8. A leíró statisztika jellemzői, diagramok. Nevezetes középértékek

Vázlat:

I. Adatsokaságok jellemzői (diagram, táblázat, osztályokba sorolás)
II. A leíró statisztika jellemzői: mintavétel, gyakoriság, relatív gyakoriság, táblázat, osztályba sorolás
III. Statisztikai mutatók: középértékek (módusz, átlag, medián, kvartilisek), terjedelem, szórás, átlagtól való abszolút eltérés
IV. Diagramok: kör-, oszlop-, vonal-, sodrófa (boxplot) diagram, gyakorisági diagram
V. Nevezetes középértékek (számtani, mértani, harmonikus, négyzetes)
Közepek közti összefüggések
VI. Nevezetes középértékek alkalmazása szélsőérték-feladatokban
- összeg állandósága esetén szorzat maximalizálása
- szorzat állandósága esetén összeg minimalizálása

VII. Alkalmazások, matematikatörténeti vonatkozások

Kidolgozás:

I. Adatsokaságok jellemzői

DEFINÍCIÓ: A statisztika feladatai közé tartozik, hogy bizonyos egyedek meghatározott tulajdonságairól tájékozódjék, majd a szerzett (általában számszerű) adatokat feldolgozza, elemzi. Az elemzéshez összegyűjtött adatok halmazát adatsokaságnak, mintának, a meghatározott tulajdonságot ismérvnek, változónak nevezzük. A sokaság elemeinek az ismérv szerinti tulajdonságát statisztikai adatnak, az adatsokaság elemeinek számát a sokaság méretének nevezzük.

II. A leíró statisztika jellemzői

A leíró statisztika a tömegesen előforduló jelenségekkel, a jelenségekből nyert adatok vizsgálatával, elemzésével (leírásával) foglalkozik.
A statisztika egyik fontos feladata az adatok összegyűjtése. Ha a vizsgálandó egyedek száma nagyon nagy, akkor nem minden egyedet vizsgálunk meg a tulajdonság alapján, hanem az adatsokaságnak vesszük egy részhalmazát, vagyis az egyedek közül **mintát veszünk**. A megfelelően kiválasztott minta elemzéséből következtethetünk a sokaság adataira.
A **reprezentatív mintavétel**nél törekedni kell arra, hogy a vizsgált tulajdonság előfordulása a mintában közelítse a sokaságban való előfordulását. Pl. közvélemény-kutatás.
Véletlenszerű mintavételnél a sokaság elemei egyenlő valószínűséggel kerülnek a mintába. Pl. urnából húzás.

DEFINÍCIÓ: Az egyes adatok előfordulásának a száma a **gyakoriság.** Az adatok összehasonlíthatósága miatt sokszor a gyakoriságnak a teljes adatsokasághoz viszonyított arányával, a **relatív gyakorisággal** dolgozunk, azaz a gyakoriságot osztjuk az adatok számával.

Az adatokat megadhatjuk **táblázat**os formában, így az adatok áttekinthetően láthatók. Táblázat használatának előnye, hogy nagyobb adathalmazokat tömören, helytakarékosan ábrázolhatunk.
Leggyakrabban a gyakorisági táblázatot használjuk, ez a lehetséges adatokat és a hozzájuk tartozó gyakoriságokat tartalmazza.
Osztályokba soroljuk az adatokat, ha nagy méretű (sok adatból álló) adatsokasággal dolgozunk, vagy ha sok különböző érték van közel azonos gyakorisággal a sokaságban, akkor az egymáshoz

közeli értékek összevonásával az adatokat osztályokba rendezzük. Az osztályba sorolásnál fontos szempont, hogy az osztályoknak diszjunktaknak (különállóknak), de hézagmentesnek kell lennie.
Egy **osztályköz hossza** az osztály felso és alsó határának különbsége. Gyakran azonos hosszúságú osztályokkal dolgozunk. Az **osztályközép** az osztály alsó és felso határának számtani közepe. Ekkor minden, az osztályba tartozó adatot úgy tekintünk, mintha értéke az osztályközép lenne. Az egyes osztályokba tartozó adatok száma a **kumulált gyakoriság**. Ha osztályközepekkel számolunk statisztikai mutatókat, akkor gyakoriságnak mindig a kumulált gyakoriságot használjuk.

