A biostatisztika alapjai

FOK Biofizika 2021/2022 II. félév

Liliom Károly

biostatisztika előadások tematikája

1. leíró statisztika (adatok jellemzése)
2. hipotézisvizsgálatok (adatok összehasonlítása)
3. korreláció és regresszióanalízis

Javasolt tankönyv:

- **Orvosi Biofizikai Gyakorlatok, Statisztika pótfejezet** – Medicina Kiadó, 2017

Javasolt olvasmányok:

- Herényi Levente: Statisztika és Informatika, Medicina Kiadó 2016

- Harvey Motulsky: Intuitive Biostatistics – A Nonmathematical Guide to Statistical Thinking, Oxford University Press
- Nature Collection: Statistics for Biologists

BEVEZETÉS

Ismerjük annak az embernek a történetét, aki hétfőn whiskyt ivott szódával, kedden gint szódával, szerdán pedig rumot szódával. Mivel az eredmény mindig azonos volt, arra a következtetésre jutott, hogy a szódától rúgott be.

Ha nem is a whisky és szóda viszonylatában, de nagyon hasonlóan viselkedünk mindannyian. Sokunkra jellemző, hogy **könnyen vonunk le megalapozatlan következtetéseket**, és így hozunk döntéseket. Egyedi esetek alapján általánosítunk, a rendelkezésünkre álló ismeretekből is önkényesen válogatunk vélt igazunk alátámasztására. Az így kialakított álláspontunkhoz, véleményünkhöz pedig sokszor megingathatatlanul ragaszkodunk.

Első megközelítésként azt mondhatjuk, hogy **a statisztika** ennek az általános „kórnak" a leküzdésére alkalmas tudomány. **Segítséget nyújt ahhoz, hogy** kritikusabban rendezzük el gondolatainkat, és **kételkedésünket** — ami minden igényes értelmiségi tevékenység alapja — mindig **ébren tartsuk**.

Mivel nap mint nap találkozunk statisztikai kimutatásokkal, úgy tűnhet, hogy ismerjük az előállításukra szolgáló módszereket is. Ez részben igaz, hiszen például, **aki** Magyarországon az utóbbi néhány évtizedben **járt iskolába**, az életében már **néhányszor biztosan használt statisztikai módszereket**. Elég csak azt az esetet említeni, **amikor** egy kisdiák **kiszámolja** az egyes tantárgyakból szerzett **érdemjegyeinek átlagait** abból a célból, hogy megtudja, milyen osztályzatokra számíthat a félévi vagy év végi bizonyítványában. Ilyenkor a tanuló a statisztikai **becslés** egy tipikus esetét alkalmazta anélkül, hogy tudott volna róla.

Egyébként a „statisztika" szó több jelentéssel is bír. Az itt használt értelmét leginkább a latin eredetre visszanyúlva a „status" szóból ismerhetjük meg. A status eredeti jelentése: állapot, helyzet, a dolgok állásának módja. Ennek megismeréséhez, illetve leírásához biztosítanak lehetőséget az **adatok**. Általánosságban a környező világ **minőségi és mennyiségi** jellemzőit nevezhetjük adatoknak. 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.

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. (Ez legfeljebb a titkosrendőröknél szokásos.) Az összegyűjtött, de rendezetlen adatok önmagukban, sok esetben teljesen használhatatlanok. Gondoljunk például arra, hogy egy telefonkönyv adatait a központba való beérkezésük sorrendjében adnák közre, mire mennénk vele? Sokszor kell az adatokat értékelni, például fontosságuk szerint. Ezt teszi az orvos is, amikor felállít egy diagnózist, vagy amikor a beteg állapotáról nyilatkozik.

Az **adatokat** tehát **össze kell gyűjteni**, legtöbbször **fel kell dolgozni** őket, és szükség esetén **következtetéseket kell levonni** belőlük, illetve időnként **döntéseket kell hozni** azok alapján. A **statisztika** az a tudomány, ami minderre megtanít bennünket. A „statisztika" szó előtt gyakran szereplő „**bio-**„ előtag arra utal, hogy bizonyos statisztikai módszereket az **élővilággal kapcsolatos jelenségek** elemzésére használnak. Ehhez hasonlóan az **orvosi statisztika** módszerei a felhasználások még konkrétabb körére, nevezetesen orvosi problémák megoldására lettek kidolgozva.

Nem könnyű első éves orvos-egyetemisták számára meggyőző érveket hozni annak alátámasztására, hogy **a statisztika** számukra **is fontos**, mondhatni **nélkülözhetetlen** tudomány. Ezért most csak néhány példát említünk.

Az orvosi fizikai gyakorlatok nagy részén, de az elméleti modul többi tárgyának részeként is, sőt az **egyetemi évek során** később is a hallgatók egyéni méréseket végeznek. A **mérési adatokból megbízható következtetéseket** pedig csak bizonyos statisztikai ismeretek birtokában lehet **levonni**.

A **kórtörténeti leírásokban**, a laboratóriumi naplókban rengeteg adat szerepel. Nagyon fontos, hogy ezen **adatok helyes értékelését, a helyes következtetések levonásának módszereit és az eredmények megbízhatóságának vizsgálatát** az orvosok, fogorvosok, gyógyszerészek **megismerjék, és** munkájukban **alkalmazzák**. A statisztika óvhat meg bennünket az új eljárások, gyógyszerek reklámáradatának esetleges hazugságaitól is.

Megemlíthetjük **az orvosi irodalom értő olvasásának** nehézségeit is. Példaként álljon itt egy rövid kivonat egy orvosi közleményből. *„Az 1. csoportban az átlagos –3,94 ± 1,3 dioptriás preoperatív fénytörési hiba az 1 éves követése során –0,47 ± 0,54 dioptriára ... csökkent."*; majd később *„A statisztikai eredményeket kétmintás t-próba és regressziós analízis segítségével értékeltük."* A kérdés ugye csak az: mit jelentenek a számok, illetve mik ezek a módszerek?

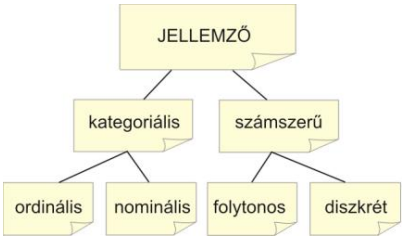
Végül, de nem utolsósorban a **statisztika** egy **sajátos szemléletet**, gondolkodási módot **közvetít**, ami nagyon **hasonlít az orvosi gondolkodáshoz** is, így megismerése ebből a szempontból is kívánatos. Ennek illusztrálására szintén bemutatunk egy példát. Tegyük fel, hogy *páciensünk fejfájásra panaszkodik, a kérdés az, hogy mi lehet ennek az oka.* Orvosi tanulmányaink alapján eszünkbe kell, hogy jusson az „összes" lehetséges ok. Ezek közül néhány:

1. *Magas a vérnyomása.*
2. *Rossz a szemüvege.*
3. *Túl magas a szemnyomása.*
4. *Koponyaűri daganata van.*
5. *A nyaki csigolyái meszesek.*
6. *Érzékeny a frontokra.*
7. *Oxigénszegény környezetben dolgozik.*

Mivel itt általában több helyes válasz is létezhet egyszerre, ezért a válaszokat egyesével vizsgáljuk meg, és ellenőrizzük, hogy a gyanú alapos, vagy alaptalan.

Ez az eljárás alapjaiban megegyezik a statisztikában használatos, e rövid összefoglalóban is bemutatásra kerülő egyik módszerrel, a „hipotézisvizsgálattal”. Első közelítésben a statisztikai tevékenységeket négy csoportba sorolhatjuk, de ezek között nincs éles határ: **adatgyűjtés**, **az adatok áttekinthetővé tétele**, az adatok **elemzése**, és a **következtetések**.

