A leíró statisztikák

A leíró statisztikák fogalma, haszna

Gyakori igény az, hogy egy adathalmazt elemei egyenkénti felsorolása helyett néhány jellemző tulajdonságának megadásával jellemezzünk. Ezeket az adatokból viszonylag könnyen kiszámítható paramétereket **leíró statisztikáknak** (vagy ritkán, de pontosabban: leíró statisztikai függvényeknek) nevezzük. Sok ilyen van, két legfontosabb csoportjuk az ún. **elhelyezkedési** és a **szóródást jellemző paraméterek**. Az elhelyezkedési paraméterek azt az értéket igyekeznek megadni, ami körül a mintánk elemei csoportosulnak (ilyen pl. átlag, medián) míg a szóródási paraméterek azt igyekeznek jellemezni, hogy értékeink mennyire szorosan vagy lazán helyezkednek el ekörül a pont körül (pl. szórás). Előfordul, hogy a minta elemeiről nem csak egyfajta adattal rendelkezünk. Kétféle adat esetén, így összetartozó értékpárok jönnek létre (pl. emberek mintájában a testsúly és testmagasság). Az értékpárok közötti összefüggésről adnak információt a **kapcsolatot jellemző paraméterek.**

A legfontosabb leíró statisztikák

Elhelyezkedést	Szóródást	Kapcsolatot
jellemző statisztikák		
átlag	szórás (tapasztalati)	korrelációs együttható (r, r^2)
medián	interkvartilis terjedelem	rangkorreláció

A leíró statisztikák közül azok a legfontosabbak, amelyek a mintánkat adó populáció elméleti eloszlásfüggvényének valamelyik paraméterére adnak jó becslést a mintánkból. A leíró statisztikák gyakorlati alkalmazhatóságának ez az elméleti alapja. Itt csak annyit jegyzünk meg, hogy pl. a mintánkból meghatározott számtani átlag a populáció eloszlásfüggvényének várható értékére ad -> torzítatlan becslést. A mintából számított (ún. tapasztalati) szórás pedig a populáció eloszlásfüggvényét jellemző (ún. elméleti) szórás paraméter becslését adja.
A képet tovább bonyolítja, hogy a statisztikák a minta választásának esetlegessége miatt maguk is valószínűségi változók, melyeknek meghatározható az eloszlásfüggvénye, sőt ennek paraméterei

becsülhetők, éspedig ismét valamilyen statisztikával. Ezt a következő példán illusztrálhatjuk. Nagyon gyakori, hogy összekeverik a mintából számított tapasztalati szórást (SD) az ugyancsak a mintából számítható 'átlag szórása' (SE) nevű paraméterrel. Sokan úgy gondolják, hogy a kettő lényegében ugyanaz, csak éppen az SE kisebb, mint az SD, ezért jobban fest a grafikonokon. Valójában az SE a mintaátlag (mint statisztika) elméleti eloszlásfüggvénye ismeretlen szórásparaméterének a becslése. Azt is mondhatjuk, hogy az SD egyszerű statisztika, az SE pedig egy statisztika statisztikája, tehát egy fokkal bonyolultabb fogalom.

A statisztikák fogalma általában

Matematikailag **statisztikai függvénynek** vagy röviden **statisztikának** neveznek minden olyan (rendszerint skaláris, olykor vektorértékű) függvényt, amelynek értelmezési tartománya a mintatér. (Magyarul statisztika az, ami az adatainkból egy képlettel kiszámítható, vagy más módon meghatározható.) Az említett leíró statisztikákon kívül igen fontosak még a hipotézisvizsgálatoknál használt statisztikák (pl. t, F statisztika).
Hipotézisvizsgálathoz használt statisztikák --> hipotézisvizsgálatok

A leíró statisztikák

A minta elemszáma (mintanagyság)
Ez a legegyszerűbb, s egyben egyik legfontosabb leíró statisztika. Rendszerint n betűvel jelöljük (latin numerus=szám).

Maximum
A legnagyobb előforduló számérték.

Minimum
A legkisebb előforduló számérték.

Mintaterjedelem
A legnagyobb (maximum) és legkisebb (minimum) előforduló számérték különbsége. Akkor használjuk csak, ha hangsúlyozni kívánjuk a mintánkban előforduló extrém értékeket (vagy éppen ellenkezőleg, az igen kicsi szóródást).

Számtani átlag
Az értékek összege, osztva az elemszámmal. A legjobban ismert, leggyakrabban használt paraméter az eloszlás elhelyezkedésének becslésére. Érdemes tudni, hogy erősen érzékeny a mintában esetleg előforduló értékekre. Ilyenkor célszerűbb a medián használata. Ugyancsak félrevezető lehet az átlag erősen ferde eloszlás esetén.

Variancia, tapasztalati szórásnégyzet
Az adatoknak az átlagtól való négyzetes eltéréseinek átlaga (pontosabban az elemszám helyett n-1-gyel szokás osztani a torzítatlan becslés érdekében.). Bár az elméleti statisztikában fontos fogalom, a gyakorlatban helyette az SD használatos.