III. Statisztikai mutatók

A középértékek

Az adatsokaság egészét csak leegyszerűsítéseket alkalmazva tudjuk jellemezni. Ezt a célt szolgálják a **középértékek**, amelyek egyetlen számmal írnak le egy adathalmazt.
Ezek előnye, hogy megfelelően alkalmazva jól jelenítik meg az egész adatsokaság valamilyen tulajdonságát, ugyanakkor hátrányuk, hogy nem nyújtanak képet az egyes adatokról.

DEFINÍCIÓ: Egy adatsokaságban a leggyakrabban előforduló adat a minta **módusz**a.

Ha a legnagyobb gyakoriság csak egyszer fordul elő az adatsokaságban, akkor az egymóduszú, ha többször is előfordul, akkor többmóduszú, tehát a módusz több elem is lehet, ha ugyanakkora a gyakoriságuk.

A módusz előnye:
– könnyen meghatározható

A módusz hátránya:
– semmitmondó, ha az adatok közel azonos gyakorisággal fordulnak elő
– csak akkor ad használható jellemzést a mintáról, ha a többi adat gyakoriságához képest sokszor fordul elő egy adat, de ekkor sem mond semmit a többiről.

DEFINÍCIÓ: Az adatok összegének és az adatok számának hányadosa **a minta átlaga (számtani közepe).**

Ha egyes adatok többször is előfordulnak, akkor az összegben szorozni kell őket a gyakoriságukkal és az összeget a gyakoriságok összegével osztjuk. Ez a **súlyozott számtani közép**.

Az átlag előnye:
– a nála nagyobb adatoktól vett eltéréseinek összege egyenlő a nála kisebb adatoktól vett eltéréseinek összegével.

Az átlag hátránya:
– egyetlen, a többitől jelentősen eltérő adat eltorzíthatja, így ekkor már nem jól jellemzi a mintát.

DEFINÍCIÓ: Az adatok **medián**ja a nagyság szerinti sorrendjükben a középső adat. Páratlan ($2n + 1$ darab) adat esetében a medián a középső (az $n + 1$-edik) adat, páros ($2n$ darab) adat esetén a két középső (az n-edik és az $n + 1$-edik) adat átlaga.

A definícióból adódik, hogy az összes előforduló ismérvérték (adat) fele kisebb vagy egyenlő, fele nagyobb vagy egyenlő, mint a medián.

A medián előnye:
– az adatoktól mért távolságainak összege minimális,
– valóban középérték, hiszen ugyanannyi adat nagyobb nála, mint ahány kisebb.

DEFINÍCIÓ: Kvartilisek azok a helyzetmutatók, amelyek a nagyság szerint növekvő sorrendbe rendezett adatokat négy, lehetőleg egyenlő mennyiségű részre osztják.

Alsó kvartilis (Q_1) az a szám, amelynél az adatok kb. negyede kisebb. Meghatározása: a mediánnal kettéosztott adatok alsó részének a mediánja.

Felső kvartilis: (Q_3) az a szám, amelynél az adatok kb. negyede nagyobb. Meghatározása: a mediánnal kettéosztott adatok felső részének a mediánja.

A szóródás jellemzői

DEFINÍCIÓ: Az adatok legnagyobb és legkisebb elemének a különbségét a **minta terjedelmé**nek nevezzük.
Minél kisebb a minta terjedelme, annál jobban jellemzi a mintát.
A terjedelem előnye:
– szemléletes, egyszerűen számolható
A terjedelem hátránya:
– egy-két szélsőséges adat elronthatja.

DEFINÍCIÓ: A **félterjedelem** (interkvartilis terjedelem) a felső és alsó kvartilis különbsége. Az adatok felének elhelyezkedését mutatja meg.

Sokszor tapasztalunk **kiugró adat**okat az adatsokaságban, ezek jelentősen eltérnek a többi adattól. A jelentős eltérés szubjektív, nincs rá meghatározás, általában kiugró adatnak tekintjük a felső kvartilistől a félterjedelem 1,5-szeresével „felfelé", vagy az alsó kvartilistől a félterjedelem 1,5-szeresével „lefelé" eltérő adatot. A kiugró adatok torzítják a mintát jellemző mutatókat, ezért sokszor kihagyjuk őket a minimum és maximum számolásakor.

A mintát jellemző számötös: minimum, alsó kvartilis, medián, felső kvartilis, maximum.