ADATGYŰJTÉS, AZ ADATOK FŐBB TÍPUSAI



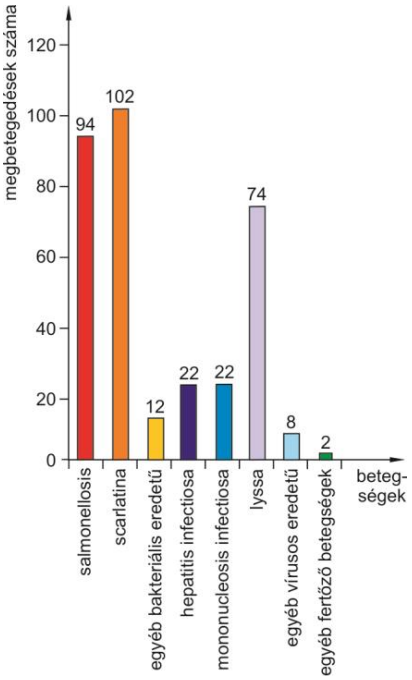
1. táblázat. *A jellemzők osztályozása.*

Mint már említettük, az adatgyűjtés valamilyen cél eléréséhez szükséges. Vannak adatok, amik csak bizonyos dolgok azonosítására, megkülönböztetésére szolgálnak. Nagyobb számban adatokat akkor gyűjtünk, ha azt reméljük, hogy ezek segítségével valamilyen korábban feltett kérdésünkre feleletet kapunk. Az adatgyűjtés módját, **az adatokhoz való hozzájutást** a továbbiakban általánosan **„kísérletnek”**, ill. ha emberen végzik **„vizsgálatnak”** nevezzük. Az adatok egy része ismert, csak meg kell kérdezni valakitől, másik részét meg kell mérni valahogy, de a fenti értelemben kísérletnek tekinthető egy természeti jelenség megfigyelése, vagy egy dobókocka feldobása is. **A kísérlet, vagy vizsgálat eredménye,** az adat lehet **minőségi,** legfeljebb kategóriákba sorolható, azaz **kategoriális**, vagy **mennyiségi,** tehát számmal jellemzett, azaz **számszerű** jellemző. A betegségek neve, ami azonosításuknál többre nem nagyon alkalmas, minőségi (kvalitatív) jellemzést jelent. Ugyanígy a betegség súlyossága, a fertőzést előidéző kórokozó típusa szintén minőségi jellemző, de például az esetleges kiütések nagysága, vagy a betegség lefolyásának ideje már mérőszámmal (és mértékegységgel) is megadható, tehát mennyiségi (kvantitatív) jellemző. A minőségi vagy kategoriális jellemzőket két csoportba sorolhatjuk aszerint, hogy valamilyen természetes sorba rendezhetők-e vagy sem. Sorba rendezhető, azaz **ordinális** jellemző például egy betegség lefolyásának foka: *enyhe, közepes, súlyos*; de nem ilyen, azaz **nominális** például a vércsoport: *A, AB, B, 0*. A számszerű jellemzőkön belül megkülönböztetünk **folytonos**, azaz bizonyos határokon belül tetszőleges értéket felvehető, illetve **diszkrét**, azaz csak bizonyos értékeket felvehető jellemzőt. A testsúly, testmagasság, vérnyomás folytonos, de például egy családon belül a gyermekek száma csak egész szám lehet, tehát az diszkrét jellemző. (A jellemzők főbb típusait az 1. táblázatban foglaltuk össze.) Itt jegyezzük meg, hogy a folytonos jellemző csak elvileg folytonos, a gyakorlatban mindig diszkrét értékekkel dolgozunk (ellenkező esetben végtelen tizedes törtek alkalmazására is szükségünk lenne).

abszolút gyakoriság
absolute frequency
absolute Häufigkeit

relatív gyakoriság
relative frequency
relative Häufigkeit

AZ ADATOK ÁTTEKINTHETŐVÉ TÉTELE, GYAKORISÁGOK ÉS ÁBRÁZOLÁSUK



1. a. ábra. *Oszlop diagram. Abszolút gyakoriságok a betegségek, mint kategóriák függvényében.*

A mindennapi életben is gyakran előfordul, hogy egy probléma kapcsán viszonylag sok adat áll rendelkezésünkre. Ilyen esetekben **szükséges, hogy az adatokról valamilyen áttekintésünk legyen**.

KÓROKOZÓ	BETEGSÉG	abszolút gyakoriság		relatív gyakoriság	
baktérium	salmonellosis (szalmonella fertőzés)	94	208	0,280	0,619
	scarlatina (skarlát)	102		0,304	
	egyéb bakteriális eredetű	12		0,036	
vírus	hepatitis infectiosa (fertőző májgyulladás)	22	126	0,065	0,375
	mononucleosis infectiosa (mirigyláz)	22		0,065	
	lyssa (veszettség)	74		0,220	
	egyéb vírusos eredetű	8		0,0238	
egyéb	egyéb fertőző betegségek	2	2	0,006	0,006
összesen:		336	336	1,000	1,000

2. táblázat*. Egy összesítés a fertőző megbetegedésekről.*

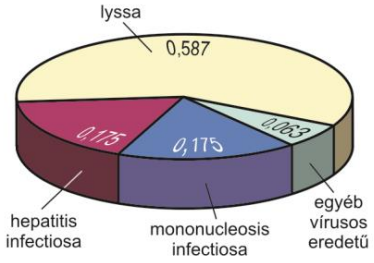
A fenti 2. táblázat a Budapesten, 2000 októberében bejelentett fertőző megbetegedések összesítését mutatja.

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 előforduló összes fertőző betegségek 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. Százzal való szorzással %-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 betegségek számával (208) kell elosztani. Az így kapott hányados (0,452) is relatív gyakoriság, de most nem az összes betegséghez (336), hanem csak a bakteriális eredetű betegségekhez (208) viszonyítottunk.

Az abszolút és relatív gyakoriságok ábrázolására sok lehetőség kínálkozik. Ezek közül mutatunk be kettőt az 1. a. és b. ábrán.



1. b. ábra. *Torta diagram. Relatív gyakoriságok a vírusos eredetű fertőző betegségek (kategóriák) megoszlásáról.*

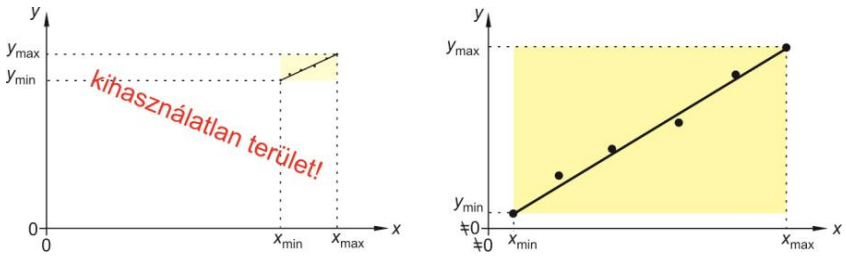
Ellenőrző kérdés:
Nincs-e ellentmondás a következő állításokban: az én csoportomban a lányok relatív gyakorisága kisebb, mint a barátomében, de nálunk mégis több lány van, mint náluk?

AZ ADATOK KÖZÖTTI KAPCSOLAT SZEMLÉLETES BEMUTATÁSA, GRAFIKUS ÁBRÁZOLÁS

Kísérleteink, méréseink során gyakran előfordul, hogy **többféle jellemzőt** is meghatározunk, illetve, hogy **egy jellemzőt egy előre meghatározott paraméter**, mondjuk az idő **függvényében vizsgálunk**. (Ez utóbbira példa a rendszeres lázmérés a kórházakban.) Ilyen esetekben a különböző adatsorok közötti kapcsolatra, az esetleges összefüggésekre vagyunk kíváncsiak. Ezért a jobb áttekinthetőség érdekében célszerű az eredményeket **grafikusan** ábrázolni. Fontos megjegyeznünk, hogy a „**kapcsolat**” szó a legáltalánosabb értelemben használandó, tehát **nem jelent ok-okozati összefüggést**.

Grafikus ábrázolásnál először két fontos dologban kell döntenünk. Az egyik a **tengelyek beosztása** (skálázás), a másik a **kezdőértékek megválasztása**. Nem szükségszerű ugyanis grafikonunkon az origót minden esetben feltüntetni. Ha például koncentráció meghatározás érdekében vizes fehérjeoldatok törésmutatóját mérjük, előre tudjuk, hogy a desztillált víz törésmutatójánál (1,333) kisebb értéket nem fogunk kapni.

A fenti kérdések eldöntésében általános irányadó szempont az lehet, hogy grafikonunk töltse ki a rendelkezésre álló terület nagy részét (lásd 2. ábra).



2. ábra. *A grafikon helytelen és helyes elhelyezése a milliméterpapíron.*

Szükségtelen azonban a skálabeosztásokat olyan mértékben megnövelni, hogy a grafikon kiértékelése nagyobb pontossággal történhessék, mint amilyenre a mérőeszköz maga egyáltalán lehetőséget nyújt. A pontosság ilyen látszólagos növelése megtévesztő lehet.

Méréseinknél bizonyos hibával mindig számolnunk kell. Ezért a görbe „kihúzásánál” ne ragaszkodjunk szigorúan a mérési pontokhoz. A görbe alá és fölé kerülő mérési pontok száma, illetve a görbétől való eltérésük mértéke körülbelül

egyezzen meg. Ily módon ritka kivételektől eltekintve sima lefutású, folytonos görbét illeszthetünk adatainkhoz (lásd lineáris regresszió).

STATISZTIKAI KÖVETKEZTETÉS ÉS A VALÓSZÍNŰSÉGSZÁMÍTÁS

A statisztikai módszerek végső célja a következtetés. A statisztikai következtetések sémája nagyon hasonlít a logikai következtetések sémájához. A logikában a szillogizmus a következtetés egyik fajtája, amelyben bizonyos dolgok megállapításából **szükségszerűen következik** valami más. (A klasszikus példa szerint: *minden ember halandó, Levente ember, tehát Levente halandó*.)

