Biostatisztika és informatika alapjai

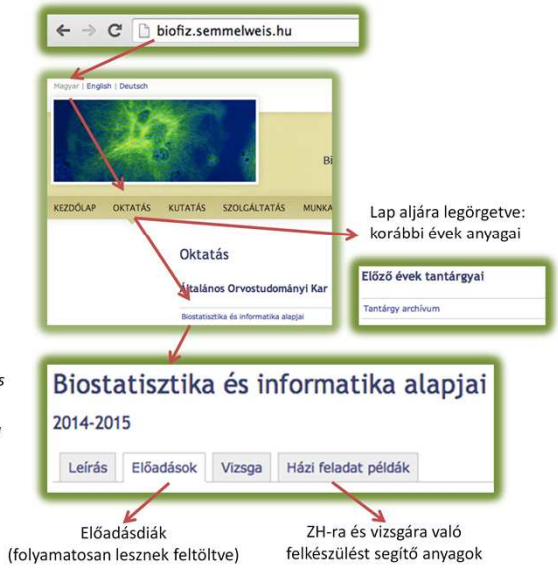
1. előadás: Bevezetés

2014. szeptember 8.

Agócs Gergely

Honnan készüljünk fel?

- egyetem = önálló tanulás
- források:
  - előadáson készített jegyzetek *(hétfő 17^{10}–17^{55}; EOK Szent-Györgyi Albert előadó)*
  - számítógépes laborgyakorlatokon készített jegyzetek *(heti 1 alkalom, 90 perc; EOK első emelet „B" folyosó)*
  - „Orvosi biofizikai gyakorlatok" című gyakorlati könyv:
    - Biostatisztika fejezet *(40-oldalas elméleti összefoglaló)*
    - Feladatok fejezet *(96–104, csak a régebbi kiadásban, az új kiadásban a Biostatisztika fejezet végén)*
  - honlap: biofiz.semmelweis.hu
    - tantárgyi követelmények
    - előadástematika és diák
    - gyakorlati tematika
    - gyakorlófeladatok (házifeladatok)
    - korábbi évek anyagai



Itt foglalom össze a legfontosabb tudnivalókat, részletek a honlapon, illetve a gyakorlatvezetőtől is kaptok információkat.

A statisztika tanulásához a legtöbb infomrációkat az előadásokon és számítógépes laborgyakorlatokon készített saját jegyzeteitekből meríthetitek. Ezt egészítik ki az előadásdiák, amik inkább a szemléltetést segítik, ehhez megkapjátok a leiratot ezekben az annotációkban. 'rott forrásként a biofizikai laborgyakjegyzetben található kb. 40 oldalas összefoglalót lehet használni (Biostatisztika fejezet) és az új jegyzetben ehhez a fejezethez csatolva, a régebbi kiadásokban pedig a biofizika feladatokkal vegyesen találhatók ezek meg. Eza pár feladat azonban kevés a gyakorláshoz, a házi feladatok között találhattok majd további feladatokat.

Az előadások menetrendjét, és a diákat (esetleg kiegészítő anyagokat), gyakorlati menetrendet, házi feladatokat, egyéb információkat a Biofizika Intézet honalpján találjátok meg a fent jelölt módon. Hasznos lehet a tantárgy archívumban a tavalyi előadásokat megkeresni, ha pl. a gyakorlatokon más fejezeteket is érintetek.

Az egyetemünkön végzett orvosképzés alapja a tudományos orvoslás oktatása. Ehhez meg kell értenünk a tudomány szó jelentését. Tudományosnak akkor nevezhetünk egy megállapítást, ha azonos jelenségek megfigyeléséből egymsától függetlenül azonos következtetéseket vonunk le. Tehát a tudományos eredmények ellenőrizhetőek, reprodukálhatóak, nem függenek a megfigyelést végző, illetve a következtetést levonó személyétől – így jelentősen különbözik a puszta véleménytől, hittől vagy a szokásoktól. A tudomány célja a valóság, a tőlünk függetlenül létező valóság megismerése.

Ha egy mondatba akarnánk tömöríteni a tudományos megállapítások lényegét, a jogból vehetünk párhuzamot: egy megállapítás addig nem tekinthető érvényesnek, amíg annak igaz voltát be nem bizonyították.

Az orvoslás nagyon sokáig nem a tudományos szempontokon alapult, részben a gyógyítást gyakorlók ismerethiánya miatt, részben a betegek tájékozatlansága és kiszolgáltatottsága miatt. Habár ma már a tudományos, vagyis a bizonyítékon alapuló orvoslás lett az uralkodó a világban, még mindig jelentős szerephez jutnak a tudományos alapokkal nem rendelekző módszerek, ezek között olyanok is vannak, amelyeknek hatásosságát nagy volumenű kutatások sem tudták kimutatni. Hogy mennyire ellentmondásos, sőt indulatos tud lenni a viszony, arra elég két példa:

1) az egyik a “német akupunktúra vizsgálatok” (GERAC-Studien, az akupunktúra hatásosságával kapcsolatban végzett egyik legkiterjedtebb vizsgálat), amelynek során 4 betegség esetén vizsgálták az akupunktúra hatásosságát, de egyik területen sem tudták kimutatni, ennek ellenére 2 betegség esetén mégis beemelték az akupunktúrát a

társadalombiztosítás által finanszírozott kezelések körébe;

2) a másik Jacques Benveniste-nek a « víz emlékezetével » (a vízemlékezet-teória a homeopátia egyik leggyakrabban használt magyarázata) kapcsolatban a Nature című tudományos lapban megjelent kísérlete, amit azóta se sikerült senkinek se reprodukálnia, még Benveniste laboratóriumában sem. Ennek ellenére a nagyközönség felszines informálódását kihasználva máig gyakran citálják a homeopátia hívei a Benveniste-kísérletet mint a homeopátiát alátámasztó (!) vizsgálatok egyikét.

Emiatt már az orvostudományokkal való legelső találkozásunkkor meg kell értenünk, miért szükséges a tudományos megközelítés. Szemben a hiten vagy hagyományokon alapuló orvoslással, a tudományos orvoslás tudományos bizonyítékokkal alátámasztott módszereket alkalmaz. Ezen módszerek megismeréséhez elengedhetetlen a szakirodalom kritikus befogadására való képesség. Ráadásul a gyógykezelések körének bővítése további bizonyítottan hatásos módszerek kutatását követeli meg. A statisztika a tudományos ismeretek létrehozásának és befogadásának, megértésének eszköze. A statisztika alapvetően matematikai tudomány: a valószínűségszámításon és a logikán alapul.

A statisztika tehát a tudomány létrehozásának és megismerésének, befogadásának eszköze. Azonban a tudományos kérdések nagyon komplexek, így szükség van azok leegyszerűsítésére, elemi kérdésekre bontására. A kérdésfelvetések leginkább a dolgok leírására és összehasonlítására vonatkoznak.

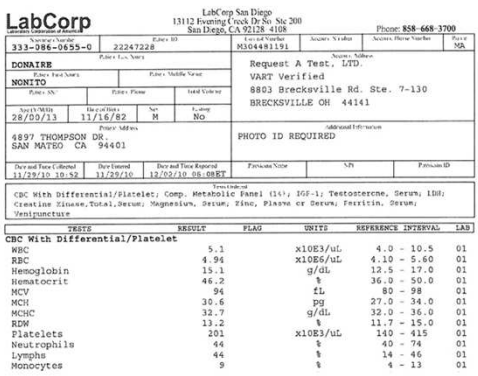
Vegyünk például egy tojást. Egy tojásnak meg tudjuk határozni a tömegét, hosszát, összetételét. De egyetlen tojás megismerése nem segít abban, hogy a tojásokról általános megállapításokat tegyünk, ehhez többet is meg kell vizsgálnunk. Ha megnézzük a következő képet, azon sok tojást látunk, amelyek hasonlítanak egymásra, de szemmel láthatóak pl. a méretbeli eltérések. A következő képen ugyacsak egy halom tojást látunk – az első tojás akár innen is származhatott volna – ezek is hasonlítanak valamennyire, de a színük eltér.