DEFINÍCIÓ: Az adatok átlagtól való eltérések négyzetének átlaga a **minta szórásnégyzete**, ennek négyzetgyöke a **minta szórása**: $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$.

A szórás megmutatja, hogy a minta adatai mennyire térnek el az átlagtól. Minél kisebb a szórás, annál jobban jellemzi az átlag az adatsokaságot.

DEFINÍCIÓ: Az **átlagtól való abszolút eltérés**:

$$\frac{\sum_{i=1}^{n}\left| x_i - \overline{x} \right|}{n}.$$

Előnye: az abszolút érték miatt nem egyenlítődnek ki a pozitv és negatív eltérések.

IV. Diagramok

Az adatok grafikus megjelenítése diagramon történik, amelynek típusát a feladat határozza meg.

Oszlopdiagram: az adatok egymáshoz való viszonyát ábrázolja. Nem célszerű használni, ha az adatok közt van 1-2 kiugró érték (túl nagy: nem fér rá a diagramra, túl kicsi: eltörpül a többi oszlop közt), vagy ha az adatok közötti eltérés nagyon kicsi (közel azonosnak látszanak az értékek). A vízszintes tengelyen az adatfajtáknak megfelelő intervallumokat jelöljük, ezek fölé olyan téglalapokat rajzolunk, amelyeknek területe arányos az adatfajta gyakoriságával.

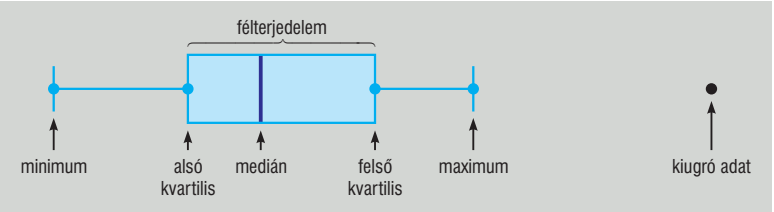
Hisztogram (gyakorisági diagram): az adatok gyakorisági eloszlását oszlopdiagramon ábrázolja úgy, hogy az oszlopok hézagmentesen helyezkednek el.

Sávdiagram: fordított oszlopdiagram, amelyben a két tengely helyet cserél, az oszlopok vízszintesek, azaz sávok.

Kördiagram: a részadatoknak az egészhez való viszonyát ábrázolja. Alkalmas %-os formában megadott adatok ábrázolására. A teljes szög (360º) 100%-nak felel meg, a megfelelő százalékérték egyenesen arányos a körcikk középponti szögével. Nem célszerű használni, ha nagyon sok az adat (túl kicsik a középponti szögek, nem összehasonlíthatók)

Vonaldiagram: koordináta-rendszerben pontként ábrázolja az összetartozó számpárokat, és ezeket töröttvonallal köti össze. Különböző adatok (pl. időbeli) változását ábrázolja. A gyakoriságok vonaldiagramját gyakorisági poligonnak nevezzük.

Sodrófa diagram (dobozdiagram, boxplot): a mintát jellemző szám-ötös (minimum, alsó kvartilis, medián, felső kvartilis, maximum) segítségével ábrázolunk. Képe a minimum és az alsó kvartilis között egy szakasz, az alsó kvartilis és a felső kvartilis között egy téglalap (doboz), benne behúzva a medián, a felső kvartilis és a maximum között szintén egy szakasz. Ha egy két nagyon kiugró adat van az adatsokaságban, akkor azokat kiugró adatként ábrázoljuk és nélkülük határozzuk meg a minimumot, illetve a maximumot. A sodrófa diagram lehet álló, illetve fekvő helyzetű is.



V. Pozitív számok nevezetes középértékei

DEFINÍCIÓ: a_1, a_2, a_3, ..., a_n pozitív számok

számtani (aritmetikai) közepe:

$$A = \frac{a_1 + a_2 + a_3 + \ldots + a_n}{n}$$

mértani (geometriai) közepe:

$$G = \sqrt[n]{a_1 \cdot a_2 \cdot a_3 \cdot \ldots \cdot a_n}$$

négyzetes (kvadratikus) közepe:

$$Q = \sqrt{\frac{a_1^2 + a_2^2 + a_3^2 + \ldots + a_n^2}{n}}$$

harmonikus közepe:

$$H = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \ldots + \frac{1}{a_n}}, \text{ ha } a_1, a_2, a_3, \ldots, a_n > 0.$$

TÉTEL: Középértékek közti összefüggés: $H \le G \le A \le Q$.
Egyenlőség akkor és csak akkor, ha $a_1 = a_2 = a_3 = \ldots = a_n$.