A statisztikai következtetés ettől annyiban tér el (ami persze nem jelentéktelen különbség), hogy míg a **logikai következtetést teljes** (100%-os) **bizonyossággal** állíthatjuk**,** addig a **statisztikait** csak **adott,** (100%-nál mindig kisebb) **bizonyossággal**. Ebből következik tehát, hogy a statisztikai következtetéseknél **tévedhetünk**. Ha például valamit 95% bizonyossággal állítunk, akkor az azt jelenti, hogy átlagosan minden 100 eset közül 5 esetben tévedünk; a bizonytalanság 5%-os. 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! Egy bankrablást követően az eseményekről beszámoló rendőri jelentés így ír: *„… 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”*. **Ha tehát következtetésünk bizonytalanságát csökkentjük, akkor annak általában az az ára, hogy a következtetés értéke, használhatósága is csökken**. Következtetési módszereinket tehát e két ellentétes tendencia szabályozza.

A bizonytalanság oka az, hogy a statisztikai következtetéseknél (a logikaival ellentétben) **nem tudunk minden körülményt számba venni**. A feldobott pénzérmét nem 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**.

A szerencsejátékokkal kapcsolatban már régen megfigyelték, hogy **a véletlen tömegjelenségek** is bizonyos **törvényszerűségeket követnek**. A feldobott pénzérme példájánál maradva, ha sokszor megismételjük a „kísérletet”, akkor az érme körülbelül az esetek felében esik fejre, felében írásra. Bár bizonyítani nem tudjuk, tapasztalatból állíthatjuk, hogy nagy számú (független) kísérlet esetén a fejek és írások **relatív gyakorisága**, azaz a [(fejek száma)/(fejek + írások száma)] illetve az [(írások száma)/(fejek + írások száma)] **stabilitást mutat**, mindkettő egyaránt ½ közelében lesz. (Ez a nagy számok törvénye.) Ennek alapján azt is mondhatjuk, hogy annak a **valószínűsége**, hogy egyetlen pénzfeldobáskor fejet kapjunk, éppen ½. (Az írás valószínűsége természetesen ugyanekkora.)

A **valószínűségszámítás matematikai leírást ad** az anyagi világban tömegméretekben lejátszódó **olyan jelenségek törvényszerűségeire**, melyek lefolyását **a számba vehető körülmények nem határozzák meg egyértelműe**n. kísérleteink, megfigyeléseink, mérési eredményeink éppen ebbe a kategóriába esnek, így a statisztika a valószínűségszámítás eredményeire építhet.

POPULÁCIÓ, VÁLTOZÓ, MINTA

Kísérleteket, méréseket, megfigyeléseket mindannyian végzünk, ezt illusztrálja néhány példával a 3. táblázat, (ahol az utolsó oszlopban található, zárójelben lévő számok a hallgatói mérések jegyzetbeli sorszámát mutatják).

KI MIT MÉR?		
FIZIKUS	ORVOS	ORVOSTANHALLGATÓ AZ ORVOSI FIZIKA GYAKORLATOKON
hosszúság	testmagasság	vörösvérsejt átmérő (3.)
frekvencia	pulzusszám	impulzus gyakoriság (9., 20.)
hőmérséklet	testhőmérséklet	–
koncentráció	vércukor-szint	vérplazma fehérje-koncentráció (5.)
feszültség	EKG-jel	EKG-jel (24.)
teljesítménysűrűség	hallásküszöb	hallásküszöb (22.)
nyomás	vérnyomás	–
impedancia	bőrimpedancia, (bőrellenállás)	bőrimpedancia (21.)

3. táblázat. *Mit mér a fizikus, az orvos és az orvostanhallgató?*

Ismét hangsúlyozzuk, hogy méréseink célja valaminek a megismerése, valamilyen kérdésnek a megválaszolása.

Válasszuk ki táblázatunkból példaképpen a pulzusszám mérést, amely könnyen elvégezhető az orvosi fizika gyakorlaton is. A pulzusszám a szívverés frekvenciája, (elvileg) 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, de nem szabad elfelejtenünk, hogy a végső eredményeket mindig mértékegységgel együtt kell megadnunk. Sokféle kérdést tehetünk fel, amire ettől a méréstől várjuk a választ. Például:

1. *MEKKORA* X.Y. hallgató pulzusszáma?
2. *MEKKORA* a normális pulzusszám?

3. *VÁLTOZIK-E* a pulzusszám egyperces lélegzet-visszatartás után?
4. *VAN-E KÜLÖNBSÉG* a lányok és a fiúk pulzusszáma között? , stb.

Vegyük sorra ezeket. **Statisztikai ismeretektől mentesen** bárki azt gondolhatná, hogy az első kérdésre igazán könnyű válaszolni, nevezetesen megmérem X.Y. pulzusszámát, **kijön egy eredmény és kész**. 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 „kísérlet" lefolyása a „véletlentől" is függ, tehát a legnagyobb igyekezettel sem tudunk „hibátlanul" mérni, akkor arra az elhatározásra jut, hogy **újra mér**.

A többszöri mérés végzésekor hallgatólagosan mindig feltételezzük, hogy **ugyanazt mérjük** még egyszer, ezért várhatóan **ugyanakkora** lesz az **eredmény** is. Másképpen mondva X.Y. pulzusszáma az ismételt mérések során **tartósan egyik irányban sem változik**, ennek ellenére a számba nem vehető, véletlen ingadozások folytán **mégsem kapunk azonos eredményeket**. Azt is mondhatjuk, hogy a kísérlet eredményének két része van, az egyik, — nevezzük fő résznek — determinisztikus (meghatározott), a másik pedig, — nevezzük ezt mellék résznek — sztochasztikus (véletlen). (E két részt természetesen nem tudjuk eleve különválasztani.)

A többszöri mérést úgy is tekinthetjük, hogy létezik egy halmaz, más néven **alapsokaság**, vagy **populáció**, és minden mérés során ennek a halmaznak választjuk ki egy elemét. (Ebben az esetben a halmaz elvileg végtelen számú elemet tartalmaz, de ez nem alapfeltétel.) A halmaz általános elemét **változónak** nevezzük és szokásos módon például x-szel jelöljük. Ez a változó különböző

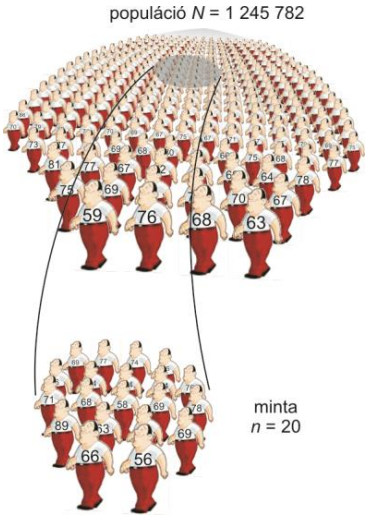
értékeket vehet fel, hogy éppen melyiket, azt mondja meg az adott mérési eredmény.

Egyetlen mérés alapján a többi feltett kérdésre sem tudunk választ adni. A második kérdés esetében úgy gondolhatjuk, hogy létezik egy determinisztikus normális pulzusszám, és az egyének pulzusszáma ekörül ingadozik véletlenszerűen. Itt a populációt például úgy lehet elképzelni (3. ábra), hogy egy adott pillanatban ismerjük sok embernek (mondjuk $N = 1245782$ egyednek) a pulzusszámát, és ezek az emberek a pulzusszámukkal együtt alkotják az alapsokaságot. (Megjegyezzük, hogy ebben az esetben az alapsokaság véges sok elemű.)

Az első esetben a méréseket ugyanazon a személyen ismételjük meg (akárhányszor), a másodikban különböző személyeken mérünk pulzusszámot. A változó tehát mindkét esetben igen hasonló, mégis két különböző populációról, alapsokaságról van szó. A harmadik, illetve negyedik kérdésre később visszatérünk.

Bár a feltett kérdések valójában a populációra vonatkoznak, a legtöbb esetben nem áll módunkban annak teljes megismerése. Emiatt az N elemű sokaságból (N végtelen is lehet) **mintát** veszünk. **A mintavétel a sokaság *n* számú elemének véletlenszerű kiválasztásából áll**. (Ezt hajtjuk végre például a többszöri mérés alkalmával.) Az egésznek természetesen csak akkor van értelme, ha n lényegesen kisebb lehet, mint N (lásd 3. ábra).

minta
sample
Stichprobe



3. ábra. *A populáció, a változó és a minta szemléltetése.*

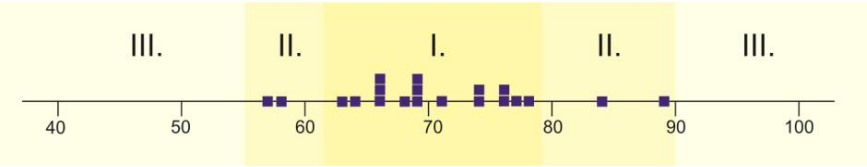
A MINTA ELOSZLÁSA, GYAKORISÁGI ELOSZLÁS, HISZTOGRAM

Ezt a fogalmat egy példa kapcsán vezetjük be. Mivel végül választ szeretnénk adni az előző részben feltett kérdésekre, most válasszuk ki a másodikat, (*MEKKORA* a normális pulzusszám?) és ennek érdekében mérjük meg egy tanulócsoport hallgatóinak pulzusszámát. A tanulócsoport a pulzusszám adatokkal együtt a feltett kérdéssel kapcsolatos alapsokaságból vett mintának tekinthető ($n = 20$) (lásd 3. ábra). Mérési adatainkat — melyeket általánosan x_i-vel jelölünk (most $i = 1,2,3, \ldots 20$) — az alábbi 4. táblázatban tüntettük fel.