Tehát látható, hogy egyetlen dolog vizsgálatával (egyetlen adattal) általában nem érdemes fogalkozni, ha több dologról gyűjtünk adatot akkor viszont el kell fogadnunk, hogy abban kisebb-nagyobb variáció, ingadozás lesz. A közös vonások és a megtűrt eltérések mértéke lesz az, aminek segítségével lényegében létrehozzuk a “tojás” általános (elvont, elméleti) fogalmát.

A statisztika pontosan sok dolog közös tulajdonságainak a számszerűsítésével, illetve a számszerűsített adatok összehasonlításával foglalkozik. Emiatt a statisztika állításai sem egyetlen dologra, hanem azok egy csoportjára, sokaságára vonatkozik.

Milyen adatokkal foglalkozunk?

A feldolgozandó adatok rendkívül változatosak …



LabCorp

LabCorp San Diego
13112 Evening Creek Dr So Ste 200
San Diego, CA 92128 4108
Phone: 858-668-3700

Account Number	Fisher ID	Local Number	Account Status	Account Phone Number	Payer
333-086-0655-0	22247228	M3044881191			MA

Fisher Last Name: DONAIRE
Fisher First Name: NONITO
Fisher Middle Name:
Fisher SSN:
Fisher Phone:
Total Volume:
Age/DOB: 28/09/13
Physical Date: 11/16/82
Sex: M
Fasting: NO
Fisher Address: 6897 THOMPSON DR.
SAN MATEO CA 94401

Account Name: Request A Test, LTD.
VART Verified
6803 Brecksville Rd. Ste. 7-130
BRECKSVILLE OH 44141

Additional Information: PHOTO ID REQUIRED

Date and Time Collected	Date Entered	Date and Time Reported	Physician Name	NPI	Physician ID
11/29/10 10:52	11/29/10	12/02/10 08:08ET			

Tests Ordered
CBC With Differential/Platelet; Comp. Metabolic Panel (14); IGF-1; Testosterone, Serum; LDH; Creatine Kinase,Total,Serum; Magnesium, Serum; Zinc, Plasma or Serum; Ferritin, Serum; Venipuncture

TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL	LAB
CBC With Differential/Platelet					
WBC	5.1		x10E3/uL	4.0 - 10.5	01
RBC	4.94		x10E6/uL	4.10 - 5.60	01
Hemoglobin	15.1		g/dL	12.5 - 17.0	01
Hematocrit	46.2		%	36.0 - 50.0	01
MCV	94		fL	80 - 98	01
MCH	30.6		pg	27.0 - 34.0	01
MCHC	32.7		g/dL	32.0 - 36.0	01
RDW	13.2		%	11.7 - 15.0	01
Platelets	201		x10E3/uL	140 - 415	01
Neutrophils	44		%	40 - 74	01
Lymphs	44		%	14 - 46	01
Monocytes	9		%	4 - 13	01

Mit mér a fizikus?	Mit mér az orvos?	Mit mér a hallgató?
hossz	testmagasság	vörösvérsejt átmérője (2)
frekvencia	pulzusszám	impulzus-gyakoriság (22)
koncentráció	vércukorszint	vérplazma fehérjekonc. (4)
feszültség	EKG-jel	EKG-jel (27)
hangintenzitás	hallásküszöb	hallásküszöb (25)
impedancia	impedancia-pletizmográfia (térfogatmérés)	bőrimpedancia (24)
nyomás	vérnyomás	
sebesség	véráramlás sebessége	

Orvosként persze nem tojásokról gyűjtünk adatokat, hanem emberekről, de az eljárás hasonló. Tekintsünk át, milyen adatokkal találkozunk pl. egy laborleleten: szerepelnek a beteg azonosítására szolgáló személyi adatok, a TAJ-szám, a kezelőorvos/beküldő orvos neve, a vérben található különféle oldott anyagok koncentrációi, az alakos elemek száma, a pH, a vércsoport stb. Már első pillantásra is látható, hogy igen változatosak az adatok, pl. van amit számmal fejezünk ki, van amit szövegesen.

A jobb oldali táblázatban a biofizika gyakorlatokon előkerülő néhány témát vethetjü össze fizikai és orvosi feladatokkal: pl. a fizikus általában hosszt mér, addig az orvos vagy a hallgató egy specifikusabb hosszt, ami a szakterületükkel kapcsolatos kérdés megválaszolására ad lehetőséget.

Statisztikai alapfogalmak

Példa: kockadobás

- **jelenség**: elvetjük a dobókockát
- **megfigyelés**: figyeljük, hogy hány pont lesz felül, mikor megáll a kocka
- példa **elemi esemény**re: az egyes kerül felülre
- **eseménytér**: az eseménytér a következő **elemi esemény**ekből áll: {1; 2; 3; 4; 5; 6}
- példa (nem elemi) **esemény**re: páros szám kerül felülre
- példa **mintá**ra: ötször dobunk a kockával és az eredmények: {4; 2; 2; 5; 6}
- **alapsokaság**: ebben az esetben végtelen elemű, hiszen végtelen dobás lehetséges
- **statisztikai változó**: a megfigyelések során felülre kerülő számok
- példa **modell**re: a statisztikai változó különféle értékei azonos gyakorisággal fordulnak elő

A statisztikának, mint minden tudománynak megvan a maga nyelvezete, aminek a legfontosabb részeit meg kell tanulni, mivel ezek jelölik a leggyakrabban előforduló fogalmakat.

