák, a vevő funkcióját pedig a rádióvevő berendezés látja el. Ez utóbbiban megy végbe a rádióhullámok hangrezgésekké való visszaalakítása, azaz a dekódolás. Az információ felhasználója pedig a rádióhallgató.

20.4. Redundancia

Az előbbiek alapján nyilvánvaló, hogy például egy hírforrás jellemzésekor különbséget kell tenni a maximális és tényleges entrópia között. A kettő hányadosát (H/H_{max}-ot) relatív entrópiának nevezzük. Ez a mennyiség olyasmi, mint valami hatásfok („hasznos/összes”), de az informatikában gyakorlati okokból inkább a „haszontalan/összes” arányt részesítik előnyben:

$$R = \frac{H_{max} - H}{H_{max}} = 1 - \frac{H}{H_{max}}, \qquad (184)$$

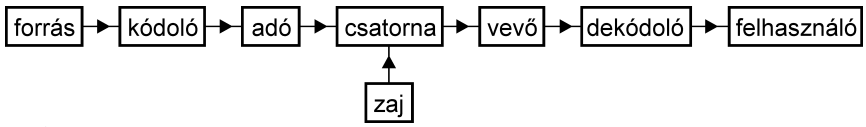
amelyet redundanciának (magyarul terjengősségnek vagy bőbeszédűségnek) neveznek (227. példa).

Ha egy **üzenet redundáns**, akkor **kevesebb az információmennyisége, mint amennyit a jelek száma alapján tartalmazhatna**. A jelenség egyik oka, mint korábban láthattuk, az, hogy a jelek előfordulási valószínűsége nem egyenlő. Azonban akkor is csökken az üzenet információmennyisége, ha a jelek között valamilyen összefüggés áll fenn. Ha egy jel bekövetkezése függ az előző jel bekövetkezésétől – például a magyar nyelvben mássalhangzó után nagyobb valószínűséggel következik magánhangzó és fordítva (vö. 4.0. rész) –, akkor a jel bekövetkezésére vonatkozólag már rendelkezünk bizonyos mennyiségű információval, így az üzenetek redundanciája nagyobb lesz (228. példa).

A redundancia biztosítja az üzenet egyértelmű értelmezhetőségét abban az esetben, ha valamilyen okból a jelsorozat sérül. Gyakran találkozunk újságokban, könyvekben nyomdahibával, amikor például felcserélődik két betű, vagy éppen kimarad egy, de ettől még az esetek többségében helyesen értelmezhető az eredeti tartalom, mert például a szövegkörnyezet egyértelművé teheti a mondanivalót. Általában azt mondhatjuk, hogy a természetes nyelvek redundanciáját nagymértékben növelik a nyelvtani szabályok is.

20.5. Üzenetek továbbítása, hírközlési rendszer

Az egyirányú hírközlési rendszerben az elküldött információ a közlőtől, az információ forrásától a fogadóhoz vagy más szóval az információ felhasználójához jut el (229. ábra).



229. ábra
Egyirányú hírközlési rendszer általános modellje.

A rendszer bemeneti oldalán kell összeállítani az ***üzenetet***, amelyet a kódoló tesz alkalmassá a továbbításra. A kód olyan, megállapodás szerinti jelrendszer, amely segítségével a jelkészlet elemeiből és a kódolás szabályainak alkalmazásával valamely információ egyértelműen megadható (230. megjegyzés). A kódolás az az eljárás, amikor az egyik jelrendszerről egy másikra térünk át, de ennek változatairól még szót ejtünk (20.6. rész). A kódolt jeleket az adó a hírközlő-csatornába küldi. A csatornából érkező jeleket a vevő fogja fel, majd a dekódolóba juttatja. A dekódoló a kódolással ellentétes műveletet hajt végre: a vevő által szolgáltatott jelekből a felhasználó számára is értelmezhető formára hozza a továbbított ***üzenetet*** (231. példa).

A csatornában haladó jelekhez **elkerülhetetlenül** nem kívánt hatások, zavaró jelek, zajok is keveredenek. (Ez természetesen a többi egységben is előfordulhat.) A zaj befolyásolhatja a továbbított üzenet információtartalmát, ezzel megnehezítve vagy akár lehetetlenné téve az információátvitelt. Az információnak a hírközlés során történő részleges vagy teljes elvesztése ellen **védekezni kell**. Egyrészt a csatorna zaj elleni szigetelésével csökkenteni lehet a beáramló zavarok mennyiségét, másrészt a továbbított jelek „„zavarállósága”" fokozható **további** például ismétlődő „felesleges” **jelek beépítésével**, azaz **az üzenet redundanciájának növelésével**.

A hírközlési rendszer egyik fontos jellemzője az, hogy mekkora a hírközlés sebessége. Az időegység alatt a csatornán átjuttatható információmennyiség felső határát nevezzük csatornakapacitásnak, amelyet [bit/s] egységekben mérünk.

20.6. A kódolással kapcsolatban felmerülő problémák

Az információ sokféle alakban jelenhet meg, de minden csatorna csak jól meghatározott típusú jeleket tud továbbítani, ezért az üzeneteket úgy kell kódolni, hogy a rendelkezésünkre álló csatornán átvihetők legyenek. Mindezt úgy kell végrehajtani, hogy az üzenet minél gazdaságosabban, minél rövidebb idő alatt és minél kevesebb veszteséggel jusson el a csatorna másik végébe. **A redundancia csökkentésével növelni lehet az átvitel sebességét, de a redundancia növelésével javítani lehet az átviteli biztonságot.** A **csatornakapacitás** fogalma úgy is megadható, hogy mekkora az az információmennyiség, amelyet egy adott csatornán **optimális kódolás** mellet időegység alatt át lehet vinni.

A 230. megjegyzésben már említettük, hogy például a Morse-ábécé szöveges információ átvitelét teszi lehetővé. Ilyenkor a legegyszerűbb kódolási eljárás a betűnkénti kódolás. A közlemény minden egyes betűjéhez hozzárendeljük az illető betű kódját. A kódolás lényege tehát az, hogy van **két jelkészlet és egy leképezési utasítás** (függvény). Információtartalmától függetlenül teljesen mindegy, hogy egy üzenetet milyen jelekkel reprezentálunk, így a különböző jelsorozatok akár **számokká** is alakíthatók. Ismét a Morse-jeleket idézve kézenfekvőnek tűnik a számok **kettes számrendszerbeli** (bináris) **ábrázolása**, hiszen például a „ti”, illetve „tá” szavaknak éppen megfeleltethető a 0 és az 1 számjegy (232. megjegyzés).

Az információt sok esetben nemcsak a csatornába való juttatás kedvéért kell átalakítani, hanem például a tárolása érdekében is. Ennek alapján beszélhetünk a tárolt adat mennyiségéről. Felhasználva a kettes számrendszerbeli kódolás egyszerűségét, amelyben csak két számjegy fordul elő, az adatmennyiséget ennek segítségével definiálhatjuk. Így **a kettes számrendszerbeli számjegy** (amely 0 vagy 1 lehet) tekinthető **az adatmennyiség legkisebb egységének**. Ezt ugyanúgy, mint az információmennyiség egységét bitnek nevezzük, de más angol eredet alapján („binary digit”) (vö. 20.2. rész). A bit valóban rendkívül kicsi egység, ezért célszerű ennek többszöröseit is bevezetni. A bájt (angolul „byte”) a „by eight”, azaz nyolcasával kifejezésből származik és 8 bitet jelent (233. megjegyzés).

Ezután azt vizsgáljuk meg, hogy hogyan kódoljuk a tízes számrendszerbeli (decimális) számokat bináris kódokkal. Először a betűnkénti kódolás mintájára válasszuk a számjegyenkénti kódolást. Mivel a tízes számrendszer számjegyei (0, 1, …, 9) között a legnagyobb szám a $9_{10} = 1001_2$, ezért minimálisan 4 bit szükséges az egyes számjegyek (bináris) kódolásához. Eszerint például az

|1|4|9|7| szám kódolásához → a |0001|0100|1001|0111| sorozat 16 bitje szükséges. Azonban, ha a decimális számjegyeket kettes csoportokba (blokkokba) foglaljuk, akkor egy-egy kettes csoport 0-tól 99-ig terjedhet és ily módon ugyanannak az

|14|97| számnak a kódolásához → a |0001110|1100001| sorozat 14 bitje is elegendő. Ez azt jelenti, hogy a második módszer gazdaságosabb, hiszen az előző 4 bit/számjegy helyett itt csak 3,5 bit jut egy számjegyre. Elvileg a csoportokba foglalás (blokkosítás) folytatható, de ennek is van egy felső határa mégpedig a $\log_2 10 \approx 3{,}322$, amely nem más, mint a 10 számjeggyel kifejezhető maximális információmennyiség.