TÉTEL: Két pozitív valós szám esetén $\sqrt{a \cdot b} \le \frac{a+b}{2}$.

BIZONYÍTÁS I.: Mivel az egyenlőtlenség mindkét oldala pozitív, ezért a négyzetre emelés az eredetivel ekvivalens állítást fogalmaz meg. Tehát

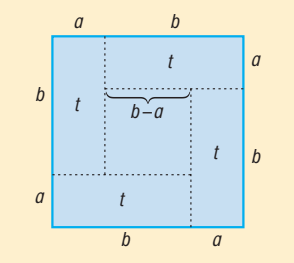
$$\begin{aligned} ab &\le \frac{a^2 + 2ab + b^2}{4} && /\cdot 4 \\ 4ab &\le a^2 + 2ab + b^2 && /-4ab \\ 0 &\le a^2 - 2ab + b^2 && /\text{nevezetes szorzattá alakítjuk} \\ 0 &\le (a-b)^2 \end{aligned}$$

Az utolsó egyenlőtlenség igaz, így az eredeti is az.

Az eredmény alapján megállapítható, hogy a két közép akkor és csak akkor lesz egymással egyenlő, ha $a = b$. Ekkor $a = \sqrt{ab} = \frac{a+b}{2} = b$.

BIZONYÍTÁS II.: Legyen $0 < a \le b$.

Vegyünk fel egy $a + b$ oldalú négyzetet, és az oldalait osszuk fel az ábrán látható módon!



A nagy négyzet területe egyenlő a keletkező részek területének összegével:

$$(a + b)^2 = 4t + (b - a)^2$$

A kis téglalap területe: $t = ab$.

Mivel $(b - a)^2 \ge 0$, ezért ezt a tagot elhagyva az $(a + b)^2 \ge 4t$ egyenlőtlenséghez jutunk.

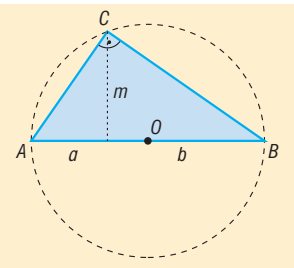
Behelyettesítve t helyére: $(a + b)^2 \ge 4ab$.

Mivel a feltétel miatt mindkét oldal pozitív, ezért gyököt vonhatunk: $a + b \ge 2\sqrt{ab}$.

Amiből $\frac{a+b}{2} \ge \sqrt{ab}$.

BIZONYÍTÁS III.: Legyen $a, b > 0$, $2r = a + b$.

Vegyünk fel egy r sugarú kört, benne egy AB átmérőt, a körvonalon egy A, B-től különböző C pontot.



A Thalész-tétel miatt $ACB\sphericalangle = 90°$.

ABC háromszögre alkalmazva a magasságtételt: $m = \sqrt{ab}$.

De a körben $m \le r$, azaz $\sqrt{a \cdot b} \le \frac{a+b}{2}$.

VI. Nevezetes középértékek alkalmazása szélsőérték-feladatokban

1. Összeg állandósága esetén a szorzatot tudjuk maximalizálni.

Pl.: Azon téglatestek közül, amelyek éleinek összege 60 cm, melyiknek a térfogata maximális?
Legyenek a téglatest élei: a, b és c.
Ekkor a téglatest térfogata $V = abc$, az élek összege: $4(a + b + c) = 60$.
Ebből $a + b + c = 15$.
A számtani és mértani közép közti egyenlőtlenséget kihasználva:

$$\frac{a+b+c}{3} \ge \sqrt[3]{abc} \Rightarrow \left(\frac{a+b+c}{3}\right)^3 \ge abc \Rightarrow \left(\frac{15}{3}\right)^3 \ge abc \Rightarrow 5^3 \ge abc \Rightarrow 125 \ge V.$$

Mivel egyenlőség csak $a = b = c$ esetén teljesül, így a térfogat az 5 cm élű kocka esetén maximális.