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

4. táblázat. *A hallgatói csoport megmért pulzusszám adatai (minta).*

Bár egy ilyen táblázatos forma talán jobban mutat, mint az adatok egyszerű felsorolása, nem nyújt kellő áttekintést az adatok egymáshoz való viszonyáról. Ha adatainkat egy számegyenes mentén ábrázoljuk (4. ábra), jobban megfigyelhető a „normális” érték körüli ingadozás:



4. ábra. *A minta elemei egy számegyenes mentén ábrázolva.*

Itt önkényesen ugyan, de háromféle tartományt különböztethetünk meg: I. sok adat, II. kevés adat, III. nincs adat.

Ha ezt a képet finomítjuk, akkor eljuthatunk a **gyakorisági eloszlás** fogalmához, amit adataink további osztályozása útján nyerhetünk. E feladat elvégzésének érdekében osszuk fel a számegyenest egyenlő részekre (intervallumokra), és számoljuk meg, hogy az így kapott **osztályokban** hány adat található. Ily módon meghatározhatjuk az egyes osztályokban a gyakoriságokat, illetve a relatív gyakoriságokat. (Megjegyezzük, hogy az osztályokat nem kell feltétlenül egyenlőnek választani, de célszerű.)

gyakorisági eloszlás
frequency distribution
Häufigkeitsverteilung

hisztogram
histogram
Histogramm

Természetesen, mivel az osztályhatárok megválasztása önkényes, ugyanazokból az adatokból többféle gyakorisági eloszlást is készíthetünk. Az alábbi 5. táblázat egyet mutat be a lehetséges esetek közül.

OSZTÁLYHATÁROK	GYAKORISÁG	RELATÍV GYAKORISÁG
$55 \le x_i < 60$	2	0,10
$60 \le x_i < 65$	2	0,10
$65 \le x_i < 70$	7	0,35
$70 \le x_i < 75$	3	0,15
$75 \le x_i < 80$	4	0,20
$80 \le x_i < 85$	1	0,05
$85 \le x_i < 90$	1	0,05
összesen:	$n = 20$	1,00

5. táblázat. *Egy, a mintából képzett gyakorisági eloszlás.*

A kapott gyakoriságokat, illetve relatív gyakoriságokat a jobb áttekinthetőség kedvéért oszlop diagrammal szemléltethetjük. Változó 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. (Az azonos osztályszélesség választás épp azért előnyös, mert ilyen esetben a téglalap területe és a magassága arányos egymással.)

Ilyen hisztogramokat láthatunk az 5. ábrán. Az első két esetben az osztályszélességek azonosak, csak az osztályhatárok különböznek, a második két esetben az osztályszélességek is különbözők. A hisztogram készítésére tehát nincs túl szigorú előírás, esztétikai szempontjaink azért lehetnek (lásd 1. megjegyzés).

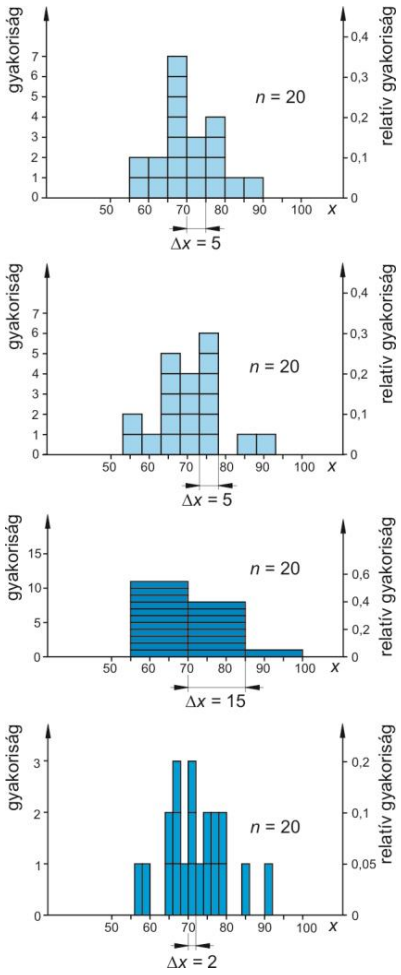
A grafikonok vízszintes tengelyén tehát a változó értéktartományai, a függőlegesen pedig az abszolút, illetve a relatív gyakoriságok vannak feltüntetve. Minden négyzet illetve téglalap egyetlen adatnak felel meg, ezért ezek száma az összes adat számával egyezik meg ($n = 20$). Azt is mondhatjuk, hogy ez éppen az oszlopokat burkoló görbe alatti terület, ami a relatív gyakoriságok szerint (az n elemszámmal való osztás miatt) éppen 1, azaz 100%.

Bár az egyes hisztogramok konkrét alakja jelentős eltérést mutat, mégis találunk közös vonásokat. Megfigyelhetjük például, hogy mindegyik „kipúposodik" a közepe tájékán nagyjából ugyanannál az értéknél, de ezen túlmenően még a „szélességük" is hasonló mértékű. Ha az adatok számát növelnénk, az osztályok szélességét pedig csökkentenénk (és ezt akármeddig folytathatnánk), akkor hisztogramjaink durva lépcsős burkoló görbéi egyre jobban kisimulnának és egyetlen folytonos, sima görbébe mennének át (lásd 6. ábra).

A SOKASÁG ELOSZLÁSA, ELMÉLETI ELOSZLÁS

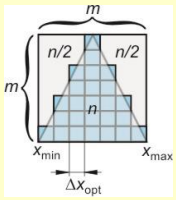
Vizsgáljuk meg közelebbről a 6. ábrán bemutatott tendenciát. Véges elemű sokaság esetén, amennyiben a minta elemszámát (n-t), növeljük akkor előbb-utóbb a sokaság „összes" eleme kiválasztásra kerül ($n = N$). Ilyenkor **a N elemű „minta" eloszlása** az osztályhatárok bizonytalanságától eltekintve **„ugyanolyan"** lesz, **mint a sokaság eloszlása**. Végtelen elemű, vagy annak tekinthető sokaságnál csak azt mondhatjuk, hogy a minta elemszámának növelése a minta eloszlását a sokaság eloszlásához egyre közelebb viszi. Ilyenkor a populáció eloszlása egy **elméleti eloszlással** írható le.

A **populáció eloszlása a változó minden jellemzőjét meghatározza**, azaz megadja, hogy a változó a lehetséges értékeit mekkora valószínűséggel veszi fel. (Ennél többet nem is mondhatunk a változóról.) Ha például kijelölünk valahol a számegyenesen egy (a, b) intervallumot, akkor az intervallumra eső görbe alatti terület azzal a valószínűséggel egyenlő, hogy egy véletlenül kiválasztott érték éppen az adott (a, b) intervallumba essen. Ha az (a, b) intervallumot olyan helyen választjuk, ahol a görbe kis értékeket vesz fel, akkor ott a görbe alatti terület kicsinysége miatt a változó lehetséges értékeinek előfordulási valószínűsége is kicsi lesz (7/1. ábra). Ha azonban az (a, b) intervallum a görbe nagyobb értékeinél van kijelölve, akkor ott a terület, és emiatt a valószínűség is nagyobb lesz



5. ábra. *A minta alapján készített különböző osztályszélességű hisztogramok. (A legfelső hisztogram készült az 5. táblázat alapján.) Minden téglalap egyetlen adatnak felel meg.*

1. megjegyzés:
A hisztogram akkor „esztétikus", ha nem hézagos, de azért van szerkezete, azaz nincs minden adat egy-két osztályba bezsúfolva.
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.