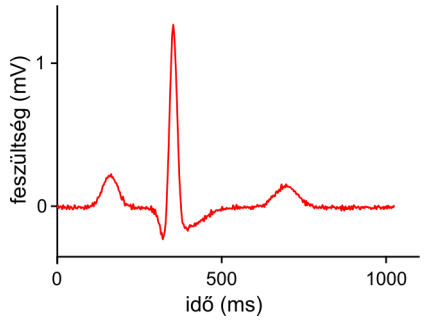
Láthattuk, hogy az ún. blokkonkénti kódolással azt nyerjük, hogy ugyanannak az információnak a megjelenítéséhez kevesebb adatmennyiség szükséges, de veszítünk is, hiszen egyetlen bit hiba esetén nemcsak egy számjegy, hanem az egész blokk sérülni fog. Például a 14 bites sorozat esetén |0001110|0100001| → |14|33|.

Kódolni azonban nemcsak szövegeket kell, hanem például képeket és hangokat is, tehát térbeli, illetve időbeli változásokat, általánosabban mondva mindenféle jeleket. A jel igen tág fogalom, például az ebéd illata és az útmentén előtűnő közlekedési jelzőtábla eléggé különböznek egymástól. Azonban minden jelben van valami közös és ez maga a változás ténye. A változásokat függvények segítségével fordíthatjuk le a matematika nyelvére, ami azért célszerű, mert utána sokkal több lehetőségünk nyílik azok elemzésére, feldolgozására.

Példaként tekintsünk egy EKG-jelet, amely egy nyugalomban lévő ember esetében például a két kar között mért elektromos feszültség az idő függvényében ($U(t)$ függvény; 234. ábra). Ez ún. analóg jel vagy másképen mondva folytonos jel, mert mind az idő, mind a feszültség folytonos változó. Ahhoz, hogy ezt a jelet például binárisan kódolni tudjuk, **mindkét változót diszkrétté kell alakítanunk**,

232. megjegyzés
A kettes számrendszer használatának további előnyeiről még szót ejtünk (vö. 246. megjegyzés). Itt csak arra emlékeztetünk, hogy a kettes (bináris) számrendszer semmivel sem bonyolultabb, mint a tízes (decimális), sőt még egyszerűbb is annál. Az viszont kétségtelen tény, hogy használata a mindennapi életben szokatlanabb és kevésbé elterjedt.
A tízes számrendszerben már megszoktuk, hogy egy szám úgy van felírva, hogy a számjegyeket a helyiértékükkel együtt kell figyelembe venni. Például a $389_{10} = 3\cdot10^2 + 8\cdot10^1 + 9\cdot10^0$. Ehhez hasonlóan $110_2 = 1\cdot2^2 + 1\cdot2^1 + 0\cdot2^0 = 6_{10}$.
Az összeadás abból áll, hogy: $0+0=0$; $1+0=1$; $0+1=1$; $1+1=10$; de a szorzás sem bonyolultabb: $0\cdot0=0$; $1\cdot0=0$; $0\cdot1=0$; $1\cdot1=1$.
Az talán a legfurcsább, hogy kettes számrendszerben már a 2-es szám is 10_2 alakú.

233. megjegyzés
A bit többszörösei még a kbit, Mbit, Gbit, stb.
A bájt jele „B”, amelynek többszörösei kB, MB, GB, stb. Itt problémát az okozhat, hogy például a kilo prefixum eredetileg 1000-et jelent, de amiatt, hogy $2^{10} = 1024$, sok esetben az egyszerűbb átszámítás kedvéért ezt használják váltószámként.



234. ábra
EKG jel ($U(t)$ függvény) nagyjából egy periódusa.

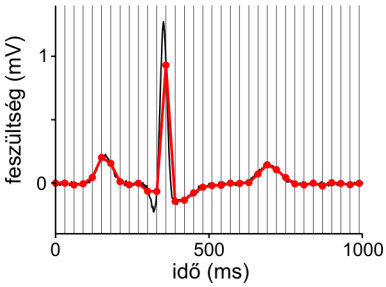
235. megjegyzés
A „digit” előtag a latin „digitus” (jelentése ujj) szóból származik. A digitális vizsgálat például az orvostudományban az ujjakkal történő tapintást jelenti. A „digit” szó az angolban az ujjon kívül számot és számjegyet is jelent, ami valószínűleg az ujjakkal való számolásból eredeztethető.

azaz az analóg jelből digitális jelet kell előállítanunk. Ez a művelet a digitalizálás (235. megjegyzés).

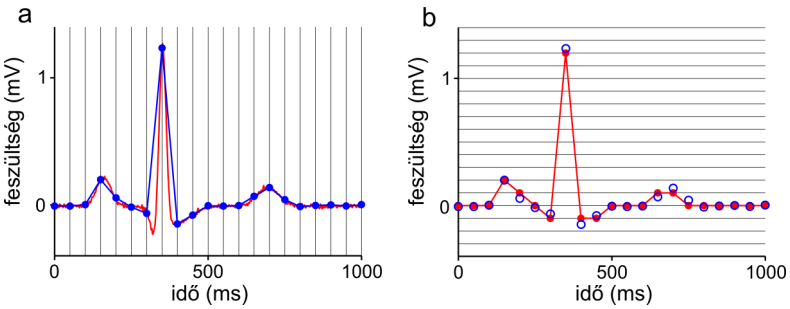
20.7. Analóg jelek digitális jelekké alakítása

Maradjunk az előbbi példánál, az EKG-jelnél, de tekintsük egy kicsit általánosabb formában. Ez azt jelenti, hogy az $U(t)$ függvény helyett vehetnénk bármilyen más $f(x)$ folytonos függvényt is.

Először a folytonos-diszkrét átalakítást az abszcissza (vízszintes tengely) mentén végezzük el. Ez egyszerűen azt jelenti, hogy a folytonos függvény függvényértékeit csak bizonyos lépésközönként olvassuk le és ezeket a leolvasott számokat rögzítjük. Ezt a folyamatot mintavételezésnek nevezzük (236a. ábra). Mivel a kapott számok akár végtelen tizedes törtek is lehetnek, amelyekhez még nem rendelhetők véges bináris kódsorozatok, ezért a folytonos-diszkrét átalakítást az ordináta (függőleges tengely) mentén is el kell végeznünk. Ez a művelet a kvantálás. Ennek során a függvény értékkészletét tartományokra osztjuk és az osztópontoknak megfelelően jelszinteket állapítunk meg. A mintavételezéskor kapott számokat a legközelebbi ilyen értékre kerekítjük (236b. ábra). Csak megjegyezzük, hogy ezzel a lépéssel akaratunktól függetlenül az eredeti jelhez zajt adunk, hiszen a kerekítés folytán egyes esetekben csökken, más esetben növekszik az eredeti függvényérték.



237. ábra
A 234. ábrán látható EKG-jel (fekete görbe) gyakoribb, 30 ms-onkénti mintavételezése.



236. ábra
A 234. ábrán látható EKG-jel (piros görbe) digitalizálásának lépései: a) az 50 ms-onkénti mintavételezés eredményeként kapott új függvény (kék pöttyök); b) a 0,1 mV-onkénti kvantálás elvégzése után kapott digitalizált jel (piros pöttyök). (A b) részben a kék karikák az ábra a) részében feltüntetett kék pöttyöknek felelnek meg.)

E két lépés elvégzése után előáll a digitalizált jel. Az átalakítással azonban sok esetben nem lehetünk teljes mértékben elégedettek, ugyanis a digitalizált jelből rekonstruált jel – még ha a hozzáadott zajtól el is tekintünk – nem hasonlít eléggé az eredeti analóg jelhez (vö. 236. ábra; piros görbék). Azt gondolhatjuk, hogy a problémát a gyakoribb mintavételezés megoldja (237. ábra). Bár az ábrán a javulás jelei felfedezhetők, a gyorsabb változások erről is lemaradtak. A kérdés tehát az, hogy milyen gyakran vegyünk mintát.

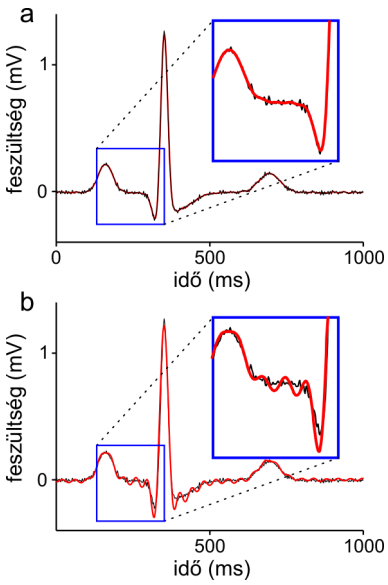
Harry Nyquist (1889 - 1976) svéd születésű amerikai elektromérnök.

Ehhez ad útmutatást a Nyquist–Shannon mintavételezési tétel. Eszerint **a mintavételezett jelből akkor állítható vissza információ veszteség nélkül az eredeti analóg jel, ha az eredeti jel leggyorsabban változó részét is legalább kétszer mintavételezzük**. Felhasználva **Fourier tételét, miszerint minden jel előállítható szinuszos jelekből**, továbbá feltéve, hogy ezek között van legnagyobb frekvenciájú (f_{max}) komponens, akkor azt is mondhatjuk, hogy a **mintavételezési frekvencia ennek a legnagyobb frekvenciának legalább a kétszerese kell, hogy legyen**. A $2f_{max}$ egyben az optimális mintavételezési frekvencia is.