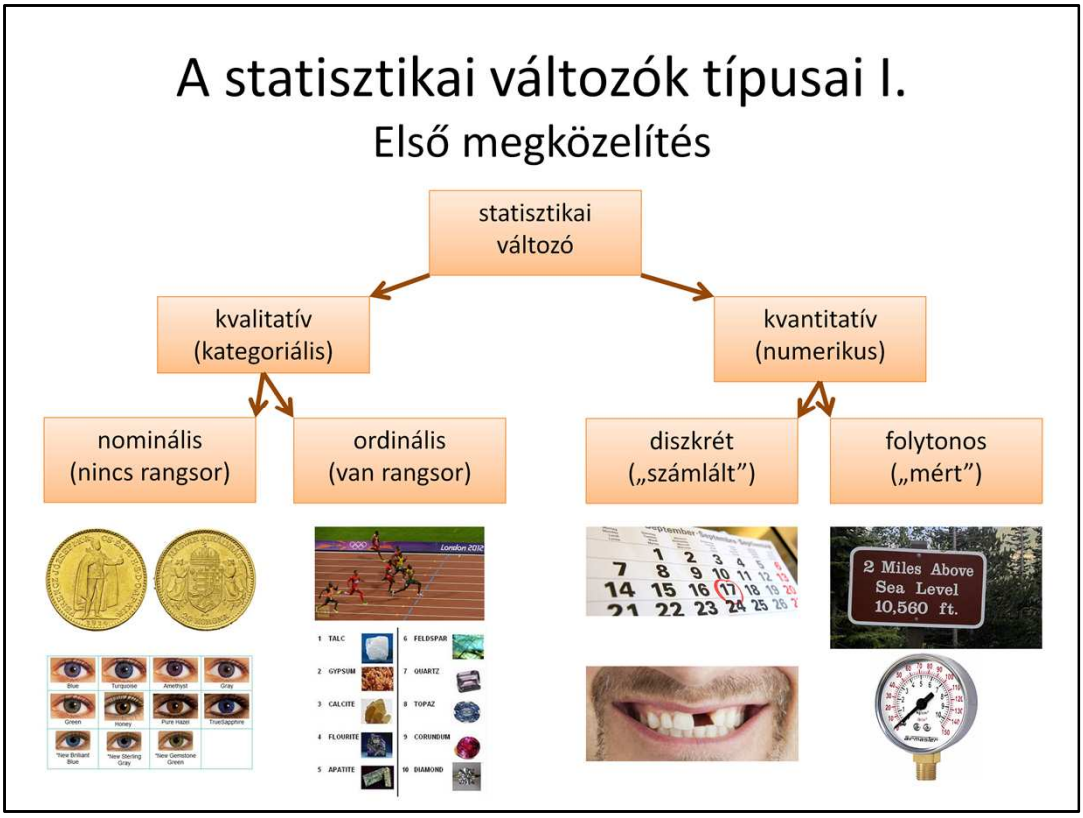
Mit vizsgálunk a statisztikával?
adat (jellemző, ismérv, statisztikai változó): valakinek vagy valaminek a megismeréséhez, jellemzéséhez hozzásegítő tény, lehet minőségi és mennyiségi
jelek: az adatok közvetítői

Ezután a kockadobás példáján szemléltetem, hogy mit jelentenek a gyakorlatban az egyes fogalmak. Alább a (a számunkra megfelelő precizitású) definíciók:

jelenség: minden, ami lényegében azonos feltételek mellett megismétlődhet, amivel kapcsolatban megfigyeléseket lehet végezni, lehet vele „kísérletezni”. (Emlékezzünk: a statisztika tömegjelenségekkel, megismételhető vizsgálatokkal foglalkozik.)
megfigyelés (vizsgálat, **mérés)**: az adatok megszerzésére irányuló tevékenység. Egy jelenség során több mindent is megfigyelhetünk, miáltal más-más adatra tehetünk szert (pl. kockadobás estén: a dobott szám vagy a kocka repülési ideje)
eredmény (**kimenetel, elemi esemény)**: olyan esemény, amelyhez egyetlen kimenetel tartozik (egyelemű halmaz, az eseménytér részhalmaza, illetve akár több esemény részhalmaza is lehet). [Elemi eseményekről szigorúan véve akkor szoktunk beszélni, ha azokból csak véges sok lehet.]
esemény: egy állítás, amely egy megfigyelés során vagy bekövetkezik, vagy nem (többelemű halmaz, az eseménytér részhalmaza). Az esemény több elemi eseményt

tartalmaz, azoknak többféle önkényesen definiált kombinációja lehet.
eseménytér: adott megfigyeléshez tartozó lehetséges eredmények (azaz elemi események) halmaza (alaphalmaz). Az összes lehetséges kimenetel együttvéve.
statisztikai változó: olyan változó, amelynek az értéke a jelenség megfigyelése során észlelt esemény. A kockadobás esetén (ha a vizsgáltat térgya a dobott szám) 1, 2, 3, 4, 5 és 6 értékeket vehet fel a változó; az érték kialakulásában főszerep jut a véletlennek.
alapsokaság (populáció): az összes elméletben kivitelezhető mérési eredményt tartalmazó halmaz; lehet véges vagy végtelen.
minta: az alapsokaság részhalmaza, azon kimenetelek összessége, amelyeket egy méréssorozat során megfigyeltünk.

modell: a valóság leírására használt rendszer, amely leegyszerűsített (azaz a valóságnak csak egy részét veszi figyelembe) és absztrakt (elvont, azaz nem az egyes dolgokra, hanem azok általános, közös tulajdonságaira vonatkozik); pl. ideális gáz, abszolút fekete test, tömegpont, faj
matematikai modell: a modell törvényszerűségeinek számszerű leírása



A statisztikai változók csoportosítása többféle szempont szerint is lehetséges attól függően, hogy mi a célunk. Első közelítésben soroljuk a változókat minőségi azaz kvalitatív és mennyiségi, azaz kvantitatív típusba. A minőségi változókra alapvetően jellemző, hogy kategóriákat állapítunk meg (ezek lehetnek természetes vagy kevésbé természetes, önkényesen kijelölt kategóriák), és ezekbe soroljuk a változó által felvett értéket. A kategóriák lehetnek egyenrangúak, ilyenkor nominális vagy névleges változóról beszélünk, ilyen a vércsoport, vagy a pénzfeldobásos kísérlet értékei. Ha a kategóriák között rangsor van, akkor ordinális, azaz sorrendi változóról van szó, amelynek értékeit rangsorolt (és rendszerint sorszámmal jelölt) kategóriákba helyezhetjük.

A számszerű változók felvehetnek diszkrét számértéket (ilyen változókat kapunk számlálással, pl. fogak száma vagy évszám), illetve folytonos számértéket (ilyen változókat méréssel kapunk, pl. nyomás vagy hossz).

Az egyes változók jellemzése, az elvégezhető műveletek és az ábrázolási lehetőségek a változó típusától függenek.

Az előbbi intuitív osztályozást célszerű finomítanunk aszerint, hogy milyen céllal akarunk foglalkozni a változókkal. Stanley Smith Stevens: On the Theory of Scales of Measurement [A mérési skálák elméletéről] a Science tudományos lapban 1946-ban megjelent cikkben ismertette a különféle mérési skálákról és az azok közötti hierarchiáról felállított rendszerét. Habár – a tudományos életben normális módon – elméletét többen vitatták, világos eligazítást nyújt a gyakorlatban használt mérési skálákról és a hozzájuk kapcsolódó mennyiségeken végezhető műveletekről.

A legprimitívebb skála, a nominális vagy névleges skála, mely a mérésszinthierarchia alján áll. Ilyen például maga a névadás, vagy a vércsoport, a hajszín, szemszín, állampolgárság stb. A skála létrehozása úgy történik, hogy kategóriákat hozunk létre, a kategóriák egyszerű névadással azonosíthatók. Az egyes megfigyelések során megállapítható, hogy két elem azonos vagy nem azonos. A kategóriák között nincs természetes sorrend, de praktikus okokból kialakíthatnak (ABC-rend, sorszámmal való jelölés), amiket a szokásoknak megfelelően használnak, hogy később könnyebb legyen az összehasonlítás. Azonban ezeknek a sorrendeknek semmiféle mennyiségi jelentése nincs. Emiatt a skála megnevezés is kissé megtévesztő, helyesebb inkább rendszert használni, ami nem enged természetes sorrendiségre következtetni. A kategóriák elhatárolása lehet könnyebb (magától értetődő, pl. fej, írás) vagy nehezebb (mesterséges, pl. szemszín).