2. Szorzat állandósága esetén az összeget tudjuk minimalizálni.

Pl.: Azon téglalapok közül, amelyeknek a területe 100 cm^2, melyiknek a kerülete a minimális? Legyenek a téglalap oldalai a és b.

Ekkor a téglalap területe $t = ab = 100$, kerülete $k = 2(a + b)$, amiből $\frac{k}{4} = \frac{a+b}{2}$.

A számtani és mértani közép közti egyenlőtlenséget kihasználva:

$$\frac{a+b}{2} \geq \sqrt{ab} \;\Rightarrow\; \frac{k}{4} \geq \sqrt{100} \;\Rightarrow\; \frac{k}{4} \geq 10 \;\Rightarrow\; k \geq 40\,.$$

Mivel egyenlőség csak $a = b$ esetén teljesül, így a kerület a 10 cm oldalú négyzet esetén minimális.

Pl.: $f\colon \mathbb{R}^+ \to \mathbb{R},\ f(x) = x + \frac{1}{x}$. Határozzuk meg az $f(x)$ függvény minimumát!

A számtani és mértani közép közti egyenlőtlenséget kihasználva:

$$\frac{x + \frac{1}{x}}{2} \geq \sqrt{x \cdot \frac{1}{x}} \;\Leftrightarrow\; x + \frac{1}{x} \geq 2 \cdot \sqrt{1} \;\Leftrightarrow\; x + \frac{1}{x} \geq 2 \;\Leftrightarrow\; f(x) \geq 2\,.$$

Ekkor az f minimumának értéke $f(x) = 2$, minimum helye: $x = \frac{1}{x} = 1$.

VII. Alkalmazások:

- Statisztika:
  - közvélemény-kutatások,
  - szavazások,
  - gazdasági mutatók,
  - osztályátlagok, hiányzási statisztikák,
  - felvételi átlagpontok
- Nevezetes középértékek:
  - számtani közép: statisztikai átlag kiszámítása,
  - mértani közép: átlagos növekedési ütem kiszámítása, magasságtétel, befogótétel,
  - négyzetes közép: statisztikai szórás kiszámítása,
  - harmonikus közép: átlagsebesség meghatározása

Matematikatörténeti vonatkozások:

- A különféle középértékeket görög **Pitagorasz** és tanítványai vezették be a Kr. e. VI-V. században. Ők foglalkoztak az $a : b = b : c$ aránypár vizsgálatával. Így jutottak el a „mértani középarányos” fogalmához. Valószínűleg az 1 és a 2 mértani közepének keresésekor találták meg az első irracionális számot, a $\sqrt{2}$ -t.
- A statisztika eredetileg „államszámtan” volt. A statisztika kifejezés a latin status (állam, állapot) és az olasz statista (köztisztviselő, politikus) szavakból származtatható. A statisztika már az ókortól kezdve arról tájékoztatta az államok vezetőit, hogy mekkora adókat vethetnek ki az alattvalóikra, azokból mennyi bevételük van, mekkora katonasággal számolhatnak egy eljövendő háborúban. **Kínában** már 4000 évvel ezelőtt összeírták a lakosságot, az ingatlanokat, az ingóságokat. **Angliában** a XI. században összeírták a földbirtokokat.
- **Magyarországon** a középkorban a dézsmajegyzékek (kilenced, tized), majd az újkorban az urbáriumok 1530-tól (tartalmazta a jobbágyok állatállományát, eszközeit, szerszámait, telkének nagyságát és milyenségét is), jobbágyösszeírások 1700-as években, népszámlálások 1800-as évektől jelentették a statisztika alapjait.
- A statisztika a polgári forradalmak után vált igazi tudománnyá. A kapitalizmusban a államok vezetőin kívül a tőkéseket is érdekelni kezdték a statisztikai felmérések, egyre komolyabb eszközöket használtak fel adataik feldolgozására hasznuk növelése érdekében.

- A XVII. század óta a matematikai statisztika a matematika önálló ágává fejlődött, amelynek fő célja minél megbízhatóbb hasznosítható információt nyerni a felmérési, megfigyelési, mérési adatokból.
- Az 1890-es Egyesült Államokbeli népszámlálásra **Hollerith** feltalálta azt a gépet, amely a statisztikai adatokat lyukkártyák elektromos leolvasásával és rendszerezésével dolgozta fel. A gép gyártására Hollerith céget alapított, amelyből később az IBM jött létre.