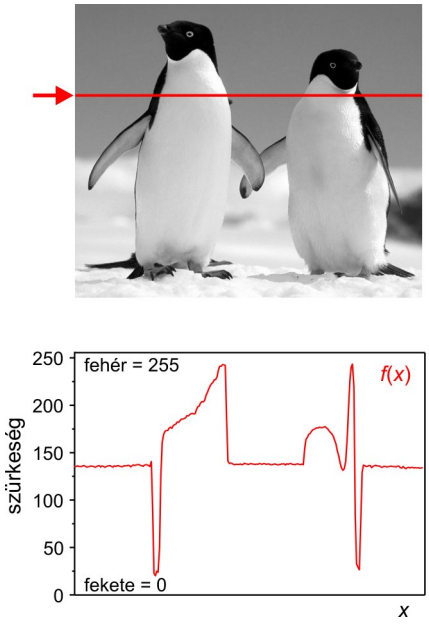
Jean Baptiste Joseph Fourier (1768 - 1830) francia matematikus és fizikus.

A Fourier-transzformációnak – nevezetesen annak, ha az $U(t)$ függvény helyett az $A(f)$ függvényt, azaz az egyes szinuszos komponensek amplitúdóját adjuk meg a frekvencia függvényében – további előnye, hogy segítségével csökkenthetjük a zajt, illetve szükség esetén a redundanciát is. A zaj csökkentésére láthatunk példát a 238. ábrán. A módszer lényege az, hogy a Fourier-transzformáció elvégzése után a nagy frekvenciájú komponenseket kihagyva végezzük el a visszatranszformálást, ezáltal „simább” lesz az visszakapott jel. Vigyázni kell azonban arra, hogy a jel „értékes” részét már ne érintse a nagyfrekvenciájú komponensek elhagyása, ellenkező esetben ugyanis éppen ezzel torzíthatjuk az eredeti jelet.

Mivel az előbbiekben sehol sem használtuk ki azt, hogy az EKG-jel időtől függő (a változás történhet térben is), általánosan bármilyen folytonos $f(x)$ függvényre



238. ábra
Zajcsökkentés Fourier-transzformáció segítségével: a) a zajos jel „kisimul” és a nagyfrekvenciájú komponensek elhagyása nem okoz torzulást; b) a túlzásba vitt „simítás” torzuláshoz vezet.

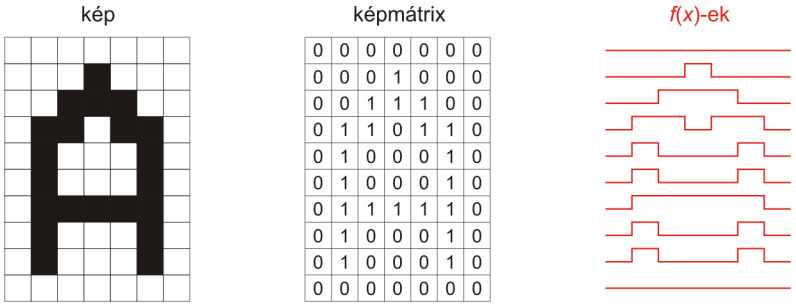


239. ábra
Képek digitalizálása. A szürkeség mint a hely függvénye ($f(x)$) a piros nyíl magasságában (piros görbe).

alkalmazhatjuk az átalakítást. Így például képeket is digitalizálhatunk. A 239. ábra azt szemlélteti, hogy például a piros nyíllal jelölt egyenes mentén hogyan változik a kép szürkesége a hely függvényében ($f(x)$). (Az csak a véletlen műve, hogy a görbe alakja emlékeztet az EKG görbére.) Természetesen a teljes kétdimenziós kép ilyen függvények sokaságaként írható fel, ami a módszer alkalmazhatóságát nem érinti, csak a műveletek elvégzése tovább tart.

A digitalizálás eredményeként képelemeket vagy képpontokat (angolul pixel a „picture element”-ből) kapunk. A képpontok együttese matematikai értelemben sorokban és oszlopokban elrendezett, a szürkeségnek megfelelő számok, azaz egy mátrix. Az ilyen mátrixokkal már mindenféle műveletet végezhetünk, így akár ki is vonhatjuk egymásból őket (vö. 1. példa).

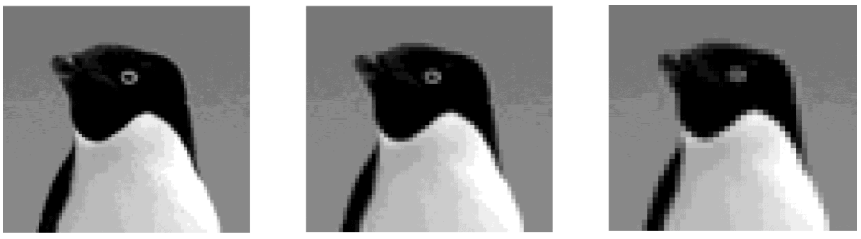
A legegyszerűbb képek a bináris (fekete-fehér) képek, ahol a „szürkeség” csak két értéket vehet fel. Ilyen képre láthatunk példát a 240. ábrán, ahol a képmátrixot és az $f(x)$ függvényeket is feltüntettük.



240. ábra
Egy „A” betű bináris digitalizált képe. Természetesen a 0-k és 1-ek felcserélésével ugyanilyen alakzatot kapunk (pozitív illetve negatív kép). (A 239. ábrával akkor kerülünk összhangba, ha a képmátrixban 0 helyett 255-öt, 1 helyett pedig 0-t írunk.)

Arról már szó esett, hogy a jelsorozatok általában redundánsak. Elég nyilvánvalóan redundáns az írott szöveg: nem lehet benne akármilyen karaktersorozat, csak olyanok, amelyekben értelmes szavak értelmes mondatokat alkotnak. A szavakon belül pedig magánhangzó után nagyobb valószínűséggel áll mássalhangzó és fordítva (vö. 4.0. rész). Hasonlóan redundáns lehet egy kép is. Például a 240. ábrán nagyobb összefüggő fehér és kisebb, de összefüggő fekete területek váltakoznak, tehát sokkal kisebb a váltás valószínűsége, mint annak, hogy egy fehér képelemet fehérek, feketét pedig feketék vesznek körül. A redundancia ebben az esetben úgy csökkenthető, hogy ismétlődő részeket keresünk és ezeket blokkokként kezeljük.

A képek digitalizálásánál a ritkább mintavételezés a kép felbontását ronthatja (241. ábra).



241. ábra
A 239. ábrán látható kép egy részlete különböző felbontásban. Az egyes képek alatta a vízszintes és a függőleges képelemek (pixelek) száma szerepel.

21.0. Az irányítás alapelvei

Az irányítás olyan tevékenység, amelynek során beavatkozunk az irányított rendszer működésébe, s ennek eredményeként a rendszer a külső feltételektől függetlenül az irányító céljainak megfelelően működik. Minden irányítás információtovábbításon alapul. Az irányítóba információ érkezik és onnan információ távozik. A beérkező információ származhat az irányított rendszeren kívüli forrásokból, de magából az irányított rendszerből is. Az irányítóból viszont az információ mindig az irányított rendszerbe áramlik. Külső forrásból érkező információ lehet pél-

dául a rendszer környezetéből érkező zavaró hatás is, amely nem kívánt módon változtathatja meg a rendszer állapotát.

Másképpen fogalmazva az irányítás célja az irányított rendszer valamely paraméterének – az irányított jellemzőnek – az előírt módon történő befolyásolása (például a hőmérséklet adott értéken tartása). Az irányítás során az irányított rendszer olyan jellemzőjére hatunk – ez a módosított jellemző – amely a leginkább alkalmas az irányított jellemző megváltoztatására (például a hőmérsékletszabályzásnál ez a fűtőteljesítmény). Az irányított jellemző olyan hatásoktól is függhet, amelyek nincsenek az irányító felügyelete alatt. Ezek a hatások a zavaró jellemzők. Az irányítás egyik fő feladata, hogy a zavaró jellemzők hatását csökkentse vagy megszüntesse (például a hőmérsékletszabályzás esetén ilyen zavaró hatás lehet a környezeti hőmérséklet ingadozása).

Az irányítás során a rendszer működésébe való beavatkozás alapvetően két formában mehet végbe: vezérlés vagy szabályozás útján. Vezérlésnek nevezzük az olyan irányítást, amelynél az irányítóhoz nem jut információ az irányított rendszerből (nincs visszacsatolás). A szabályozás éppen ebben tér el a vezérléstől, hiszen ilyenkor a visszacsatolásnak igen fontos szerepe van (242. megjegyzés).

Értéktartó a szabályozás, ha a paraméter értékét változatlan szinten akarjuk tartani. Követő szabályozásnál a paraméter értéke az éppen aktuális követelménynek megfelelően változik (243. megjegyzés). Az esetek jelentős részében a paraméter változása előre meghatározott program szerint megy végbe (vö. 21.1. rész).

242. megjegyzés
A **vezérlést** nyílt hatásláncú irányításnak is szokás nevezni, hiszen az irányított jellemző nincs (közvetlen) hatással az irányítási folyamatra. A **szabályozás** a zárt hatásláncú irányítás, ahol az irányított jellemzőnek a célkitűzéstől való eltérését használjuk fel magának az eltérésnek a csökkentésére, megszüntetésére (negatív visszacsatolás).