Az ordinális skála szintén kategóriákat jelöl ki, azonban ezek között már természetes sorrend van, ilyen például az iskolai osztályzat, a betegségek, sérülések súlyossága vagy a Mohs-skála. Az ordinális skálán theát nem csak azonosságot tudunk megállapítani,

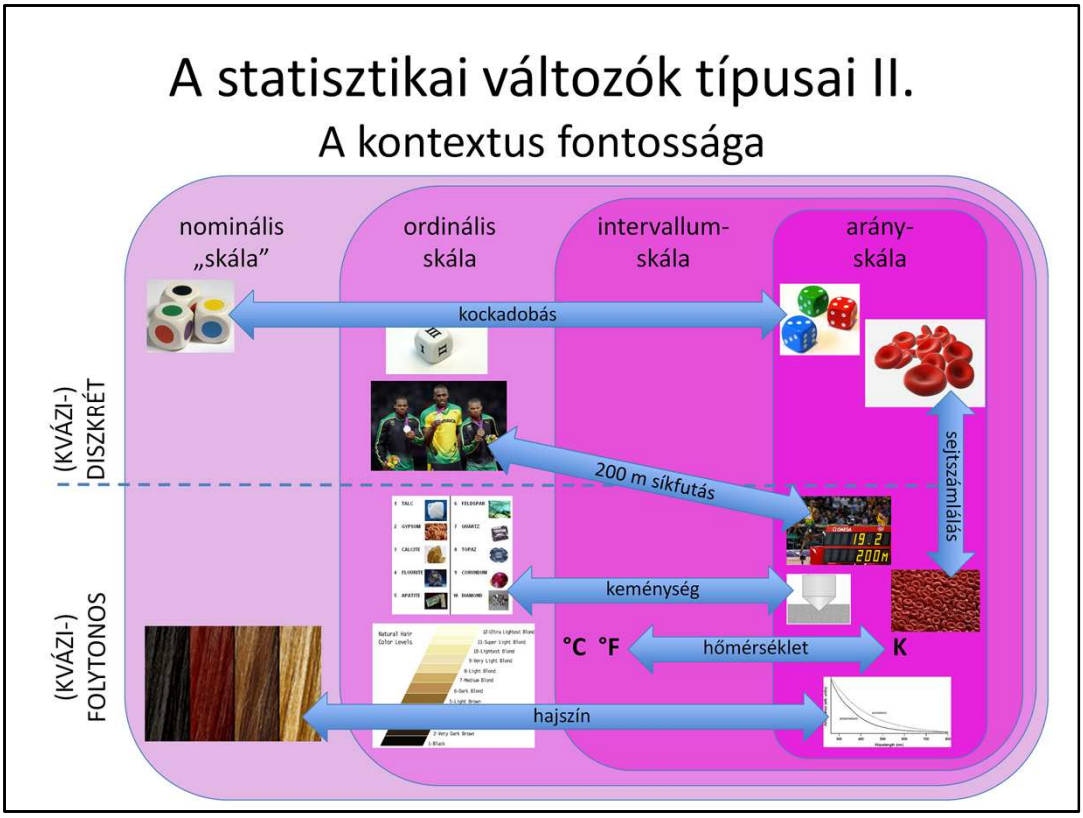
hanem kisebb/nagyobb relációt is. A skálaelemeket rendszerint sorszámmal jelölik, amit észben kell tartani, hiszen sorszámokon nem végezhetők el a szokásos matematikai műveletek. Az ordinális skála kategóriái közötti eltérés vagy távolság nem egyenlő vagy nem tudjuk megállapítani.

Az intervallumskála annyiban fejlettebb az ordinális skálánál, hogy ismert a felvehető értékek közötti távolság, vagyis már nem csak a sorrend, hanem a különbség és az összeg is értelmezhető. A mindennapi életből ismertpéldák pl. az évszám, a Celsius-fokban mért hőmérséklet vagy a tengerszinthez viszonyított magasság. A példákból látható az intervallumskálák egy további közös tulajdonsága: a nullapont kijelölése egyezmény alapján történik.

Ezen konvencionális nullaérték helyett természetes nullaértéken alapulnak az arányskálák: az arányosságot maga a természetes nullapont létezése teszi lehetővé. Az ilyen skálákon így már az arányossághoz kapcsolódó műveletek, az osztás és a szorzás is értelmezhető.

Mindegyik skálaszinten elkülöníthetők többé-kevésbé diszkrét és folytonos változók. Nominális változókat tekintve például az érmefeldobásnál egyértelműek a diszkrét kategóriák, míg a szemszín esetén eléggé homályos az egyes kategóriák határa, illetve a kategóriák száma, igazából csak rajtunk múlik, hogy mennyire finom felosztást hozunk létre.

Stevens nevével a pszichofizikai törvények kapcsán még fogunk találkozni a második félévben a biofizika tantárgy keretében.



A különféle változók nem mindig sorolhatók egyértelműen valamelyik kategóriába, s ha ez még így is van, nem ritka, hogy gyakorlati vagy történelmi okokból nem a mennyiség jellegéhez legjobban illeszkedő vagy esetleg abból egyértelműen adódó skálát használjuk.

Pl. az első hőmérsékleti skálákat még azelőtt alkották meg, hogy az abszolút nulla pont léte bebizonyosodott volna. Így bár a hőmérséklet a természetéből adódóan arányskálával mérhető, hétköznapi célokra intervallumskálát használunk. Ugyanez már a tengerszint feletti magasságra nem igaz, ott elméletileg sincs nullapont (a Föld nem tökéletes gömb, így nem igazán definiálható pl. egy középpont). Ugyanez mondható el a többi “potenciális” mennyiségről is.

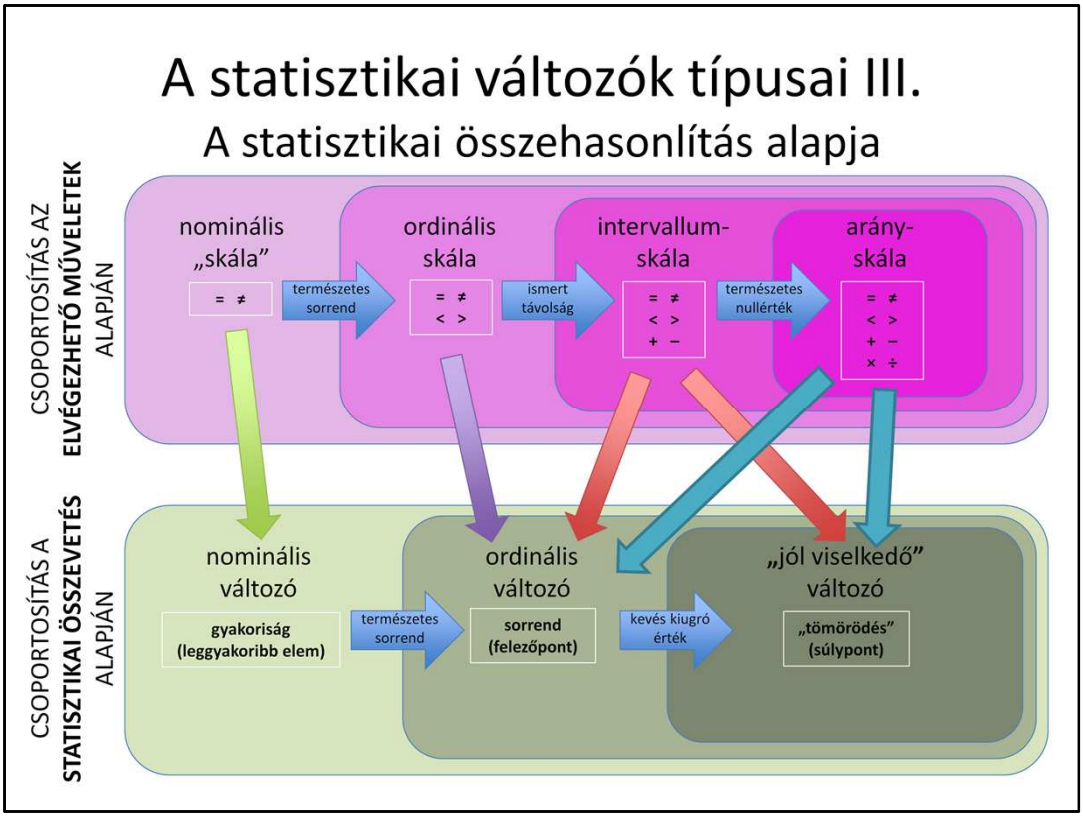
A kockadobás kimenetelének mint változónak megítélése még kevésbé egyértelmű. Alapvetően egy kategoriális változóról van szó, hiszen a kocka mint puszta geometriai test oldalai között semmiféle rangsor nincs. Az oldalakat színekkel jelölve már megkülönböztethetők lesznek, de pusztán ettél még nem alakul ki rangsor. Az egyes oldalakhoz sorszámokat is rendelhetünk, ekkor ordinális változót kapunk. Ha pöttyöket rajzolunk az oldalakra, amelyek megszámlálhatók, intervallumskálán kifejezhető változót kapunk, sőt, bár nullát nem dobhatunk, a pöttyszámlálás esetén a skála nullapontja (= nincs pötty) is jól definiált (vagyis attól, hogy egy mérés során a nulla kimenetel nem jöhet létre, magának a mérési skálának lehet – akár természetes – nullapontja). Mindezzel együtt az elvégezhető műveletek köre is bővül a korábban elmondottaknak megfelelően. Azt azonban észben kell tartani, hogy az így magasabb skálaszintre emelt változó “többletjelentése” továbbra sem a kocka tulajdonságait jellemzi, hanem az