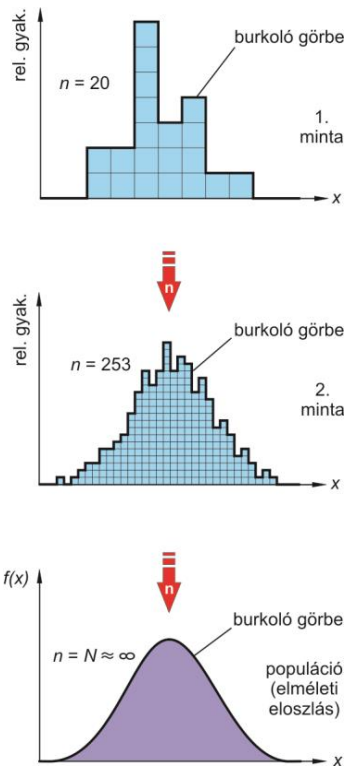
Ekkor egy elem egy négyzet alakú terü letet foglal el (a hosszúkás téglalap helyett). Az ábra alapján az optimális intervallumok száma:

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

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

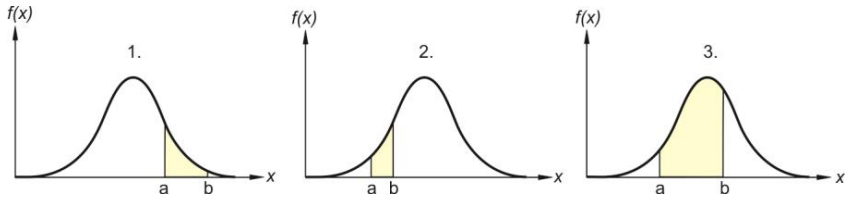
$$\Delta x_{opt} = \frac{x_{max} - x_{min}}{m}\ .$$

Az 5. ábrán a felső két hisztogram ilyen szempontok szerint készült.



6. ábra. *A minta elemszámának növelése és az osztályszélesség csökkentése kisimítja a hisztogramot burkoló görbét.*

(7/2. ábra). Ha egy ilyen helyen az intervallum nagyságát növeljük, akkor nyilvánvalóan a görbe alatti terület még nagyobb lesz, ami azt jelenti, hogy nagyobb intervallumba nagyobb valószínűséggel fog beleesni a véletlenül kiválasztott érték (7/3. ábra). Ebből az is következik, hogy ugyanúgy, mint a hisztogramoknál, **a teljes görbe alatti területnek mindig 1-t kell adnia**, hiszen, ha az (a, b) "intervallum" olyan nagy, hogy –∞-től +∞-ig terjed, akkor abban bármilyen tetszőlegesen kiválasztott érték biztosan benne van. (Lásd korábban: a következtetés bizonytalansága és értéke (használhatósága) közötti kapcsolatot.)



7. ábra. *Az elméleti eloszlás görbe alatti területének szemléletes jelentése.*

Egy fontos dolgot vegyünk azonban észre: nem véletlen, hogy az előbbiek során mindig intervallumról beszéltünk, ugyanis egyetlen kiragadott érték fölött nincs terület (olyan "területről" van szó, amelynek nulla a szélessége). Így folytonos változók esetén nulla annak a valószínűsége is, hogy egy véletlenül kiválasztott érték pontosan egy előre megadott értékkel legyen egyenlő. Ez az oka annak, hogy ilyenkor elvileg az összes mérési eredményünk különbözik egymástól Az, hogy ez a gyakorlatban mégsincs így, azzal magyarázható, hogy valójában minden mérési adat már a leolvasáskor egy intervallumot takar, ami annak az egyszerű ténynek a következménye, hogy a gyakorlatban használt számaink mindig véges tizedes törtek. Így a leírt utolsó számjegy mindig kerekített érték. (Lásd korábban: folytonos és diszkrét jellemzők.)

Az elméleti eloszlás tehát az összes lehetséges adat, vagyis a populáció jellemzésére szolgál, a hisztogram pedig csak egy ebből vett minta elemeire, a konkrét mérési adatokra vonatkozik.

A STATISZTIKA ALAPTÉTELE

Most gondoljuk meg még egyszer, hogy hogyan nyertük az elméleti eloszlást: úgy, hogy a minta elemszámát, azaz méréseink, adataink számát növeltük. A matematikai statisztika alaptételének szemléletes tartalma éppen az, hogy **nagy minták esetén a tapasztalati eloszlásfüggvény** (azaz a hisztogram burkolója) **nagyon jól megközelíti az elméleti eloszlásfüggvényt**. Azt reméljük tehát, hogy **minél gyakoribb egy értékcsoport előfordulása a mintában, annál valószínűbb a megjelenése az alapsokaságban is**.

A matematikai statisztika segítségével egy populáció vagy alapsokaság valamely jellemzőjét úgy határozzuk meg, hogy a sokaságnak csak bizonyos számú (lehetőleg kevés) elemét vizsgáljuk meg. A mintavétel feladata a megvizsgálásra szánt elemek (a minta) kijelölése úgy, hogy belőlük az egész sokaságra megbízható következtetéseket vonhassunk le. Ezt általában úgy érhetjük el, hogy **a mintaelemeket véletlenszerűen választjuk ki** (lásd 2. megjegyzés). (A mintavételnél sok esetben orvosi szempontokra is tekintettel kell lennünk.)

A NORMÁLIS VAGY GAUSS-ELOSZLÁS

Az elméleti eloszlás a vizsgált változótól függően **különböző alakú lehet**, de **az esetek többségében egyetlen csúccsal rendelkező, szimmetrikus, harang alakú görbe** — aminek okára még visszatérünk —, és amit normális vagy Gauss-eloszlásnak nevezünk. (Ilyen eloszlást tüntettünk fel már a 6. és 7. ábrán is.) Az eloszlást leíró függvény:

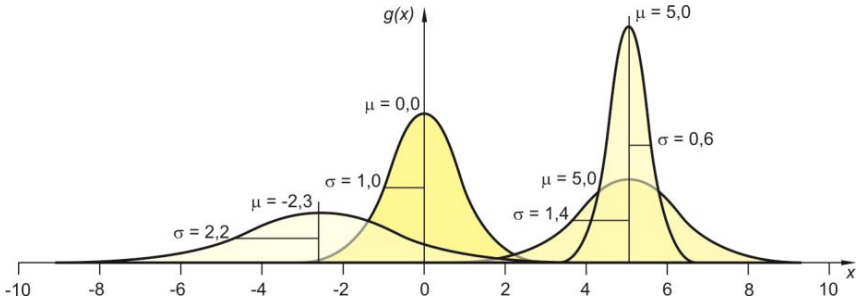
$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} . \qquad (1)$$

Ez a kifejezés elég bonyolultnak tűnik, de valójában nem más, mint az $f(x) = e^{-x^2}$

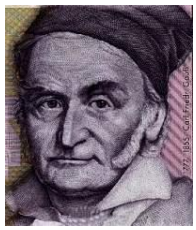
normális eloszlás, Gauss-eloszlás
normal distribution,
Gaussian distribution
Normalverteilung, Gauss Verteilung

2. megjegyzés:
A mintának az alapsokaságra nézve reprezentatívnak kell lennie, vagyis alapfeltétel, hogy a vizsgált jellemző eloszlása a mintában véletlen hibától eltekintve ugyanaz legyen, mint az egész sokaságban. Méréseink, kísérleteink tervezésekor erre is gondolnunk kell.
Ha például egy felmérés során a pajzsmirigy betegségek gyakoriságát akarjuk megállapítani hazánkban, akkor a népesség területi eloszlását figyelembe véve az egész ország területéről kell adatokat gyűjtenünk. Egy régió túlreprezentált jelenléte a mintában hamis eredményre vezethet, mivel például északi megyéink területén az ivóvíz jód-szegénysége miatt az ilyen megbetegedések előfordulása sokkal gyakoribb, mint délen.

függvény állandókkal megtűzdelt változata. Normális vagy Gauss-eloszláson tehát nem egyetlen eloszlást értünk, hanem épp a konstansok miatt egy egész eloszláshalmazt (lásd 8. ábra): az egyes eloszlások tengely menti elhelyezkedése, szélessége és magassága más és más, csak az alakjuk hasonló.

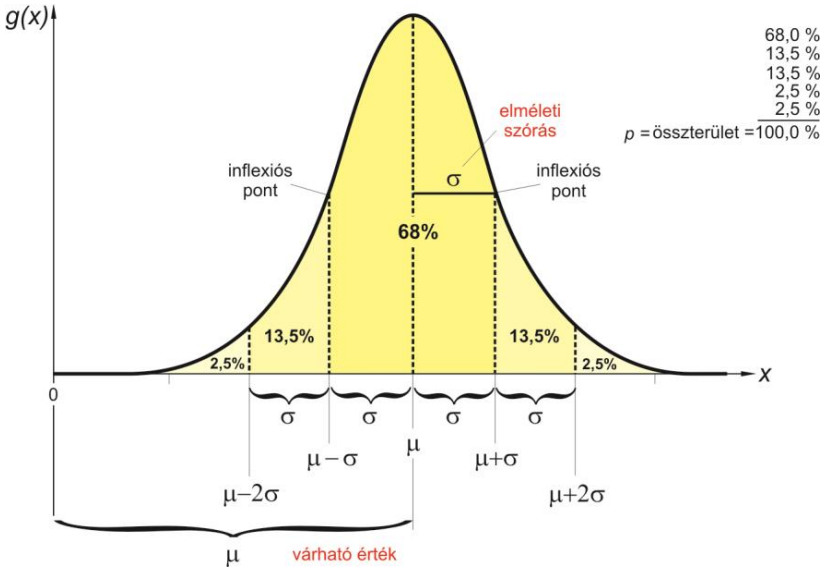


8. ábra. *Néhány eltérő helyzetű (μ), szélességű (σ), és magasságú Gauss-eloszlás.*



Carl Friedrich Gauss (1777-1855) német matematikus.