243. megjegyzés
A biológiai rendszerekben megvalósuló irányítási formák közül külön jelentősége van az adaptív szabályozásnak. Például az éleslátás érdekében a szemlencse törőképességének megváltoztatásával alkalmazkodik a szem.

21.1. Alapismeretek a számítógépekről

Bár a számítástechnika az élet csaknem minden területére behatolt, mégis azt tapasztalhatjuk, hogy a felhasználók jelentős része nincsen tisztában a számítógépek alapvető működési elveivel. A következőkben ezeket az alapelveket tekintjük át.

Kezdjük egy példával és oldjuk meg zsebszámológép segítségével a

$$2x^2 - 8x + 6 = 0 \qquad (185)$$

másodfokú egyenletet, pontosabban elégedjünk meg a nagyobbik gyök kiszámolásával. A megoldáshoz a következő ismeretekkel rendelkezünk:

1. a megoldóképlet,

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \qquad (186)$$

(és benne a betűk aktuális jelentése: $a = 2$; $b = -8$; $c = 6$);

2. a zsebszámológép gombjainak funkciói.

A gombok lenyomásának ***egy jó sorrendje*** lehet a következő:

$$8 \times 8 - 4 \times 2 \times 6 = \sqrt{} = +8 = \div 2 \div 2 = , \qquad (187)$$

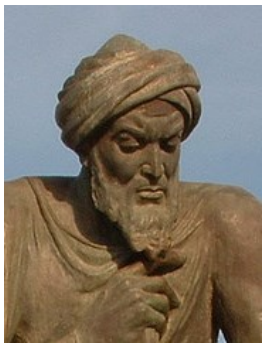
amely nem az egyetlen jó sorrend, de eredményül megkapjuk a nagyobbik gyököt, 3-at.

Ezen a példán keresztül a következőket figyelhetjük meg:

1. Van egy ***műveletsorozat***, amely előírja, hogy mit kell csinálnia a számológépnek.
2. Vannak számok, ***adatok*** (vö. 1.0. rész), amelyeken a számológép végrehajtja a műveleteket.

Általánosan az **adott cél elérése érdekében végzett műveletsorozatot** (vagy tevékenységsorozatot) algoritmusnak nevezzük. Az előbbi feladat más algoritmussal is megoldható, viszont ha már van egy jó algoritmusunk, akkor a hasonló feladatok (azaz tetszőleges együtthatójú másodfokú egyenletek) egész sorát könnyedén megoldhatjuk (csak a megfelelő pirossal jelölt számokat kell értelemszerűen kicserélni.).

Nagyon sok tevékenység algoritmizálható, sőt inkább úgy kell fogalmaznunk, hogy tevékenységeink nagyobb részét valamilyen algoritmus szerint végezzük. Például a fogmosás, az ételek elkészítése mind-mind ilyen tevékenység. A szakácskönyvekben „algoritmusokat” olvashatunk, pontosabban a hozzávalók a be-



Abu Abdalláh Muhammad ibn Músza al-Hvárizmi (780? - 845?) arabul alkotó perzsa tudós, matematikus. Ő vezette be az algoritmus fogalmát (ami miatt néhányan a számítástechnika nagyapjának nevezik), és maga az „algoritmus” szó is nevének eltorzított latin változatából ered.

244. példa

Pénzfelvétel bankautomatából:

1. Helyezze a bankkártyát a „bankkártya” feliratú nyílásba!
2. Írja be a titkos kódját!
3. Nyomja meg a „pénzfelvétel” menüpontot!
4. Válassza ki a kívánt összeget!
5. Vegye el a kártyát!
6. Vegye el a pénzt!

Neumann János (1903 – 1957) magyar származású matematikus.

245 ábra
A sokfiókos szekrény mint a processzor munkaterülete.

246. megjegyzés
Neumann-elvek (az elektronikus számítógépekkel szembeni követelmények):

- Legyen soros működésű, teljesen elektronikus. A gép egyszerre csak egy műveletet vesz figyelembe és hajt végre.
- Használjon kettes számrendszert. Elektronikusan ezt igen könnyű megvalósítani: van áram (1), nincs áram (0).
- **Tárolt program elve**. Az algoritmus is kódolható számokkal, tehát ugyanúgy kezelhető, mint az adatok. Ezek mind egy belső memóriában tárolhatók. Azáltal, hogy a számítógép belső memóriájában utasításokat is tárolhat, a számítógép önállóan képes dolgozni, mivel mindegyik lépés után memóriája utasítja a további teendőkre anélkül, hogy emberi beavatkozásra kellene várnia. Ezen túlmenően lehetőség nyílik a program módosítására annak végrehajtása közben is.
- Legyen univerzális, a speciális feladatokhoz ne kelljen más-más gépet készíteni.

menő adatok és az elkészítés módja az algoritmus. Az algoritmus legfontosabb jellemzőinek tanulmányozása érdekében vegyünk egy másik példát, nevezetesen a bankautomatából történő pénzfelvételt (244. példa).

Ebből a példából az alábbi következtetéseket lehet levonni:

1. Az algoritmus a leírás alapján végrehajtható, de a végrehajtónak természetesen sok mindent tudnia kell. (Például értenie kell a nyelvet, tudnia kell a titkos kódját stb. azaz ismernie kell az elemi részalgoritmusokat.)

2. Az algoritmus lépésekre bontva a megfelelő sorrendben hajtandó végre. Ezt a lépéssorozatot végrehajtás közben az algoritmus által leírt folyamatnak (processzusnak), végrehajtóját pedig processzornak hívjuk (esetünkben ez most a pénzfelvevő ember).

3. Az algoritmus elegendően pontos, egyértelmű és félreérthetetlen, de túlzott részletezés nincs benne.

4. Véges sok elemi lépés vezet el a célig.

5. Ugyanazt az algoritmust különböző módon is leírhatjuk (például másik nyelven), azaz különbözőképpen kódolhatjuk (vö. 20.5. rész).

A 20.6. részben már említettük, hogy a különböző jelsorozatok számokká alakíthatók. Neumann Jánostól származik az az ötlet, hogy mind az adatokat, mind az algoritmust kódoljuk azonos formában, számokkal.

Ezután vegyünk egy olyan sokfiókos szekrényt, amelynek a fiókjai sorra meg vannak számozva és a kódolás eredményeként kapott számokat megfelelő sorrendben tegyük a fiókokba (245. ábra).

Tehát minden ***fiókon*** van egy sorszám és minden ***fiókban*** van egy szám, amely lehet adatkód vagy annak részlete, illetve utasításkód vagy annak részlete. (A részlet azért van kihangsúlyozva, mert elképzelhető, hogy egy kód csak több fiókban fér el.) Ez a fiókos szekrény lesz a processzor munkaterülete. A processzor minden fiókot közvetlenül elérhet függetlenül attól, hogy az mit tartalmaz (adatkód- vagy utasításkód-részletet). Mivel a szekrényen belül az utasítások és az adatok már nem különböztethetők meg, ezért a processzornak meg kell mondanunk, hogy hol van a legelső utasítás (hányas számú fiókban kezdődik). A processzor ezután kihúzza a megadott sorszámú fiókot, kiolvassa a tartalmát, utasításként értelmezi. Ez az utasítás tartalmazhatja további fiókok sorszámait is, amelyekben például az utasítás folytatása vagy az utasítás elvégzéséhez szükséges adatok lehetnek. Az utasítás végrehajtása után, ha az nem rendelkezik másképpen, a processzor visszatolja a fiókot és kihúzza a következő, eggyel nagyobb sorszámú fiókot és így szisztematikusan halad végig.

A fiókos szekrényt nevezzük memóriának, a fiók a tárolórekesz, a ***fiókon*** lévő sorszám a tároló rekesz címe, a ***fiókban*** lévő szám, a tároló rekesz tartalma. A memória nagyságát a tárolórekeszek számával szokták jellemezni. Amennyiben a kódolást kettes számrendszerbeli számokkal végezzük, akkor általában egy tárolórekesz (egy fiók) egy bájtot jelent.

Ezek szerint a számítógépnek tartalmaznia kell legalább egy **processzort**, kell, hogy legyen **memóriája** (ebben számokká alakítva lesznek tárolva az adatok és az algoritmus is) és ezen kívül szükségesek még a környezettel való kapcsolattartás céljából a megfelelő **kommunikációs egységek**. Ezeket együttesen, azaz mindazokat a készülékeket, berendezéseket, egységeket, amelyekből a számítógép összetevődik, összefoglaló néven hardvernek nevezzük.

Manapság a számítógép elektronikus digitális számítógépet jelent, amely elektronikus eszközökből épül fel. Ezek között is talán a legfontosabbak a logikai áramköri elemek, amelyekre még visszatérünk (vö. 21.2. rész). A hardver párjaként szokták emlegetni a szoftvert, amely a számítógépen futó megfelelően kódolt algoritmusok, azaz programok összességének átfogó neve (246. megjegyzés).

