Matematikai alapok és valószínűségszámítás

Középértékek és szóródási mutatók

Középértékek

- A leíró statisztikák talán leggyakrabban használt csoportját a középértékek jelentik. Legkönnyebben mint az adathalmaz tipikus, ‘átlagos’ értékeit definiálhatjuk a középértékeket, amiket más kifejezéssel a centrális tendencia mutatóinak is neveznek.
- Centrális tendenciának, vagyis a középértékeknek három fő típusát szokás elkülöníteni:
  - A Móduszt
  - A Mediánt és
  - Az Átlagot

A Módusz

- Az adathalmaz leggyakoribb elemét *Módusz*nak nevezzük. Ez a legáltalánosabban használható középérték, bármely változótípus esetén értelmes, és sokszor érdekes lehet a kérdés, hogy mi a leggyakrabban előforduló, azaz legtipikusabb értéke a mintának.
- Szigorúan tekintve ez a mutató kevéssé méri a centrális tendenciát, sokkal inkább a tipikusságot ragadja meg.
- Ha grafikusan ábrázoljuk az adatainkat a *Módusz* a legmagasabb oszlophoz tartozó értéke a hisztogramnak, vagy az oszlopdiagramnak.
- Pl. közvéleménykutatások esetén érdekes lehet, mert ez reprezentálja a legtipikusabb véleményt.

A Medián

- A Medián kettéosztja az adathalmazt, azaz a centrális tendencia azon mutatója, az adathalmaz azon értéke, amelynél ugyanannyi kisebb, mint nagyobb érték található.
- A Módusszal ellentétben a Medián, bár továbbra is széles körűen alkalmazható, nominális adatok esetén már nem értelmezhető, mivel itt már a mérési skálának legalább a sorba rendezhetőség tulajdonsággal rendelkeznie kell.
- A Mediánt legegyszerűbben úgy találhatjuk meg, ha sorba rendezzük az adatainkat, és megkeressük a középen elhelyezkedő adatot. (Amennyiben páratlan számú adatunk van.) páros számú adat esetén a középső két adat átlagaként határozható meg. Ezt az elvet akkor is követjük, ha egy-egy elem többször fordul elő a mintában.

A Medián

- Nagyszámú minta esetén a százalékos kumulatív gyakorisági táblázatot hívjuk segítségül, ahol is az az érték lesz a középső, amelynél a kumulatív százalék éppen átlépi az 50 %-os küszöböt. Ha a minta egy adott értéknél éppen eléri az 50 %-os határt (kumulatív százalékot), akkor ezen érték és az ezt követő érték átlaga lesz a medián.

A Számtani Átlag

- Ez a legelterjedtebb középérték, nagyon széles körben használt a hétköznapi életben is. Általában ezt értjük rajta, ha valaminek az átlagos értékéről, mértékéről beszélünk.
- Meghatározása: A minta elemeinek összege osztva a minta elemszámával.
- N elemű adathalmaz esetén az X statisztikai változó populációbeli számtani átlaga:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Ahol x_i a populáció i-edik eleme (i = 1,…, N), μ a populáció átlaga.

Az Átlag

- Bár az átlag a legelterjedtebbnek tekinthető szóródási mutató, ez a legkevésbé széles körben (a különböző mérési skálákat tekintve) használható mutatója a centrális tendenciának, mivel az átlag minimum intervallum skálájú változót feltételez.
- Bár az átlag a leginformatívabbnak tekintett középérték, fontos tisztában lenni vele, hogy vannak kevésbé vonzó tulajdonságai is. Talán a legfontosabb negatív tulajdonsága, hogy nagymértékben érzékeny már alig néhány extrém érték jelenlétére is.

A középértékek összefoglalása

Milyen skálatípusnál milyen középérték használható:

	módusz	**medián**	**átlag**
nominális	Igen	Nem	Nem
ordinális	Igen	Igen	Nem
intervallum	Igen	Igen	Igen
arány	Igen	Igen	Igen

Szóródási mutatatók

- A középértékek önmagukban nem elegendőek a minta (vagy a populáció) megfelelő jellemzésére. Az átlagos, tipikus értékek ismerete mellett azt is tudnunk kell az adatok megfelelő leírásához, hogy mennyire szóródnak az értékek a középértékek körül.
- Hogy ennek fontosságát lássuk, nézzünk két példát:
  - Az 1, 99 értékeket tartalmazó kételemű átlaga 50
  - A 47, 48, 49, 50, 51, 52, 53 értékeket tartalmazó hételemű minta átlaga úgyszintén 50
- Tehát mindkét minta ugyanolyan átlaggal jellemezhető, mégis jelentősen különbözik.