általunk kreált és a kockához rendelt szabályrendszert.

Ehhez a példához valamelyest hasonló, de folytonos változóra vonatkozó példa a haj (vagy szem) színe. A hétköznapokban gyakran csak annyit mondunk, hogy valakinek a haja színe barna, fekete, szőke vagy vörös. Ha alaposabb megfigyelők vagyunk, finomíthatjuk a kategóriákat, pl. törtfehér, sötétbarna, gesztenyeszín, tűzvörös, világosbarna stb., de ez a skálaszintet még nem érinti, csak a jellemző folytonos mivoltára és a létrehozott kategóriák mesterkéltségére utal. Lehetséges azonban ezeket a kategóriákat például egy sötét-világos skála szerint sorrendbe rendezni, sőt, precíz mérésekkel színskálát is rendelhetünk a szemhez. Vagyis a hétköznapokban puszta kategóriákra egyszerűsített szemszínt egy numerikusan kifejezhető folytonos változóvá alakíthatjuk, amely sokkal precízebben felel meg a valóságnak.

Gyakorlati szempontok szabják meg azt is, hogy a vörösvérsejtszámot diszkrét vagy folytonos változónak tekintjük. Ha csupán néhány sejtet kell számlálnunk pl. egy elektronmikroszkópos felvételen, akkor célszerű diszkrét változóként kezelni, azonban egy nagyobb vérmintában több millió vérsejt lehet, vagyis gyakorlatilag folytonossá válik a változónk.

Azt, hogy – egyebek mellett – a 200 méteres síkfutást a Londoni olimpián Usain Bolt nyerte, elég sokan tudják, arra viszont már minden bizonnyal csak kevesen emlékeznek, hogy a távot 19.32 másodperc alatt futotta le. Habár a megtett idő mint arányskálán mért folytonos változó több információt tartalmaz, mint az ordinális skálán mért helyezés, utóbbi érték megjegyzése éppen ezért kevesebb erőfeszítést igényel.



Visszatérve a statisztika egyik kulcskérdésére, az azonosság és különbség megállapítására, ismertetnünk kell egy harmadik féle csoportosítást. Valamennyi változót összehasonlíthatunk az egyes értékek előfordulási gyakorisága alapján. Például két ország lakosságát összevethetjük az AB0 vércsoportok százalékos előfordulási gyakorisága alapján. Leegyszerűsíthető ez a fajta összehasonlítás úgy, hogy csak azt vizsgáljuk, hogy a leggyakoribb érték megegyezik-e vagy sem.

Ha az elemek sorrendbe állíthatók, akkor meghatározható egy felezőpont, amelynél ugyanannyi elem kisebb, mint ahány nagyobb. Sorba rendezhető elemeket tartalmazó halmazokat tehát összehasonlíthatunk a felezőpontjuk szerint.

Az összehasonlítás tovább finomítható, ha nem azt vizsgáljuk, hogy a változó melyik értékénél felezhető el a halmaz, hanem ha azt a pontot, ahol az adathalmaz egyensúlyban van, vagyis amelytől számítva a távolsággal súlyozott elemszám azonos. Ez a pont (amit átlagnak nevezünk és később fogjuk precízen bevezetni) theát nem csak attól függ, hogy nála kisebb és nagyobb elemből hány van, hanem attól is, hogy azok milyen távolságra esnek, így az előbb leírt felezőpontnál sokkal érzékenyebb. Éppen emiatt azonban csak olyan esetben van értelme használni, amikor a változó úgyszólván “jól viselkedik”, vagyis nincs sok kiugró érték, amely eltorzítaná az súlypont által adott képet. Ezért azokat az intervallum- és arányskálán mért adathalmazokat, amelyek viszonylag sok kiugró értéket tartalmaznak (vagyis “rosszul viselkednek”), az ordinális változókhoz hasonlóan kezeljük.

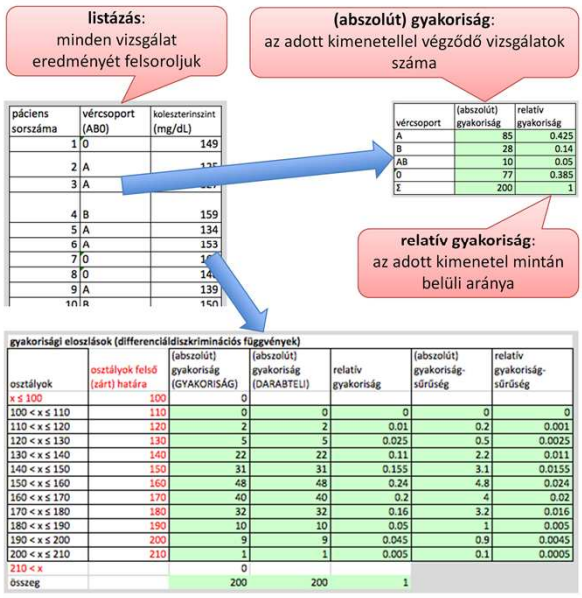
Miután áttekintettük a statisztikai változók sokféleségét és csoportosításukat, tisztázzuk az alapsokaság és a minta fogalmát.

Mint említettük, a statisztika olyan jelenségeket vizsgál, amik megismételhetők. Ez azt jelenti, hogy a jelenségek vizsgálata során sok, akár végtelensok mérést is végezhetünk. Ezen elméletileg lehetséges összes mérés kimeneteleinek, eredményeinek összefoglaló halmazát nevezzük alapsokaságnak. Elméletben a statisztikai változó teljes megismeréséhez ezt a végtelensok mérést el kellene végeznünk, de erre nyilvánvalóan nincs lehetőségünk, ezért annak csak egy részhalmazát, a mintát vizsgáljuk. A minta tehát az alapsokaság részhalmaza, amelynek létrehozására a legkézenfekvőbb módszer a véletlen kiválasztás.