A hardver szoftver nélkül nem sokat ér. Egy nagy számítógép bekapcsolás után csak a legprimitívebb alapfunkciókat képes elvégezni. Ezért létrehoztak úgynevezett rendszereket, operációs rendszereket, amelyek komfortosabbá teszik a számítógépet azáltal, hogy már elvégeznek bizonyos rutinműveleteket is. Ilyen például

247. megjegyzés
A kettes számrendszerbeli számokat sok esetben célszerű 16-os (hexadecimális) számrendszerbe átalakítani. Ez eléggé egyszerű, hiszen $2^4 = 16$, tehát amennyiben négyesével olvassuk ki a kettes számrendszerbeli számokat, akkor hexadecimális számokat kapunk. Ennek az az előnye, hogy tömörebb írásmódot tesz lehetővé. 0-tól 9-ig a számjegyek ugyanazok, mint a tízes számrendszerben, 10-től kezdve a következő jelölést használjuk:

1010 = A
1011 = B
1100 = C
1101 = D
1110 = E
1111 = F

az, hogy megkeresi a megfelelő programot egy külső memóriaegységen és azt betölti a belső memóriába, vagy ha kell, lemásolja. Egy rendszer az tulajdonképpen egy program, amit a nagy számítógépeken előre be kell tölteni a belső memóriába és természetesen minél többet tud (minél intelligensebb), annál több helyet foglal el. A kisebb számítógépek egy részénél, a rendszer vagy annak legfontosabb része fixen tehát kitörölhetetlenül be van építve a memóriába. Erre a memóriaterületre írni nem lehet csak kiolvasni, ezért ezt ROM-nak (angolul „Read-Only Memory”), a másik típusú memóriát, ahová írni is lehet RAM-nak (az angol „Random Access Memory” rövidítéseként, amelynek jelentése nagyjából „tetszőleges hozzáférésű memória”) nevezzük.

A program megírása nem más, mint az algoritmus megfelelő kódolása. Ennek legprimitívebb változata az, amikor azokat a számokat írjuk le sorjában, amelyek a memória tároló rekeszeibe kerülnek. Ezt hívjuk gépi kódú programozásnak. Ezt a processzor közvetlenül megérti.

Ha a gépi kódok (általában kettes számrendszerbeli számok; 247. megjegyzés) helyett velük kölcsönösen egyértelmű kapcsolatban lévő szimbólumokkal kódoljuk az algoritmust, akkor ezt szimbolikus gépi kódnak vagy assembler-szintű programozásnak nevezzük. Ilyenkor azonban már szükségünk van egy olyan segédprogramra, ún. fordítóprogramra, amelyik a szimbólumokat számokká visszaalakítja.

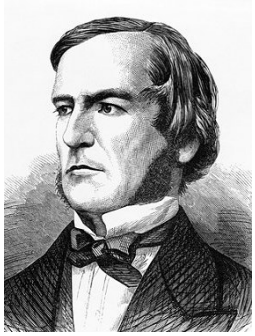
Az algoritmust azonban nem kell nekünk lebontanunk a legelemibb lépésekig akkor, ha van egy olyan fordító programunk, amelyik ezt elvégzi helyettünk. Így csak bizonyos részalgoritmusoknak kell megfeleltetnünk szimbólumokat és ezekkel kódoljuk a teljes algoritmust. Az így nyert szimbólumok összességét hívjuk algoritmikus programozási nyelvnek. Az, hogy milyen részalgoritmusokat célszerű egy szimbólummal jelölni, elsősorban a feladat típusától függ. Ezért sokféle programozási nyelv jött létre.

21.2. Matematikai logikai alapok

248. megjegyzés
Az „Őrült sikerem a tébolydában” című művében így ír Karinthy Frigyes:
„Ohó álljunk csak meg. Ön azt mondja, a rögeszmém, hogy őrült vagyok. De hiszen tényleg az vagyok, az imént mondta. De hiszen akkor ez nem rögeszme, akkor az egy logikus gondolat. Tehát nincs rögeszmém. Tehát mégse vagyok őrült. Tehát csak rögeszme, hogy őrült vagyok, tehát rögeszmém van, tehát őrült vagyok, tehát igazam van, tehát nem vagyok őrült...”

Az előző részben már említettük, hogy a számítógépben végbemenő folyamatok irányítását a processzor végzi. Ennek egyik fontos része az aritmetikai és logikai egység, amely arra képes, hogy a legelemibb számolási és logikai műveleteket végrehajtsa. Ez elegendő is, hiszen ilyen elemi műveletek segítségével elvileg bármely számítási feladat már elvégezhető. **A számítógép attól különleges, hogy az említett egyszerű műveleteket nagy sebességgel, nagy pontossággal és megbízhatóan képes akár igen hosszú időn keresztül is sokszor végrehajtani**. A legegyszerűbb kettes számrendszerbeli aritmetikai műveleteket a 232. megjegyzésben már megemlítettük, itt a legegyszerűbb logikai műveleteket tekintjük át.

A matematikai logika ún. kétértékű logika, ami azt jelenti, hogy egy megállapítást akkor tekintünk **állításnak**, ha eldönthető róla, hogy **igaz** vagy **hamis** (248. megjegyzés). (Tehát egy állítás nem lehet igaz és hamis is egyszerre, de az sem lehet, hogy egyszerre se nem igaz, illetve se nem hamis.) Az ***igaz*** logikai érték az **1**, a ***hamis*** a **0**. Az ilyen típusú változót a binárison kívül (vö. 95. megjegyzés) szokás még logikai vagy Boole-féle változónak is nevezni. A logikai változók (X, Y) közötti összefüggéseket logikai függvényeknek is nevezhetjük.



George Boole (1815 - 1864) angol matematikus és filozófus.

Egyváltozós logikai művelet a *NEM* művelet vagy tagadás (*NOT*; negáció), amely 1-hez 0-t, 0-hoz 1-et rendel. Jelölése: $\neg X$; (*NEM* X; de szokás felülvonással is jelölni). A kétváltozós logikai műveletek közül a két legfontosabb a *VAGY* művelet, vagy logikai összeadás (*OR*, diszjunkció), illetve az *ÉS* művelet, vagy logikai szorzás (*AND*, konjunkció). Jelölések: $X \vee Y$, (X vagy Y); $X \wedge Y$, (X és Y). Ezeket a műveleteket legegyszerűbben az ún. igazságtáblájukkal definiálhatjuk (vö. 3.2. rész; 17. megjegyzés):

X	Y	$\neg X$	$X \vee Y$	$X \wedge Y$
0	0	1	0	0
0	1	1	1	0
1	0	0	1	0
1	1	0	1	1

249. táblázat
A legegyszerűbb logikai műveletek igazságtáblája: *NEM* ($\neg$), *VAGY* ($\vee$), *ÉS* ($\wedge$).

Belátható, hogy a bonyolultabb logikai műveletek mindig lebonthatók csupán az említett három művelet felhasználásával.

Mintafeladat

Határozzuk meg, hogy az ***x*** változó mely értékeire ad ***igaz*** eredményt az alábbi kifejezés:

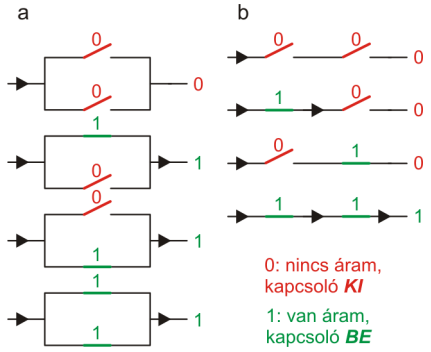
$$((x < 5) \wedge (x \geq 2)) \vee ((x > 4) \wedge (x \leq 6))\ ?$$

Megoldás: Az első *ÉS* művelet eredménye: $2 \leq x < 5$; a második *ÉS* művelet eredménye: $4 < x \leq 6$. A két eredményt *VAGY* művelet kapcsolja össze, így a végeredmény: $2 \leq x \leq 6$.

Mintafeladat

Mi a következő kifejezés logikai értéke: $((2 \leq 3) \wedge (4 > 5)) \vee ((3 \leq 4) \wedge (4 < 5)) = ?$

Megoldás: A zárójeleken belüli kiértékelés után kapjuk: $(1 \wedge 0) \vee (1 \wedge 1) = 0 \vee 1 = 1$, tehát a fenti kifejezés logikai értéke 1.



250. ábra
Egyszerű logikai „áramkörök” megvalósítása kapcsolókkal: a) *VAGY* „áramkör”; b) *ÉS* „áramkör” (vö. 249. táblázat).

251. ábra
A Pannonhalmi Bencés Főapátság könyvtára mint adatbázis.

252. megjegyzés
Ha például van két halmazunk D_1 és D_2, akkor ezek direkt szorzata a D halmaz, amelynek elemeit úgy állítjuk elő, hogy veszünk egy elemet a D_1 halmazból, egyet a D_2-ből, és ebben a sorrendben egymás után írjuk őket. Ha ezt a műveletet minden lehetséges módon elvégezzük, akkor megkapjuk a D halmaz összes elemét. A D halmaz egy R részhalmazát relációnak nevezzük.