A normális eloszlásra jellemző haranggörbe (vagy Gauss-görbe) szárai mindkét irányban a végtelenbe nyúlnak, az *x*-tengelyhez azonban annyira közel, hogy külön nem is rajzolhatók. A görbe alatti terület szükségképpen 1. (lásd korábban: elméleti eloszlás.)



9. ábra. *A Gauss-eloszlás haranggörbéje és paraméterei.*

A fenti képletben szereplő **μ és σ állandók az eloszlás paraméterei**. Ezek mutatják meg, hogy a végtelen sok lehetséges normális eloszlásból éppen melyikről van szó. A **μ paraméter a várható érték**, ami megadja az eloszlás maximumának helyét az *x*-tengelyen. A **σ paraméter az elméleti szórás**, ami az eloszlás szélességét jellemzi: a haranggörbe szélessége a magassága felénél mérve körülbelül 2σ-val egyenlő (pontosabban a görbe úgynevezett inflexiós pontjai éppen σ távolságra esnek a μ értéktől) (9. ábra). Mindezek alapján a normális eloszlás szokásos jelölése $N(\mu, \sigma)$.

várható érték
expected value
Erwartungswert

elméleti szórás
theoretical standard deviation, SD
theoretische Streuung

A haranggörbe alakja és a paraméterek közötti kapcsolatról ennél többet is tudunk mondani: a $\mu - \sigma$ és $\mu + \sigma$ értékek között van a görbe alatti terület mintegy kétharmada (kb. 68%-a), $\mu - 2\sigma$ és $\mu + 2\sigma$ között kb. 95 %-a, a $\mu - 3\sigma$, $\mu + 3\sigma$ intervallumon kívülre pedig mindössze csak 2 ezrelék jut. A görbe gyakorlatilag egy 6σ hosszúságú szakaszon helyezkedik el a várható érték körül. Csak megjegyezzük, hogy a **görbék magassága** valójában lényegtelen paraméter, **σ-val fordítottan arányos**, ami a rögzített (egységnyi) görbe alatti terület következménye.

A könnyebb kezelhetőség kedvéért ki szokás emelni egy speciális eloszlást a végtelen sok normális eloszlás közül: ez a $\mu = 0$, $\sigma = 1$ paraméterekkel jellemzett

standard normális eloszlás, a bevezetett jelölés szerint $N(0, 1)$. (a 8. ábrán balról a második görbe).

A normális eloszlás kiemelkedő jelentőségére a valószínűségszámítás egyik nevezetes tétele, a **centrális határeloszlás tétel** mutat rá. Eszerint a sok apró, **egymástól független hatás eredményeképpen kialakult értékek normális eloszlást követnek.** Ezzel magyarázható, hogy a természetben előforduló változók jelentős része normális eloszlású, ezért a továbbiakban főként csak ezzel az eloszlástípussal foglalkozunk.

„Orvosi" példaként megemlíthetjük, hogy Gauss eloszlású a **testmagasság**, vagy a **vérnyomás** is. Magyarországon a felnőtt férfiak testmagasságának eloszlása napjainkban (cm-ben kifejezve) nagyjából az $N(171, 7)$ eloszlásnak felel meg. Iskoláskorú fiúk diasztolés (alsó érték) vérnyomása Hgmm-ben az $N(58, 8)$, dohányzó fiatal férfiaké pedig az $N(84, 10)$ eloszlással adható meg.

A testmagasság példájánál maradva, ahol $3\sigma = 21$ (cm) azt mondhatjuk, hogy a felnőtt férfiak nagy többségének (több mint 99%-ának) testmagassága 150 és 192 cm között van. Vannak persze 2 m magas férfiak is, de ez nagyon nem tipikus érték. A nagyon tipikus a 170 cm körüli magasság, de azt is megfigyelhetjük, hogy viszonylag gyakori a 160, illetve 180 cm-es férfi is. Ez is azt mutatja, ami az **élővilág** egyik fontos jellegzetessége, hogy **vannak ugyan tipikus értékek**, **de a sokféleség**, az egyedek közötti **különbözőség legalább olyan fontos jellemző**.

Ha a másik példát tekintjük, első ránézésre látható, hogy a dohányzó fiatal férfiak vérnyomása nemcsak magasabb, de nagyobb a szórása is ($10 > 8$). Ha azonban kiszámítjuk a **relatív** (elméleti) **szórásokat**, azaz a σ/μ hányadosokat, akkor ebben a paraméterben már fordított a helyzet ($10/84 \approx 0{,}12 < 8/58 \approx 0{,}14$). Sok esetben a relatív szórás, amit százalékban is megadhatunk ($(\sigma/\mu)\cdot(100\%)$) többet árul el az abszolút szórásnál. A szóródás tehát lehet kicsi vagy nagy, de talán még fontosabb, hogy mihez képest, így μ és σ meghatározása egyaránt fontos célunk. Könnyen belátható azonban, hogy a paraméterek „pontos" meghatározása igen fáradságos, esetenként — például végtelen elemű alapsokaság esetén — lehetetlen feladat. Így megelégszünk azzal, ha meg tudjuk becsülni őket.

3. megjegyzés:
A μ várható érték becslésére kínálkozó további lehetőségek:
1. A **modus** az az érték, amelyikből a „legtöbb" van a mintában, vagyis a gyakorisági eloszlásból készített hisztogram maximumának megfelelő érték. Mivel — mint láthattuk — a gyakorisági eloszlás nem egyértelmű (függ az osztályok megválasztásától), ezért a modus sem az. Különösen kevés adat esetén nem igazán jó jellemző. (A 4. táblázatban megadott minta modusa az 5. legfelső ábra alapján: 65 és 70 között van.)
2. A **medián** a nagyság szerint sorba állított adatok közül a középső, vagy a középső kettő átlaga. Vegyük észre, hogy ennek az értékét a szélső adatok egyáltalán nem befolyásolják. Éppen ezért amikor a szélsőséges adatok például méréstechnikai okokból nagyon megbízhatatlanok, akkor ez a legjobb becslése a várható értéknek. (A 4. táblázatban megadott minta mediánja a 4. ábráról leolvasható: 69.)
(Gauss-eloszlás esetén, nagy elemszámú mintákra igaz, hogy az
átlag ≈ modus ≈ medián.
Vannak azonban olyan, például nem szimmetrikus eloszlások is, amelyekre ez nem teljesül.)

átlag
mean, average
Durchschnitt

tapasztalati szórás
empirical standard deviation, SD
empirische Streuung

variancia
variance
Varianz

szabadságfok
degree of freedom
Freiheitsgrad

A PARAMÉTEREK BECSLÉSE, A MINTA STATISZTIKAI JELLEMZŐI

Láttuk, hogy a Gauss-görbét két paramétere (μ *és* σ) egyértelműen jellemzi, ezért célunk az, hogy e paramétereket egy minta alapján történő becsléssel minél jobban megközelítsük.

A μ **várható értéket** leggyakrabban **az átlaggal** ($\bar{x}$) **közelítjük**, amely az adatokból (a minta elemeiből) képzett **számtani közép**:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (2)$$

Ez **a legstabilabb középérték** (normális eloszlás esetén, lásd 3. megjegyzés), amely az összes mintaelem számszerű értékére támaszkodik, és így a minta változására legkevésbé érzékeny. A többi középérték közül kiemelkedő jelentőségűvé még az teszi, hogy ez az a szám, amelytől az adatok eltérésének összege éppen 0-val egyenlő, ugyanis a pozitív és negatív irányú eltérések éppen kiejtik egymást:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = 0. \qquad (3)$$

Ennek alapján azt is mondhatjuk, hogy az átlag a számegyenesen, az adatok között a „súlypontnak" megfelelő helyen van.

A σ **elméleti szórást** leggyakrabban a **tapasztalati szórással** (s) **közelítjük**, amely az átlagtól való átlagos eltérésekkel kapcsolatos definíció:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \qquad (4)$$

(Használatos még az s_x jelölés is, ahol a félreértések elkerülése végett a változót indexként feltüntetjük.) Ennek négyzete, a tapasztalati **szórásnégyzet**, más néven **variancia**:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \qquad (5)$$

Mivel a (4) illetve (5) képlet számlálójában szereplő négyzetes (idegen szóval kvadratikus = „Quadratikus") kifejezés, vagy hozzá nagyon hasonló, még később is fog szerepelni, ezért külön jelölést (Q) vezetünk be rá, továbbá kis átalakítással a kiszámítására alkalmasabb alakra hozzuk:

$$Q_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}. \qquad (6)$$

A (5) képlet nevezője (az $(n-1)$ kifejezés) a **szabadságfok**. Ez a paraméterek becslésével kapcsolatos matematikai statisztikai fogalom szoros összefüggésben van az adatok számával, de láthatóan nem mindig egyenlő vele. A számítás kezdetekor n adatból álló minta szabadságfokainak száma n. Ha egy mintából valamely paramétert úgy kell becsülnünk, hogy ahhoz ugyanebből a mintából már előzetesen meghatározott paramétereket fel kell használnunk, akkor annyit kell levonnunk az eredetileg n szabadságfokból, ahány korábbi paramétert a becslés közben felhasználunk. Mivel a tapasztalati szórás becslésénél az n adaton kívül az ugyanabból a mintából már meghatározott átlagot is fel kell használnunk, ezért a tapasztalati szórás szabadságfokainak száma $n-1$. Bonyolultabb esetekben a szabadságfokot külön képlet segítségével kell kiszámítanunk. (A 4. megjegyzésben a 4. táblázatban megadott minta néhány fontos statisztikai jellemzőjét foglaltuk össze.)