A létrehozott mintán méréseket végzünk, a keletkező mérési eredmények halmazát szinén mintának nevezzük. (Magyarán: kevésbé precíz fogalmazással pl. az évfolyam mint alapsokaság egyik csoportjának tagjait tekinthetjük az évfolyamból vett mintának; precízebben fogalmazva az évfolyam hallgatóinak vércsoportadatai jelenthetnek alapsokaságot, míg az egyik csoport tagjainak vércsoportadatai egy lehetséges mintát.) A mintát jellemezhetjük grafikusan és számszerűen, majd a minta így megállapított tulajdonságait extrapolálhatjuk, kiterjeszthetjük az alapsokaságra. Az előbbi példánál maradva: amilyen arányban a mintában előfordulnak az egyes vércsoportok, kb. olyasmit várunk el az egész alapsokaságtól is. Mivel a minta összeállítása véletlenszerűen történik, nem biztos, hogy tökéletesen reprezentálja az alapsokaságot, a különböző értékek alapsokaságon belüli előfordulási arányát. Így a minta alapján levont következtetésekhez mindig társul valamekkora bizonytalanság.

A minta szemléltetése I.

1) *A minta elemeinek egyszerű listázása*
2) *A gyakoriságok táblázatos összefoglalása*
   - abszolút gyakoriság (Δn) és relatív gyakoriság ($\Delta n/n$)
   - kvalitatív változók esetén a kategóriák adottak [Excel: =DARABTELI() vagy =COUNTIF() függvény]
   - kvantitatív változók esetén a kategóriákat (osztályokat) magunk definiáljuk [Excel: =GYAKORISÁG() vagy =FREQUENCY() függvény is használható]
   - gyakoriságsűrűség ($\Delta n/\Delta x$) és relatív gyakoriságsűrűség ($[\Delta n/n]/\Delta x$)



páciens sorszáma	vércsoport (AB0)	koleszterinszint (mg/dL)
1	0	149
2	A	
3	A	
4	B	159
5	A	134
6	A	153
7	0	
8	0	
9	A	139

vércsoport	(abszolút) gyakoriság	relatív gyakoriság
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

gyakorisági eloszlások (differenciáldiszkriminációs függvények)

osztályok	osztályok felső (zárt) határa	(abszolút) gyakoriság (GYAKORISÁG)	(abszolút) gyakoriság (DARABTELI)	relatív gyakoriság	(abszolút) gyakoriság-sűrűség	relatív gyakoriság-sűrűség
x ≤ 100	100	0				
100 < x ≤ 110	110	0	0	0	0	0
110 < x ≤ 120	120	2	2	0.01	0.2	0.001
120 < x ≤ 130	130	5	5	0.025	0.5	0.0025
130 < x ≤ 140	140	22	22	0.11	2.2	0.011
140 < x ≤ 150	150	31	31	0.155	3.1	0.0155
150 < x ≤ 160	160	48	48	0.24	4.8	0.024
160 < x ≤ 170	170	40	40	0.2	4	0.02
170 < x ≤ 180	180	32	32	0.16	3.2	0.016
180 < x ≤ 190	190	10	10	0.05	1	0.005
190 < x ≤ 200	200	9	9	0.045	0.9	0.0045
200 < x ≤ 210	210	1	1	0.005	0.1	0.0005
210 < x		0				
összeg		200	200	1		

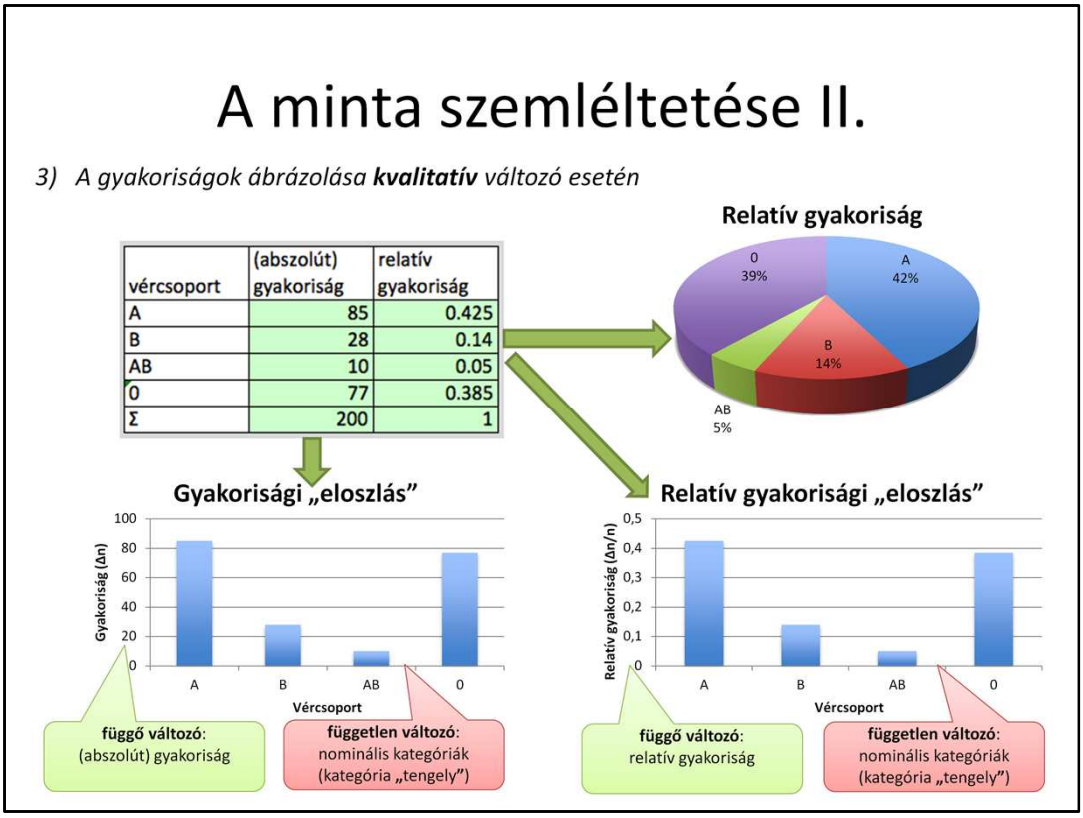
Egy minta bemutatásának legegyszerűbb módja a minta elemeinek felsorolása (prakatikusan a mérés sorrendjében). Ez a fajta listázás a minta elemeinek puszta létezésén túl nem sok mindent mutat. Nagy elemszámú minta esetén is így tároljuk az adatokat, de szemléltetésre már nem alkalmas.

Szemléletesebb a minta elemeit kategóriákba sorolni, majd az előfordulási gyakoriságot kategóriánként megadni. A besoroláshoz szükséges kategóriák minőségi változók esetén az elemi események alapján egyértelműen adódnak, bár összevonásra szükség lehet (pl. ha egy kategóriába kiugróan kevés elem kerülne). Mennyiségi változóknál már sokszor nekünk kell önkényes kategóriákat (másnéven intervallumokat vagy osztályokat) definiálnunk, és ezekbe sorolhatjuk az adatokat. Az intervallumok kialakításának vannak praktikus szempontjai, amit elsősorban a minta terjedelme és az elemszám szab meg (lásd később). Nagyobb elemszám esetén például nagyobb számú, így kisebb (keskenyebb) intervallumokat célszerű definiálni. Azonban tartsuk észben, hogy az intervallum határok kijelölésével – akár ugyanolyan széles intervallumokat véve – mi magunk is befolyásoljuk a gyakoriságok értékeit.