A matematikai logika címszó alatt gyakran szerepel a Boole-algebra kifejezés, amelyre joggal tekinthetünk úgy, mint a digitális számítógépek kifejlesztésének egyik elvi alapjára, hiszen olyan bonyolultabb állítások valóságtartalmát vizsgálja, amelyek helyes vagy hamis elemi állításokból tevődnek össze.

A Boole-algebra másik interpretációja az ún. kapcsolási algebra. Alapjául olyan kapcsolási elemek (áramkörök) szolgálnak, amelyek csupán két, egymástól különböző állapotot vehetnek fel, például egy áramkörben vagy folyik áram, vagy nem. A kapcsolási algebra azt vizsgálja, hogy az ilyen kapcsolási elemekből összeállított hálózat kimenetén a lehetséges két állapot melyike valósul meg, ha az elemek az egyik vagy másik lehetséges állapotban vannak. Ezért a **Boole-algebra** egyben az **elektronikus digitális számítógép konstruálásának** nélkülözhetetlen **elméleti alapja** is (250. ábra).

22.0. Adatbázisok

Az adatbázis köznapi értelemben az adatok rendszerezett gyűjteménye, amely nem feltétlenül számítógépen kerül tárolásra (251. ábra). Az adathalmaz csak akkor válik adatbázissá, ha az olyan szisztéma szerint épül fel, amely lehetővé teszi az adatok értelmes kezelését. Így az adatbázisok legfontosabb jellemzője az, hogy **nem csak az adatokat**, hanem **az adatok közötti kapcsolatokat, összefüggéseket is képes tárolni**. Hangsúlyozzuk, hogy nem az a fontos, hogy ezt milyen módon valósítjuk meg (például kartotékokkal, papíron vagy számítógéppel): amíg az adatokat konkrét célra valamilyen rendezett formában gyűjtjük és tároljuk, adatbázisról beszélhetünk. A továbbiakban persze azt fogjuk feltételezni, hogy az adatgyűjtésre és az adatok kezelésére **számítógépet** használunk. Ennek legalább két előnye van, nevezetesen a **kisebb helyigény** és az **egyszerűbb karbantartás**.

Ebben az esetben az adatbázisokat adatbázis-kezelő szoftverrel hozzuk létre, illetve annak segítségével használjuk. Egy ilyen rendszer által az információ

- könnyen tárolható,
- rendezhető,
- visszakereshető és
- megjeleníthető.

Az adathalmaz és annak elemei között fennálló kapcsolatok strukturált leírását adatmodellnek nevezzük. Az adatmodellezés az adatbázisok tervezésének egy lehetséges módszere. Az egyik legelterjedtebb adatbázismodell a relációs adatmodell, amely halmazelméleti alapokra épül (252. megjegyzés). A modellben a reláció gyakorlatilag egy táblázat. **Minden reláció egyedi nevet kap**, sorai a logikailag összetartozó adatokat tartalmazzák, tehát egy-egy objektumot, egyedet írnak le, az oszlopaiban pedig az egyedek azonos tulajdonságaira vonatkozó adatok találhatók. Az **oszlopok szintén egyedi nevet kapnak a reláción belül**, de másik relációban már ismét előfordulhatnak.

A táblázat elemeinek elnevezése nem egységes: általában egy sor és egy oszlop metszetét cellának (ritkábban mezőértéknek vagy adatmezőnek) nevezzük, a cella tartalmazza az adatot. Az adatbázisok kapcsán reláció helyett gyakran tábla, sor helyett rekord, oszlop helyett pedig mező kifejezéssel élünk. Sok esetben a mező

253. megjegyzés
Érdemes megemlíteni, hogy a relációs adatbázis-kezelők leggyakrabban használt szabványos programozási nyelve az SQL (az angol „Structured Query Language”-ből: strukturált, több elemből felépített lekérdező nyelv).

helyett használjuk az attribútum vagy tulajdonság megnevezést is. A relációs adatmodell lényege tehát az, hogy az **egyedeket, a tulajdonságokat és a kapcsolatokat egyaránt táblázatok**, adattáblák (relációk) segítségével kezeli (253. megjegyzés). Az adatbázis ebben a modellben a táblák összességét jelenti.

Példaként tekintsünk egy ilyen táblát (254. ábra). A rekordok egy-egy beteget, a mezők pedig egy-egy tulajdonságot adnak meg. A táblázat első sora a tábla nevét, a második a mezők nevét tartalmazza. **Ha a mezők sorrendjét vagy a rekordok sorrendjét felcseréljük, akkor a tárolt adatok helyessége nem sérül**.

Beteg				
TAJ szám	Név	Születési dátum	Kezelő orvos	Kórterem
232 861 245	Kond Előd	1961.03.21	Dr. Kiss	3
152 351 254	Tas Huba	1972.04.22	Dr. Kiss	2
342 864 213	Álmos Emese	1971.05.23	Dr. Nagy	1

rekord
mező
cella

254. ábra
A „Beteg” tábla, amelyen kiemeltünk egy cellát, egy rekordot és egy mezőt.

255. megjegyzés
A „Név” mellett gyakran használatos az „Anyja neve” mező is, ilyenkor a „Név”-vel együtt összetett kulcsként lehet ez is elsődleges kulcs. Ugyanis azonos nevű betegek még viszonylag gyakran előfordulhatnak akár ugyanazon a kórházi osztályon is, de olyanok, akiknek még az anyjuk neve is megegyezik, már gyakorlatilag nem.

Az viszont **egyetlen táblában sem fordulhat elő, hogy két sor teljes mértékben megegyezzen**. Ennek megfelelően minden táblában létezik egy oszlop (vagy az oszlopoknak egy olyan halmaza), amely a tábla bármely sorát egyértelműen meghatározza, azonosítja. Ezt az oszlopot vagy oszlopkombinációt hívjuk a tábla elsődleges kulcsának. Ezen kívül az adott táblában természetesen előfordulhatnak olyan oszlopok (vagy oszlopcsoportok) is, amelyek egy másik táblában elsődleges kulcsként szerepelnek. Az olyan azonosítót, amely egy másik táblában az elsődleges kulcs szerepét tölti be, de az adott táblában nem, kapcsoló kulcsnak nevezzük, de használatos még a külső vagy idegen kulcs elnevezés is. A 254. ábrán a „Beteg” táblában a TAJ szám lehet elsődleges kulcs, hiszen ebből nincs két azonos (255. megjegyzés).

Ha egy adott tábla rekordjai valamilyen módon egy másik tábla rekordjaihoz társíthatók, azt mondjuk, hogy a táblák között kapcsolat áll fenn. Két tábla között háromféle kapcsolat állhat fenn: egy az egyhez (1:1), egy a többhöz (1:*n* vagy *n*:1) vagy több a többhöz (*m*:*n*) kapcsolat. Az első kettő értelemszerűen következik a legáltalánosabb, több a többhöz kapcsolat meghatározásából.

Két tábla akkor áll több a többhöz kapcsolatban, ha az első tábla egy rekordja a második tábla több rekordjához kapcsolódik, és a második tábla egy rekordja is több rekorddal áll kapcsolatban az első táblából. Ezt a kapcsolatot egy kapcsolótábla segítségével hozzuk létre. Ily módon egyszerűen társíthatunk rekordokat az egyik táblából a másik tábla rekordjaihoz, és a kapcsolótábla arról is gondoskodik, hogy a kapcsolódó adatok hozzáadása, törlése vagy módosítása során semmilyen probléma ne lépjen fel. Például, ha egy könyvtári adatbázis két táblája az „Olvasó” és a „Könyv”, akkor ezek összekapcsolásával nyilvántarthatjuk, hogy melyik könyv kinél van (256. ábra).

256. ábra
A kapcsolótábla szerepe egy könyvtári adatbázisban.

Kölcsönzés		
Kölcsönzés-azonosító	Olvasóazonosító	Könyvazonosító
001	10102	00125
002	10102	00123
003	10101	00124
004	10101	00126
005	10103	00123

Olvasó		
Olvasóazonosító	Név	Lakcím
10101	Kovács József	9985 Felsőszölnök, Templom út 112.
10102	Asztalos János	7345 Alsómocsolád, Kossuth u. 132.
10103	Szekér Júlia	4732 Cégénydányád, Rákóczi Ferenc u. 223.

Könyv			
Könyvazonosító	Szerző	Cím	Kiadó
00123	Arany János	Toldi	Szépirodalmi Könyvkiadó
00124	Petőfi Sándor	János vitéz	Szépirodalmi Könyvkiadó
00125	Bernolák Kálmán	A fény	Műszaki Könyvkiadó
00126	Helmut Vogel	Kompakt fizika	Springer Hungarica Kiadó

A legfontosabb adatbázis-funkciók a rendezés, a szűrés és a lekérdezés. Rendezés segítségével könnyen megkereshetünk egyszerű feltételeknek megfelelő rekordokat az adattáblában.

Az adatok szűrése azt jelenti, hogy megadunk egy logikai feltételt (szűrőfeltételt), és a program csak azokat a rekordokat jeleníti meg, amelyek ennek a feltételnek eleget tesznek. Ha például egy adott mező szerinti feltételt állítottunk be, akkor már csak az annak megfelelő rekordok jelennek meg, de ezt tovább szűrhetjük egy másik mező szerinti feltétel megadásával stb.