A paramétereknek a mintából kiszámított, becsült értéke és a „valódi" értéke között általában több–kevesebb eltérés mutatkozik. Ez az eltérés a becsült paraméter hibája, amire még visszatérünk. Ezt a hibát **pontatlanság** és **torzítás** okozhatja. **Pontatlanságnak** tekintjük a hibát, ha **az eltérés** a valódi értéktől **pozitív és negatív irányban véletlenszerű**. **Torzításnak** tekintjük a hibát, ha **a paraméter becsült értéke szisztematikusan kisebb vagy nagyobb** a „valódi" értéknél. Míg a pontatlanság számszerűen mérhető, addig **a torzítás többnyire nem mérhető** becslési hiba.

Az átlag és a tapasztalati szórás fenti definícióira igaz a következő állítás (amennyiben a torzításoktól eltekintünk): Ha a minta elemszáma végtelenhez tart ($\rightarrow$), akkor az átlag a várható értéket, a tapasztalati szórás pedig az elméleti szórást közelíti egyre nagyobb pontossággal, azaz

$$n \rightarrow \infty \text{ esetén } \bar{x} \rightarrow \mu \text{ és } s \rightarrow \sigma. \qquad (7)$$

Az s **tapasztalati szórás azt mutatja meg**, hogy **az adatok** átlagosan **mennyire térnek el az átlagtól**, azaz — a Gauss-eloszlásnál (9. ábra) ismertetettekhez hasonlóan — a **minta elemeinek** kb. 68%-a található az $\bar{x} \pm s$ intervallumban, kb. 95%-a az $\bar{x} \pm 2s$ és több mint 99%-a az $\bar{x} \pm 3s$ intervallumban.

A nagy elemszámú ($n \approx 1000$) mintából számolt $\bar{x} \pm k\,s$ tartományt, amelyben a minta elemeinek pontosan 95%-a található ($k \approx 2$), **referencia tartománynak** vagy **normál tartománynak** is szokás nevezni, amit elsősorban a laboratóriumi diagnosztikában alkalmazunk. (Egyes orvosi alkalmazások során a normál tartomány ettől eltérő értelmezése is előfordul). Ezt a gyakorlatban úgy lehet felhasználni, hogy amennyiben egy laboratóriumi adat a normál tartományon belüli, akkor arról 95%-os bizonyossággal azt mondhatjuk, hogy nem utal kóros elváltozásra (lásd 5. megjegyzés).

Ilyen esetben a torzítás nem 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

4. megjegyzés:
Még egyszer az adatok, $n = 20$:

	x_i
x_1	66
x_2	56
x_3	89
x_4	63
x_5	66
x_6	69
x_7	71
x_8	68
x_9	58
x_{10}	69
x_{11}	78
x_{12}	66
x_{13}	64
x_{14}	84
x_{15}	74
x_{16}	76
x_{17}	69
x_{18}	77
x_{19}	74
x_{20}	76
$\sum x_i = 1413$	

Az **átlagos** pulzusszám:
$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1413}{20} \approx 71$ (1/perc)
(egészre kerekítve).
A **kvadratikus összeg**:
$Q_x = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} =$
$= 101075 - \frac{1413^2}{20} = 1246{,}55$
A **variancia**:
$s^2 = \frac{Q_x}{n-1} = \frac{1246{,}55}{19} \approx 66$ (1/perc)^2
(egészre kerekítve).
A **szórás**:
$s = \sqrt{\frac{Q_x}{n-1}} = \sqrt{\frac{1246{,}55}{19}} \approx 8$ (1/perc)
(egészre kerekítve).
A **szabadságfok**: 19.

5. megjegyzés:

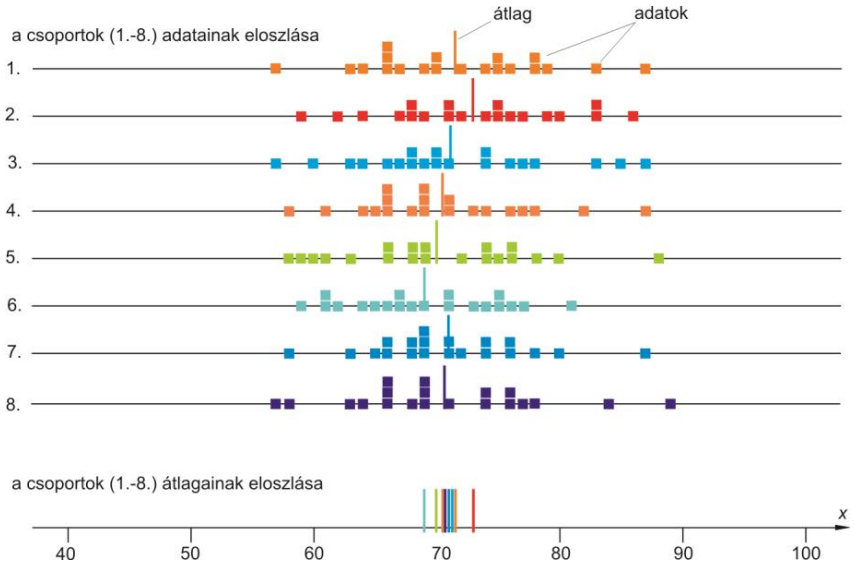
Egy ilyen állítás értékére talán jobban rávilágíthatunk a következő példával. Annak a valószínűsége, hogy egy dobókockával 6-ost dobjunk 1/6, kb.17%, tehát annak a valószínűsége, hogy ne dobjunk 6-ost 5/6, ami (5/6 ≈ 0,83) kb. 83%-nak felel meg.

Ha szabályos dobókocka helyett egy ikozaédert (20 lapú szabályos testet) használnák, aminek minden lapja meg van számozva, akkor annak a valószínűsége, hogy 20-ast dobjunk 1/20 = 0,05 = 5%. A nem 20-as dobás valószínűsége pedig 19/20 = 0,95 = 95%.
Eszerint az az állítás, hogy „*egy laboratóriumi adat a normál tartományon belüli*" nagyjából azzal ekvivalens bizonyosságú, hogy az említett ikozaéderrel elvégzett dobás eredménye nem 20.

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.

KONFIDENCIA INTERVALLUMOK, A BECSÜLT PARAMÉTER VÉLETLEN HIBÁJA, PONTATLANSÁGA

Elöljáróban még egyszer hangsúlyozzuk, hogy a becsült paraméter hibája lehet torzítás is, ami általában nem mérhető, így az alábbiakban a hiba csak a pontatlanságot jelenti, azaz a **véletlen hibát**. Az átlag az előzőek szerint ugyan n növelésével egyre jobban közelíti a „meghatározni" kívánt értéket (lásd (7) kifejezés), de arra a kérdésre, hogy egy adott n elemszámú minta esetén mennyire tér el a populációra jellemző várható értéktől, azaz, hogy **mekkora az átlag hibája**, még nem kaptunk választ.



10. ábra. *Adatok és átlagaik: nyolc 20 fős tanulócsoport hallgatóinak pulzusszámát és azok átlagait tüntettük fel. Figyeljük meg, hogy az átlagok jóval kevésbé szóródnak, mint az adatok.*

Az előző részben azt is említettük, 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 az eredménye, hogy a különböző mintákból számolt átlagok csak kevéssé térnek el egymástól (lásd 10. ábra). Azt is mondhatjuk, hogy a minták átlagai ($\bar{x}_1, \bar{x}_2, \ldots \bar{x}_i$) jóval kevésbé „szóródnak" a várható érték körül, mint az adatok. Ezt a „szóródást" fejezi ki a **standard hiba** (más néven az átlag szórása):

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}. \qquad (8)$$

Ennek segítségével a végső eredményt a következőképpen szokás megadni:

$$\bar{x} \pm k \cdot s_{\bar{x}}. \qquad (9)$$

Végeredményként tehát mindig egy intervallumot adunk meg, azaz pozitív és negatív irányban egy-egy határoló értéket, azzal a céllal, hogy **ezek a** keresett **várható értéket közrezárják**. Problémánk most már csak az, hogy hol vonjuk meg a két határt, azaz mekkora legyen k értéke? A határok megvonása ugyanis nem egyértelmű. Ha nagyon tág határokat adunk meg (k nagy), akkor várható ugyan, hogy ezek a valódi értéket közrefogják, és az a következtetésünk, hogy a valódi érték a két határ között fekszik, feltehetően nem lesz téves (lásd korábban: statisztikai következtetés). A nagyon tág határoknak azonban többnyire kevés a gyakorlati haszna. Ha ezzel szemben szűkítjük a határokat, fokozottan növekszik annak a kockázata, hogy a határokat tévesen adjuk meg, azok már nem fogják