Az egyes kategóriákba tartozó elemek számát abszolút gyakoriságnak nevezzük. Azonban különböző méretű minták összehasonlítását megkönnyíti, ha nem az abszolút, hanem a relatív gyakoriságokat (vagyis a mintán belüli [százalékos] arányt) használjuk. Így könnyen összehasonlítható pl. egy kisváros és egy ország lakosságának vércsoport szerinti megoszlása.

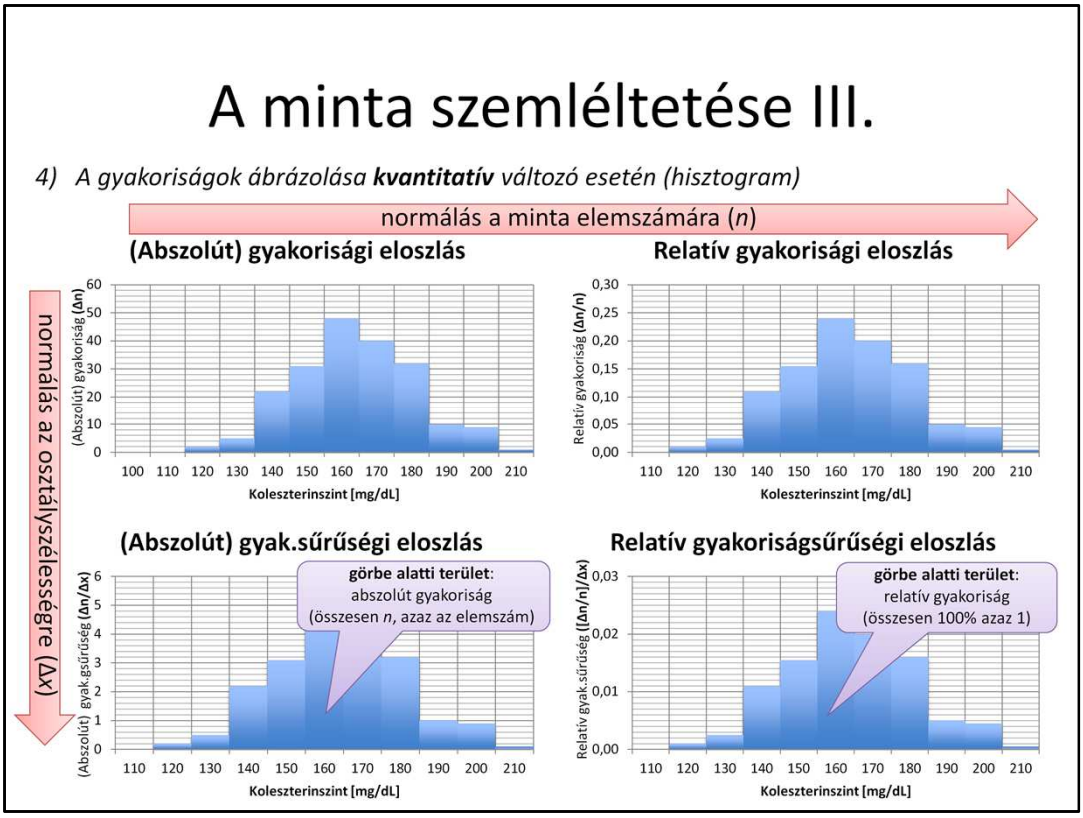
Számszerű változók esetén azonban nem csak a minta elemszámában adódhatnak

jelentős eltérések, hanem a kialakított osztályok szélességében is. Nyilvánvaló, hogy ha szélesebb osztályokat használunk, akkor ugyanazon minta esetén is eltérők lesznek a gyakoriságok és a relatív gyakoriságok. Ezért az összehasonlítás érdekében célszerű ezeket a gyakoriságokat az osztály szélességére vonatkoztatni (magyarán azzal elosztani). Az így kapott mennyiségek: a(z abszolút) gyakoriságsűrűség és a relatív gyakoriságsűrűség.



Kvalitatív mennyiségből álló minta esetén oszlopdiagramot vagy kördiagramot célszerű használni az adatok ábrázolásához. Az oszlopdiagram elsőre egy klasszikus függvény koordinátarendszerére hasonlít, azonban fontos eltérés, hogy az x tengelyen a kvalitatív (a példában nominális kvalitítv) változót veszünk fel. A nominális “skála” korábban részletezett tulajdonságaiból fakadóan sem a kategóriák sorrendjének, sem a köztük lévő távolságnak vagy a szélességüknek nincs jelentése, ezeket praktikus (pl. ábécérend, gyakoriság szerinti sorrend stb.) és esztétikai szempontok szerint alakítjuk ki. A függő változó természetesen az abszolút vagy relatív gyakoriság (utóbbit gyakran százalékos formában tüntetik fel).

A kördiagram esetén a kör a teljes mintának felel meg, míg az egyes szeletek mérete a változó különféle értékeinek mintán belüli előfordulási gyakoriságával arányos. A szeletek sorrendjét itt is praktikus szempontok szabják meg, matematikai jelentésük nincs.



Kvantitatív változó esetén az x tengelyen az intervallumskálát vesszük fel, amelyet praktikus szélességű önkényes kategóriákra, osztályokra osztunk fel.

Az y tengelyen felvehetjük az abszolút gyakoriságot, ekkor az oszlopok magasságáról közvetlenül leolvasható az adott osztályba tartozó elemek száma. Ha a relatív gyakoriságot ábrázoljuk, akkor az oszlopok magassága az adott osztályba tartozó elemek mintán belüli hányadának felel meg.

Ha az abszolút gyakoriságot normáljuk az osztályszélességre, akkor megkapjuk a(z abszolút) gyakoriságsűrűséget. Ezt felvéve az y tengelyen a oszlopok (téglalapok) területe fog megfelelni az adott osztály elemszámának. A teljes Hasonlóan, a relatívgyakoriságsűrűség

Megjegyzés: a sűrűség kifejezés általános jelentése, hogy valamilyen egységre (egységnyi térfogatra, felületre vagy szakaszra) vonatkoztatunk. Háromdimenziós példát említve a hagyományos “sűrűség” (tömegsűrűség) az egységnyi térfogatra vonatkoztatott tömeget adja meg. Két dimenzióban használt mennyiség a felületi sűrűség (például papírt szokták vele jellemezni), amely az egységnyi felületre vonatkoztatott tömeg (pl. egy négyzetméter papírlap tömege). A gyakoriságsűrűségnél lényegében a Δx szakaszhosszra vonatkoztatjuk a gyakoriságot («tömegét»), vagyis ez egy egydimenziós sűrűségnek tekinthető.

Ellenőrző kérdések I.

- Mik a tudomány meghatározó tulajdonságai?
- Mik a tudományos gondolkodás jellemzői?
- Mi különbözteti meg a tudományos orvoslást a kuruzslástól?
- Hogyan győződhetünk meg arról, hogy egy orvosi módszer tudományos bizonyítékokon alapul?
- Mondj példát nem bizonyítékon alapuló gyógymódokra!
- Kit terhel a bizonyítási kényszer: Aki egy új módszerről azt állítja, hogy hatásos, vagy aki azt, hogy hatástalan?
- Mi a statisztika?
- Mi a statisztika célja?
- A matematika melyik területére támaszkodik leginkább a statisztika?
- Mi a modellalkotás célja?
- Milyen kapcsolat van a valóság és a modell között?
- Mit kell egy gyógyszerről/gyógymódról bizonyítani: hogy hatásos vagy hogy hatástalan?
- Mi a probléma a következő állítással: „Erről a módszerről még senki se bizonyította be, hogy hatástalan, így igazságtalan lenne a felhasználását korlátozni.”
- Hogyan csoportosíthatjuk a beteg megfigyelése során keletkező adattípusokat?
- Sorolj fel nominális statisztikai változókat!
- Sorolj fel ordinális statisztikai változókat!
- Sorolj fel diszkrét statisztikai változókat!
- Sorolj fel folyotonos statisztikai változókat!
- Sorolj fel jól viselkedő statisztikai változókat!
- Sorolj fel nem jól viselkedő statisztikai változókat!
- Mondj példát nominális „skálára”!
- Miért helytelen a nominális „skála” kifejezésben a skála szó?