Adott tulajdonságú adatok listázásának másik módja a lekérdezés. A lekérdezések segítségével az adatbázisból

- megjeleníthetjük,
- módosíthatjuk,
- törölhetjük

az adott feltételeknek megfelelő adatokat. A kapcsolótáblák épp azt teszik lehetővé, hogy két vagy akár több táblából egyetlen lekérdezéssel jelenítsük meg az összetartozó adatokat.

257. megjegyzés
Egy adatbázist általában többen is használnak, ezért az adatokhoz való hozzáférés felhasználói jogosultságokhoz köthető.

Fontos megemlíteni még a jelentéseket, amelyek az adatbázis adatainak rendezett, megfelelően csoportosított, nyomtatható formában történő megjelenítésére szolgálnak (257. megjegyzés).

Igen fontos adatbázis-funkciókat segítő megoldás a mezők indexelése. Például a nagyméretű, esetleg több millió rekordot tartalmazó táblák fizikai átrendezése nem célszerű, ehelyett a kiválasztott mezőkhöz indextáblát készítenek, amely az adott mező szerint tartja nyilván a rekordok sorrendjét, és szükség esetén a program ennek megfelelően jeleníti meg az adatokat. Az indextábla az adatok bevitelekor, törlésekor, módosításakor folyamatosan frissül. Az indexek alapján a rekordok gyorsan sorba rendezhetők, és egy másik index aktivizálásával könnyen át lehet állni egy másik szempont szerinti sorrendre. Az adatbáziskezelő-szoftver a kereséseket is az indextáblában végzi, amely sokkal gyorsabb, mint ha sorban kellene végignézni az összes rekordot.

A legfontosabb irodalmi források:

Reiczigel J., *Harnos A.*, *Solymosi N.*: Biostatisztika nem statisztikusoknak, Pars Kft. Nagykovácsi 2007

Prékopa A.: Valószínűségelmélet, Műszaki Könyvkiadó, Bp. 1980

Vetier A.: Szemléletes mérték- és valószínűségelmélet, Tankönyvkiadó, Bp. 1991

Farkas M.: Matematikai Kislexikon, Műszaki Könyvkiadó, Bp. 1974

Hajtman B.: Bevezetés a matematikai statisztikába, Akadémiai Kiadó, Bp. 1972

Sváb J.: Biometriai módszerek a kutatásban, Mezőgazdasági Kiadó, Bp. 1981

Juvancz I., Paksy A.: Az orvosi biometria alapjai, Medicina Könyvkiadó, Bp. 1981

Canavos G. C.: Applied probability and statistical methods, Little, Brown and Company Ltd., Boston, Toronto 1984

Dawson B., Trapp R. G.: Basic and Clinical Biostatistics, Lange Medical Books/McGraw-Hill 2001

Armitage P., Berry G.: Statistical methods in medical research, Blackwell scientific publications, Oxford 1994

Mendenhall W.: Introduction to probability and statistics, Duxbury Press, Boston 1987

Norman T., Bailey J.: Statistical methods in biology, Cambridge university press 1993

Rontó Gy., *Tarján I.* (szerk.): A biofizika alapjai 8. fejezet Semmelweis Kiadó, Bp. 2002

Damjanovich S., *Mátyus L.* (szerk.): Orvosi biofizika 11. fejezet, Medicina Könyvkiadó, Bp. 1981

Harnos Zs., *Herdon M.* (szerk.): Informatika, Debreceni Egyetem Agrár- és Műszaki Tudományok Centruma, Agrárgazdasági és Vidékfejlesztési Kar, Debrecen, 2007

http://tanulnijo.uw.hu/adatbazis/ab_tartalom.html

Név- és tárgymutató

abszolút értelemben használt változó 37
abszolút gyakoriság 4
adaptív szabályozás 86
adat 2
adatbázis 89
adatbázis kezelő szoftver 89
adatmennyiség 83
adatmező 89
adatmodell 89
adatrendszer 13
adatrendszer eloszlásfüggvénye 13
adatsűrűség 17
additív hibatag 64
adó 82
alapsokaság 33
algoritmikus programozási nyelv 88
algoritmus 86
al-Hvárizmi 86
állandó hozzáadása (eltolási transzformáció) 31
állandóval való szorzás (nyújtási transzformáció) 31
álnegatív arány 74
álpozitív arány 74
alternatív hipotézis 48
analóg jel 83
AND művelet 88
aritmetikai és logikai egység 88
assembler-szintű programozás 88
aszimptotikusan torzítatlan becslés 39
asszociáció 63
átlag 22
átlagos négyzetes eltérés 39
attribútum 90
bájt 83
Bartlett-próba 70
becslés 36
– elégséges 39
– hatásos 38
– konzisztens 38
– torzítatlan 38
becsléses illeszkedésvizsgálat 57
Bernoulli 26
bináris változók 37
Binomiális eloszlás 25
bit 80
biztos esemény 7
Boole-algebra 89
Boole-féle változó 88
cella (táblázaté) 89
centrális határeloszlás tétel 29
cím (tároló rekeszé) 87
csatorna 82
csatornakapacitás 82
csoportok (varianciaelemzésben) 69
dekódoló 82
determinációs együttható 67
determinisztikus törvényszerűség 3
diagnosztikai effektivitás 74
diagnosztikai paraméterek 74
diagnosztikai tesztek 73
dichotom változók 37
digitális jel 84
digitalizálás 84
digitalizált jel 84
diszjunkció 88
diszkrét változó 13
döntés 46
egyed 89
egyenletes eloszlás 25
– – (diszkrét) 25
– – (folytonos) 27
egyesítés (eseményeké) 7
egymást kizáró események 8
egyoldalú próba 51
egyszerű lineáris regresszió 65
együttes eloszlásfüggvény 30
együttes sűrűségfüggvény 30
elégséges becslés 39
elektronikus digitális számítógép 87
elfogadás (nullhipotézisé) 48
elfogadási tartomány 51
ellentett esemény 7
elméleti eloszlásfüggvény 35
eloszlás 12
eloszlástól független módszerek 53
elsődleges kulcs 90
elsőfajú hiba 51
elterjedtség (prevalencia) 74
eltolási transzformáció 31
elvetés (nullhipotézisé) 48
elvetési tartomány 51
értéktartó szabályozás 86
érzékenység (szenzitivitás) 74
ÉS művelet 88
esélyérték 9
esélyhányados 71
esemény 3
eseménytér 7
eset (eset-változó táblázatban) 37
exponenciális eloszlás 27
feltételes relatív gyakoriság 5
feltételes sűrűségfüggvény 30
feltételes valószínűség 11
Fisher-féle egzakt próba 57
folytonos változó 13
folytonos jel 83
folytonossági korrekció 60
fordítóprogram 88
Fourier 84
független minta 34
független valószínűségi változók 30
függetlenség 9
függetlenségvizsgálat 56
Galton 68
Gauss 28
gépi kód 88
Gosset (Student) 44
gyakorisági eloszlás 18
gyakoriságsűrűség 18
halmaz 5
halmaz elemei 5
hardver 87
hartley 80
hatásos (hatásosabb) becslés 38
hatékonyabb próba 54
hipotézisvizsgálat 36
hisztogram 18
homogenitás-vizsgálat 56
idegen kulcs 90
illeszkedésvizsgálat 56
indexelés 91
indextáblat 91
induktív statisztika 4
információ 78
– felhasználó 82
– forrás 82
információtechnológia 78
informatika 78
interkvartilis terjedelem 24
intervallum 37
intervallum becslés 43
irányító 85
irányított jellemző 86
– rendszer 85
jel 2
jelenség 3
jelentés 91
kapcsolási algebra 89
kapcsolat 90
kapcsoló kulcs 90
kapcsoló tábla 90
kapcsolt rangok 53
kategoriális változó 37
kategorizáló transzformáció 53
kétoldalú próba 51
kezelés (varianciaelemzésben) 69
kiegyensúlyozott kísérleti elrendezés 70
kimenetel 3
kísérlet 3
kísérlet entrópiája 78
kód 82
kódolás 82
kohort 72
kommunikáció 78
konfidencia 43
– határok 43
– intervallum 43
– szint 43
konjunkció (ÉS művelet) 88
kontingencia tábla 58
konzisztens 38
– próba 52
korrekt klasszifikáció 74
– negativitás 74
– pozitivitás 74
korrelációs együttható 32
korrelációs t-próba 63
korrelálatlan valószínűségi változók 32
korrigált tapasztalati szórásnégyzet 41
kovariancia 32
követő szabályozás 86