Szórás, tapasztalati szórás
A variancia négyzetgyöke. Jelölésére az angol kifejezés rövidítését (SD) használjuk. Mint fentebb említettük, nem tévesztendő össze az átlag szórásával. Az SD a legfontosabb, adataink szóródását jellemző paraméter. Fontos tudnunk azonban, hogy értéke függ adataink mértékegységétől, így két adathalmaz szórása csak akkor hasonlítható össze, ha ugyanazt a mértékegységet használtuk. Egységfüggetlen mérőszám viszont a következő statisztika.

Variációs koefficiens (CV)
A szórás százalékos aránya az átlaghoz viszonyítva. Méréskor ez nem más, mint a relatív hiba. Dimenzió nélküli szám, bármely adathalmaz variációs koefficiense összehasonlítható.

Rendezett minta
Az eredeti minta, az előforduló értékek nagysága szerint sorba rendezve. (pl. egy iskolai osztály a tornasorban, ha a tanulók magasságát vizsgáljuk). Önmagában nem használjuk, de több fontos további statisztika meghatározásához nélkülözhetetlen. Ilyenek pl. a következőkben ismertetendő kvantilisek. A rendezett minta és a belőle származtatott további statisztikák értelmezéséhez nem szükséges, hogy adataink numerikusak legyenek, elég, ha ordinális skálán mérhetők.
Kvantilisek
A rendezett mintából tovább származtatott statisztikák összefoglaló

neve, amikor a rendezett mintát több egyenlő részre osztjuk, és a részhatárokon levő mintaelemek értékét tekintjük.

Medián
A medián annak az adatnak a számértéke, amelyik a rendezett minta közepén van (pl. egy iskolai osztályban a magasságértékek mediánja a tornasor közepén álló tanuló magassága). Mint említettük, jó tulajdonsága, hogy sokkal kevésbé érzékeny a kilógó értékekre, mint az átlag, továbbá ferde eloszlások esetén is használhatóbb. Ordinális skála esetén az átlag értelmezhetetlen, míg a medián igen.

Kvartilisek
Az alsó kvartilis (latin quarta pars = negyedrész) a legkisebb és a medián között középen elhelyezkedő adat számértéke a rendezett mintában. (A tornasorban a legkisebb és a középső diák között középen levő tanuló magassága).
A felső kvartilis hasonlóan a medián és a legnagyobb érték között van középen. A kvartilisek az SD-hez hasonlóan az adatok szóródásáról tájékoztatnak, elsősorban ferde eloszlás esetén érdemes őket használni. (A kvartilisek mutatják a ferdeséget, az SD nem).

Percentilisek
Ha elég adatunk van, akkor percentilisek is definiálhatók. Pl. az n%-os (vagy n-edik) percentilis azt jelenti, hogy az adatok n%-a kisebb, mint ez az érték. (Így a medián az 50%-os percentilisnek, az alsó és felső kvartilisek pedig a 25% ill. 75%-os percentilisnek felelnek meg.) A percentiliseknek óriási jelentősége van a 'mit tekintünk normálisnak?' kérdés eldöntésében. Az alsó és felső néhány percentilis közötti részt (2.5% - 97.5% vagy 5% - 95%) szokás normális (referencia) értéknek elfogadni. Akkor szokás pl. egy gyermekről feltételezni, hogy elmaradt a növekedésben, ha magassága (vagy súlya) nem éri el az azonos korú társaira jellemző 5%-os percentilis értéket. A laboratóriumi normálértékeket is a megfelelő percentilisek alapján definiálják.

A percentilisek összessége valójában a tapasztalati eloszlásnak felel meg. Ilyen alapon - ha tetszik - a tapasztalati eloszlásfüggvényt (és az abból származtatott dolgokat, pl. a hisztogramot) is tekinthetjük

statisztikának.

Interkvartilis terjedelem
A felső és alsó kvartilis különbsége. Ugyanakkor használatos, amikor a kvartilis.

Korrelációs együttható
Pearson féle korrelációs együtthatónak is nevezik. Összetartozó értékpárok lineáris kapcsolatát jellemző, dimenzió nélküli szám. Kétféle módon adják meg: r a jele a tulajdonképpeni korrelációs együtthatónak, míg r^2 (az előbbi négyzete) hivatalos megnevezése: coefficient of determination. A tökéletes pozitív lineáris összefüggés esetén r = 1, tökéletes negatív lineáris összefüggés esetén r = -1, míg függetlenség esetén r = 0. A korrelációs együtthatóval kapcsolatban gyakoriak a félreértések. Ezek részletesen olvashatók a "Kapcsolat változók között" című fejezetben. Fontos tudni, hogy a korrelációs együttható értéke erősen függ a kilógó értékektől.

Rang
Ezt a statisztikát úgy kapjuk, ha a rendezett mintában minden elem értékét a rendezésben elfoglalt sorszámával helyettesítünk. Mint a rendezett mintát, ezt a statisztikát sem önmagában használjuk, hanem további statisztikákat származtatunk belőle.

Rangkorreláció
A rangokból számított korrelációs együttható (Spearman féle korrelációs együtthatónak). Akkor használjuk az eredeti Pearson féle korrelációs együttható helyett, ha az adatpárok közül legalább egy nem numerikus, hanem ordinális skálájú, vagy ha az eloszlás nagyon ferde, esetleg kilógó értékek vannak a mintában.