A kérdések megválaszolhatók az előadáson elhangzottak, a gyakorlatvezetővel folytatott konzultációk, illetve saját utánaolvasás segítségével. Az ellenőrző kérdések egyben példák arra, hogy milyen tesztkérdések (feleletválasztós formában) fordulhatnak elő.

Ellenőrző kérdések II.

- Mondj példát ordinális skálára!
- Milyen lényeges eltérés van a nominális és az ordinális skála között?
- Mondj példát intervallumskálára!
- Milyen eltérés van az ordinális és az intervallumskála között?
- Mondj példát arányskálára!
- Miért fontos a statisztikai változó pontos definiálása?
- Szemléltesd példával a statisztikai változó kontextusának jelentőségét!
- Hogyan csoportosíthatók a statisztikai változók a statisztikai összehasonlíthatóság szemszögéből?
- Milyen kapcsolat van a Stevens-féle méréselmélet skálaszintjei és a statisztikai összehasonlítás alapján kialakított változóhierarchia között?
- Mit jelent az, hogy egy változó "jól viselkedik"?
- Hogyan jellemezhetők a statisztikai összehasonlíthatóság alapján létrehozott hierarchia egyes szintjein az adathalmazok?
- Egy (Stevens szerinti) intervallumskálán mért mennyiség "jól viselkedőnek" számít a statisztikai összehasonlítás szempontjából?
- Mit nevezünk alapsokaságnak?
- Mit értünk minta alatt?
- Hogyan választjuk ki a mintát az alapsokaságból?
- Mi az oka annak, hogy a minta nem tökéletesen képviseli az alapsokaságot?
- Mi az oka annak, hogy a minta alapján az alapsokaságról levont következtetéseket bizonytalanság terheli?
- Hogyan jellemezhetjük a mintánkat?
- Mit értünk gyakoriság és abszolút gyakoriság alatt?
- Ha csak annyit mondunk, hogy gyakoriság, az alatt mi értendő: relatív vagy abszolút gyakoriság?
- Hogyan lehet egy adathalmaz gyakoriságait táblázatosan összefoglalni?

Ellenőrző kérdések III.

- Mire kell figyelnünk numerikus változók gyakoriságainak táblázatos összefoglalásakor?
- Hogyan tehetők különféle méretű minták gyakoriságai összehasonlíthatóvá?
- Hogyan tehetünk különféle osztályszélességekkel készült gyakoriságokat összevethetővé?
- Mitől relatív a relatív gyakoriság?
- Mit jelent a gyakoriságsűrűség kifejezésben a sűrűség?
- Milyen célszerű ábrázolási lehetőségek vannak kvalitatív változók esetén?
- Mik szerepelnek egy kvalitatív adathalmazból készült oszlopdiagram vízszintes tengelyén?
- Milyen jelentése van egy kvalitatív adathalmazból készült oszlopdiagramon a kategóriatengelynek?
- Mi és hogyan olvasható le egy gyakorisági eloszlás grafikonjáról?
- Mi és hogyan olvasható le egy relatív gyakorisági eloszlás grafikonjáról?
- Mi és hogyan olvasható le egy gyakoriságsűrűségi eloszlás grafikonjáról?
- Mi és hogyan olvasható le egy relatív gyakoriságsűrűségi eloszlás grafikonjáról?
- Mekkora a teljes görbe alatti terület egy gyakoriságsűrűségi és egy relatív gyakoriságsűrűségi eloszlás grafikonja estén?

Ellenőrző kérdések IV.

- Sorold be a Stevens-féle skálaszintek alapján a következő statisztikai változókat, illetve mérési skálákat!
  - szemszín
  - hajszín
  - beteg neve
  - beteg neme
  - a Beaufort-skála szerinti szélerősség
  - szélsebesség kilométer per órában mérve
  - iskolai osztályzatok Magyarországon számmal kifejezve (5; 4; 3; 2; 1)
  - iskolai osztályzatok Magyarországon szövegesen kifejezve (jeles, jó, közepes, elégséges, elégtelen)
  - iskolai osztályzatok az Egyesült Államokban (A; B; C; D; F)
  - ásványok Mohs-féle keménységi skála szerinti keménysége
  - a közvéleménykutatáshoz használt Likert-skála (1 = teljesen egyetért, 2 = inkább egyetért, 3 = egyet is ért meg nem is, 4 = inkább nem ért egyet, 5 = egyáltalán nem ért egyet)
  - intelligenciahányados (IQ)
  - név sorszáma egy listán
  - fogak száma
  - gyermekek száma egy családban
  - vércukorszint
  - testmagasság
  - testtömeg
  - beteg állampolgársága
  - gyógyszertabletta tényleges hatóanyagtartalmának eltérése a névleges hatóanyagtartalomtól
  - sérülés súlyossága (I.–IV. fokú égési sérülés)
  - háztartási készülékek energiafogyasztás szerinti besorolása az Európai Unióban (A+++; A++; A+; A; B; C; D)
  - egy labdarúgócsapat helyezése a tabellán
  - pulzusszám
  - katonai rangjelzés
  - egy *steak* készültsége, átsütöttsége (*blue rare, rare, medium rare, medium, medium well, well done*)
  - dobókockával dobott szám
  - az érmefeldobás eredménye
  - vércsoport (AB0)
  - vércsoport (Rhesus-faktor)
  - Kelvin-skála
  - Celsius-skála
  - Fahrenheit-skála
  - Rankine-skála

Függelék

Műveletek gyakoriságokkal

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		

Abszolút gyakoriság: az adott halmazba tartozó elemszám
Relatív gyakoriság: a teljes mintára vonatkoztatott elemszám
Feltételes relatív gyakoriság: a minta egy részhalmazára vonatkoztatott elemszám.

A mintán belül az abszolút gyakoriságok összeadhatók.
Példa: mirigyláz és fertőző májgyulladás összesen 22 + 22 = 44 embert érintett.

A mintán belül a relatív gyakoriságok összeadhatók.
Példa: mirigyláz és veszettség az embereknek összesen 6,5% + 22% = 28,5%-t érintette.

A részhalmazon belül az arra vonatkoztatott feltételes relatív gyakoriságok összeadhatók.
Példa: A vírusos megbetegedéseken belül 17,5% a mirigyláz és 6,3% egyéb. Ezek együttesen a vírusos megbetegedések 23,8%-át teszik ki.

A feltételes relatív gyakoriságot megszorozva a vonatkoztatási részhalmaz relatív gyakoriságával megkapjuk a relatív gyakoriságot.
Példa: A mirigyláz vírusos megbetegedéseken belüli aránya 17,5%. A vírusos betegségek mintán belüli hányada 37,5%. A mirigyláz mintán belüli aránya tehát 17,5%*37,5% = 0,175*0,375=0,065=6,5%