közös rész (eseményeké) 8
kritikus értékek 51
– tartomány 51
Kruskal 62
különbsége (eseményeké) 8
külső kulcs 90
kvantálás 84
kvantilis 23
kvartilis 23
legkisebb négyzetek módszere 65
lehetetlen esemény 7
leíró statisztika 4
lekérdezés 91
Levene-próba 70
logikai változó 88
logikai áramköri elemek 87
log-normális eloszlás 68
maga után von (egyik esemény a másikat) 8
magyarázó változó 64
magyarázott változó 64
Mann 62
MA-regresszió 67
másodfajú hiba 51
matematikai modell 2
medián 23
megbízhatóság 36
megfigyelés 3
megfigyelési egység 33
memória 87
mező 89
mezőérték 89
minta 33
-vétel 33
-vételezés 84
modell 2
módosított jellemző 86
módusz 24
multiplikatív hiba 68
nat 80
negáció 88
negatív korreláció 32
– visszacsatolás 86
NEM művelet 88
nemparaméteres eljárások 53
Neumann 87
nominális 37
normális eloszlás 28
normált eloszlás 12
NOT művelet 88
nulleloszlás 54
-hipotézis 48
nyílt hatásláncú irányítás 86
Nyquist 84
nyújtási transzformáció 31
operációs rendszer 87
OR művelet 88
ordinális változó 37
osztályok 18
összeg (valószínűségi változóké) 31
összetett kulcs 90
összetett megfigyelés 9
paraméter (eloszlásoké) 25
paraméteres eljárások 53
párosított minták 53
Pearson-féle korrelációs együttható 64
peremeloszlás 30
Poisson-eloszlás 26
pontbecslés 43
pozitív korreláció 32
PP ábra 57
prevalencia (elterjedtség) 74
próba ereje 52
processzor 87
program 87
program szerinti szabályozás 86
QQ ábra 57
RAM 88
rangok 53
rangsor transzformáció 53
redundancia 82
referencia tartomány 45
regresszió 33
regressziós egyenes 65
– görbe 33
rekord 89
reláció 89
relációs adatmodell 89
relatív entrópia 82
– gyakoriság 5
– gyakorisági eloszlás 18
– gyakoriságsűrűség 18
– kockázat 72
relevancia 74
rendezés 91
reprezentatív minta 33
rétegzett mintavétel 34
reziduális szórás 66
reziduum 65
robusztusság 54
ROM 88
Shannon-féle entrópia 81
Shapiro–Wilk-próba 70
SMA-regresszió 67
Spearman-féle rangkorrelációs együttható 64
specificitás 74
SQL 90
standard hiba 39
standard normális eloszlás 31
standardizálás 31
statisztika 38
– alaptétele 35
statisztikai próbák 49
statisztikus kapcsolat 32
– törvényszerűség 3
Student–eloszlás 44
sűrűségfüggvény 19
szabadságfok 41
szabályozás 86
számítástechnika 78
számszerű változó 37
szegregancia 74
szenzitivitás (érzékenység) 74
szignifikancia szint 51
szignifikáns 50
szisztematikus mintavétel 34
szoftver 87
szórás 24
szórásnégyzet 24
szorzata (valószínűségi változóké) 31
szűrés 91
t-eloszlás 44
tábla 89
tapasztalati eloszlásfüggvény 35
tároló rekesz 87
– – tartalma 87
– – címe 87
teljes eseményrendszer 8
teljes valószínűség tétel 11
tényező (varianciaelemzésben) 69
terjedelem 24
tévedési valószínűség 36
téves figyelemfelkeltő arány 74
– megnyugtatási arány 74
tiszta illeszkedésvizsgálat 57
torzítás 39
torzítatlan becslés 38
többszörös lineáris regresszió 67
t-próba 54
tulajdonság (attribútum) 89
u-próba 50
VAGY művelet 88
valószínűség 7
valószínűségeloszlás 12
valószínűségi változó 13
– – eloszlásfüggvénye 15
valószínűségszámítás 3
változó (halmaz általános eleme) 5
változó (statisztikában) 34
változók (eset-változó táblázatban) 37
várható érték 23
variancia 25
varianciaelemzés 55
véletlen hiba 39
– mintavétel 34
vetületeloszlás 30
vevő 82
vezérlés 86
visszacsatolás 86
visszatevés nélküli mintavétel 34
visszatevéses mintavétel 34
Wallis 62
Welch-próba 55
Whitney 62
Wilcoxon 61
zajok 82
zárt hatásláncú irányítás 86
zavaró jellemző 86
χ-eloszlás 42
χ^2-eloszlás 42

Statisztikai táblázatok

t-eloszlás

ν	p (egyoldalú próba)						
	0,4	**0,25**	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**
	p (kétoldalú próba)						
	0,8	**0,5**	**0,2**	**0,1**	**0,05**	**0,02**	**0,01**
1	0,325	1,000	3,078	6,314	12,70	31,82	63,65
2	0,289	0,816	1,886	2,920	4,303	6,965	9,925
3	0,277	0,765	1,638	2,353	3,182	4,541	5,841
4	0,271	0,741	1,533	2,132	2,776	3,747	4,604
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032
6	0,265	0,718	1,440	1,943	2,447	3,143	3,707
7	0,263	0,711	1,415	1,895	2,365	2,998	3,499
8	0,262	0,706	1,397	1,860	2,306	2,896	3,355
9	0,261	0,703	1,383	1,833	2,262	2,821	3,250
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169
11	0,260	0,697	1,363	1,796	2,201	2,718	3,106
12	0,259	0,695	1,356	1,782	2,179	2,681	3,055
13	0,259	0,694	1,350	1,771	2,160	2,650	3,012
14	0,258	0,692	1,345	1,761	2,145	2,624	2,977
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947
16	0,258	0,690	1,337	1,746	2,120	2,583	2,921
17	0,257	0,689	1,333	1,740	2,110	2,567	2,898
18	0,257	0,688	1,330	1,734	2,101	2,552	2,878
19	0,257	0,688	1,328	1,729	2,093	2,539	2,861
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845
21	0,257	0,686	1,323	1,721	2,080	2,518	2,831
22	0,256	0,686	1,321	1,717	2,074	2,508	2,819
23	0,256	0,685	1,319	1,714	2,069	2,500	2,807
24	0,256	0,685	1,318	1,711	2,064	2,492	2,797
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787
26	0,256	0,684	1,315	1,706	2,056	2,479	2,779
27	0,256	0,684	1,314	1,703	2,052	2,473	2,771
28	0,256	0,683	1,313	1,701	2,048	2,467	2,763
29	0,256	0,683	1,311	1,699	2,045	2,462	2,756
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750
40	0,255	0,681	1,303	1,684	2,021	2,423	2,704
60	0,255	0,679	1,296	1,671	2,000	2,390	2,66
120	0,254	0,677	1,289	1,658	1,980	2,358	2,617
∞	0,250	0,674	1,282	1,645	1,960	2,326	2,576

χ^2-eloszlás

ν	p					
	0,99	**0,975**	**0,95**	**0,05**	**0,025**	**0,01**
1	0,000015	0,000098	0,000393	3,84	5,02	6,63
2	0,0201	0,0506	0,103	5,99	7,88	9,21
3	0,115	0,216	0,352	7,81	9,35	11,34
4	0,297	0,484	0,711	9,49	11,14	13,28
5	0,554	0,831	1,15	11,07	12,83	15,09
6	0,872	1,24	1,64	12,59	14,45	16,81
7	1,24	1,69	2,17	14,07	16,01	18,47
8	1,65	2,18	2,73	15,51	17,53	20,09
9	2,09	2,70	3,33	16,92	19,02	21,67
10	2,56	3,25	3,94	18,31	20,48	23,21
11	3,05	3,61	4,57	19,68	21,92	24,72
12	3,57	4,40	5,23	21,03	23,34	26,22
13	4,11	5,01	5,89	22,36	24,74	27,69
14	4,66	5,63	6,57	23,68	26,12	29,14
15	5,23	6,26	7,26	25,00	27,49	30,58
16	5,81	6,91	7,96	26,33	28,85	32,00
17	6,41	7,56	8,67	27,59	30,19	33,41
18	7,01	8,23	9,39	28,87	31,53	34,81
19	7,63	8,91	10,12	30,14	32,85	36,19
20	8,26	9,59	10,85	31,41	34,17	37,57
21	8,90	10,28	11,59	32,67	35,48	38,93
22	9,54	10,98	12,34	33,92	36,78	40,29
23	10,20	11,69	13,09	35,17	38,08	41,64
24	10,86	12,40	13,85	36,42	39,36	42,98
25	11,52	13,12	14,61	37,65	40,65	44,31
26	12,20	13,84	15,38	38,89	41,92	45,64
27	12,88	14,57	16,15	40,11	43,19	46,96
28	13,56	15,31	16,93	41,34	44,46	48,28
29	14,26	16,05	17,71	42,56	45,72	49,59
30	14,95	16,79	18,49	43,77	46,98	50,89
31	15,66	17,54	19,28	44,99	48,23	52,19
32	16,36	18,29	20,07	46,19	49,48	53,49
33	17,07	19,05	20,87	47,40	50,73	54,78
34	17,79	19,81	21,66	48,60	51,97	56,06
35	18,51	20,57	22,47	49,80	53,20	57,34
40	22,16	24,43	26,51	55,76	59,34	63,69
50	29,71	32,36	34,76	67,51	71,42	76,15
60	37,48	40,48	43,19	79,08	83,30	88,38
100	70,06	74,22	77,93	124,3	129,5	135,8