A terjedelem

- A terjedelem (a szóródás terjedelme) a minta legkisebb és legnagyobb értékének különbsége

$$R = X_{max} - X_{min}$$

- Könnyen számolható, és hasznos információt kínál arra vonatkozóan, hogy milyen terjedelemben szóródnak az értékek.
- Meghatározásából fakadóan a terjedelem meglehetősen érzékeny a kiugró értékekre (outlierekre), amik nagymértékben növelik a minta terjedelmét.

Az interkvartilis félterjedelem

- Mint a középértékeknél láttuk, megtalálható az az érték, amely éppen kettéosztja a mintát. Ez az érték a medián. Ennek mintájára megtalálható az az érték is, amely ¼ , ¾ arányban osztja fel a mintát, illetve az az érték is, amely ¾ , ¼ arányban osztja fel a mintát. Ezt a három értéket kvartiliseknek nevezzük, melyek 4 egyenlő részre osztják a mintát. A medián egyben a második kvartilis.
- A minta első és harmadik kvartilis közé eső részét interkvartilis terjedelemnek, ennek felét interkvartilis félterjedelemnek (IKF) nevezzük, és az értékek szóródásának jellemzésére használjuk.
- Természetesen, minél nagyobb az IKF, annál nagyobb szóródást mutatnak az értékek.

Átlagos abszolút eltérés

- Ha egy adathalmazban alacsony a variabilitás akkor az adathalmaz értékei egymáshoz közel állnak nagyságuk tekintetében. Egy kevésbé variábilis, kisebb szóródású mintában az értékeket kivonjuk egy tetszőleges konstansból, az így kapott különbségek kisebb változatosságot mutatnak majd, mint ha egy nagy szóródással jellemezhető minta esetén járunk el hasonló módon. Tehát, a változatosság egy adathalmazon belül kifejezhető, ha meghatározzuk az értékek távolságát valamilyen fix ponttól.
- Jó ötletnek tűnhet tehát kiszámolni az adatok átlaguktól, mint fix ponttól való átlagos eltérését. Azonban az átlagtól való átlagos eltérés, $(x_i - \bar{X})/N$, definíció szerint mindig nulla lesz, tehát ez a mutató ebben a formában nem használható.

A Négyzetösszeg

- Jó megoldás lehetne az átlagtól való eltérések abszolút értékét használni, de az abszolút érték kedvezőtlen matematikai tulajdonságai miatt célszerűbb a különbségek négyzetét, illetve ezen négyzetes különbségek összegét venni. Ezt nevezzük négyzetösszegnek.

- Formalizálva:
  $\Sigma(X-\mu)^2$

- Azonban a négyzetösszeg nagysága nem csak az átlagtól való eltérések nagyságától függ, hanem a minta elemszámától is.

A Variancia (Szórásnégyzet)

- Ha a négyzetösszeget elosztjuk a elemszámmal, N-el, akkor egy, az elemszámtól független statisztikát definiáltunk, ami a Variancia, vagy szórásnégyzet néven ismert szóródási mutató, ami tehát a négyzetösszegek átlaga, vagy átlagos négyzetes eltérés.

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

- A variancia lehetővé teszi, hogy összehasonlítsuk különböző adathalmazok változékonyságát. Azonban, mivel a varianciát a különbségek négyzeteinek összegeként kapjuk, a variancia nem tükrözi a nyers adatok mérési egységeit.

A minta varianciája

- A minta varianciájának számítási módja némileg különbözik a populáció varianciájának számításától:

$$s^2 = \frac{\Sigma (x_i - \bar{X})^2}{N - 1}$$

A Szórás

- Ha a változékonyság mértékét ugyanolyan egységekben kívánjuk kifejezni, mint a nyers adatok mérési egységei, akkor egyszerűen vegyük a variancia négyzetgyökét, így képezve a szórást.

$$\sigma = \sqrt{\frac{\Sigma (X - \mu)^2}{N}}$$

- A varianciához hasonlóan a szórás számítása is eltér kicsit, ha nem a populációra, hanem a mintára számoljuk az értékét:

$$s = \sqrt{\frac{\Sigma (x_i - \bar{X})^2}{N - 1}}$$

Statisztikák a populációban és a mintában

- Ahogy fentebb már volt róla szó, minden statisztikának meghatározható a populációbeli és a mintabeli értéke.
- Az egyes statisztikák populációbeli értékeit elméleti értékeknek (elméleti átlag, elméleti variancia, elméleti szórás) míg a megfelelő mintabeli statisztikákat tapasztalati értékeknek nevezzük (tapasztalati átlag, tapasztalati variancia, tapasztalati szórás).
- Mivel a populációbeli statisztika értékeit általában nem ismerjük, ezért az elméleti értékeket a tapasztalati értékekkel becsüljük.