6. megjegyzés:
A pulzusszámra vonatkozó mintából (4. táblázat) már kiszámítottuk az átlagot (71 (1/perc)) és a szórást (8 (1/perc)). A standard hiba

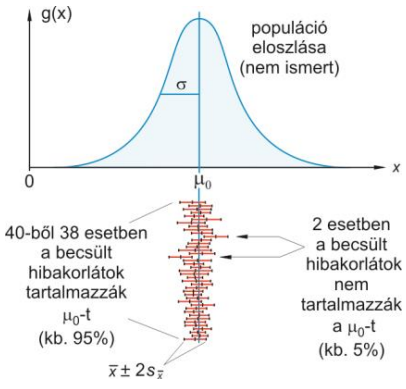
$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{20}} \approx 2 \text{ (1/perc)};$$

a **hibakorlát** pedig (kb. 95% konfidencia szintnél):

$$\bar{x} \pm 2s_{\bar{x}} = 71 \pm 4 \text{ (1/perc)}.$$

(mindenhol egészre kerekítve)
A mérést az alábbiak szerint jellemezhetjük: **„az elvégzett mérések alapján kb. 95 % bizonyossággal állíthatjuk, hogy a populáció várható értéke a 67 - 75 (1/perc) tartományban található"** (11. ábra).
Ha ennél nagyobb bizonyossággal, egyúttal nagyobb pontossággal akarjuk határok közé vonni a pulzusszám várható értékét, akkor növelnünk kell a minta elemszámát.
Meddig növeljük a minta elemszámát?
Általános szabály nincs, de a konkrét esetről a következőket mondhatjuk. Mivel a pulzusszámot egészre kerekítve szokás megadni ezért a pontosság ± 1 (1/perc)-en túli növelésének nincs sok értelme. A bizonyosság kérdése nem ennyire egyértelmű, de ritkán van szükségünk 99%-os, vagy pláne annál nagyobb bizonyosságra. Eszerint (a biztos hibakorlátot választva) addig kell növelnünk n-t, amíg a hiba le nem csökken annyira, hogy a $3s_{\bar{x}} \leq 1$ (1/perc) feltétel teljesüljön.

standard hiba
standard error
Standardfehler



11. ábra. *A mintából számolt hibakorlát 95% bizonyossággal tartalmazza a populáció várható értékét.*

közre a várható értéket, következtetésünk biztonsága, bizonyossága ezért csökken. Így tehát egyrészről a bizonyosság fokozása, illetve a téves következtetés kockázatának csökkentése tág határokat, másrészről a szakmai értelmezhetőség szűk határokat igényel.

konfidencia intervallum
confidence interval
Konfidenzintervall

konfidencia szint
confidence level
Konfidenzniveau

A statisztikai módszerek lehetővé teszik a bizonyosság mértékének rögzítését. Ez azt jelenti, hogy két olyan határoló értéket állapítunk meg, melyek a valódi értéket meghatározott bizonyossággal (konfidencia) fogják közre. Az így megállapított határoló értékeket **konfidencia határoknak**, a két konfidencia határ által közrefogott szakaszt **konfidencia intervallumnak**, a **bizonyosság mértékét** pedig **konfidencia szintnek** nevezzük.

Bár k értéke a minta elemszámától (szabadságfoktól) is függ, **nagy minták esetére** a következőket mondhatjuk: ha $k = 1$, a konfidencia szint kb. 0,68, ha $k = 2$, akkor kb. 0,95, ha $k = 3$, akkor nagyobb, mint 0,99 (lásd 6. táblázat).

konfidencia szint (kb.)	68%	**95%**	99%
konfidencia intervallum	$\bar{x} \pm s_{\bar{x}}$	$\bar{x} \pm 2s_{\bar{x}}$	$\bar{x} \pm 3s_{\bar{x}}$
		hibakorlát	biztos hibakorlát

6. táblázat. *Konfidencia szintek és intervallumok.*

Ezek ismeretében a „*MEKKORA*?” kérdésre a választ, (illetve a mérés végeredményét) a (9) szerint kell megadnunk (lásd 6. megjegyzés, 11. ábra.).

Az átlag szórásának definíciójából (8) kitűnik, hogy a hiba az adatok számának növekedtével egyre csökken:

$$n \to \infty \quad \text{esetén} \quad s_{\bar{x}} \to 0\,, \tag{10}$$

ami rámutat a sokszori mérés értelmére. Így az adatok számának növelésével elérhető az, hogy rögzített konfidencia szint mellett a konfidencia intervallum tetszőleges mértékben csökkenjen (lásd 6. megjegyzés).

GRAFIKUS ADATFELDOLGOZÁS

Az adatok közötti kapcsolat grafikus bemutatásáról már ejtettünk néhány szót. Itt a grafikonok különböző transzformációit mutatjuk be.

Az egyenes egyszerűségéből adódó könnyebbségek (például az, hogy a mérési pontok egyenes vonalzóval „összeköthetők”) olyan nagymértékűek, hogy még ott is egyenessel szeretnénk dolgozni, ahol a változók közötti kapcsolat nyilvánvalóan nem lineáris. Ilyen esetekben olyan transzformációkat hajtunk végre adatainkon, hogy a transzformált adatok végül egyenest határozzanak meg. Így pl. az

$$y = a \cdot e^{bx} \tag{11}$$

exponenciális függvény **logaritmikus transzformáció** után

$$\lg y = (b \cdot \lg e) \cdot x + (\lg a) \tag{12}$$

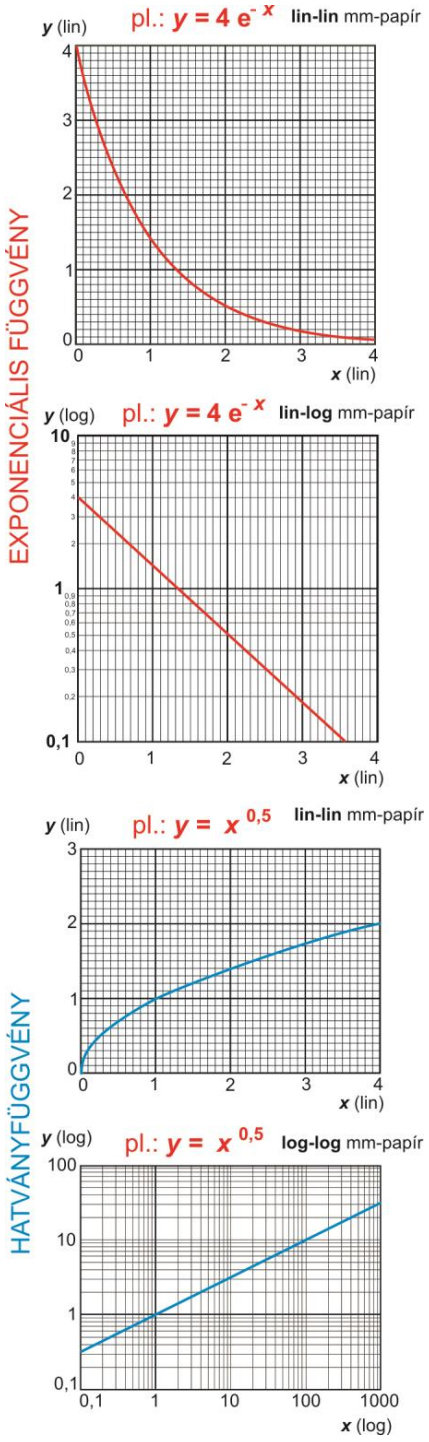
alakú lesz, ahol a **lg *y*** és az ***x*** közötti összefüggés már lineáris. Az egyenes meredeksége (b lg e), tengelymetszete pedig (lg a) lesz. Ha egy speciális, **lin - log** koordinátarendszerű **milliméterpapírt** használunk, az y értékek logaritmálását nem kell elvégeznünk, hiszen a logaritmikus beosztású y tengely ezt „automatikusan” megvalósítja (lásd 12. felső ábrák).Vagy az

$$y = a \cdot x^b \tag{13}$$

hatványfüggvény, szintén logaritmikus transzformáció után,

$$\lg y = (b) \cdot \lg x + (\lg a) \tag{14}$$

alakú lesz, ahol a **lg *y*** és a **lg *x*** között van lineáris összefüggés. Az egyenes meredeksége (b), tengelymetszete pedig (lg a) lesz. A logaritmálás elvégzését itt is elkerülhetjük, egy másik, speciális, **log - log milliméterpapír** használata esetén.



12. ábra. *Az exponenciális, ill. a hatványfüggvény „kiegyenesítése” lin-log, ill. log-log mm-papír segítségével.*