Matematikai alapok és valószínűségszámítás

Statisztikai becslés

Statisztikák eloszlása

Mintavétel

A statisztikában a cél, hogy az érdeklõdés tárgyát képezõ populáció bizonyos paramétereit a populációból vett minta segítségével becsüljük.

A minta akkor jó, ha reprezentatív a populációra nézve, amit a legjobban akkor biztosíthatunk, ha véletlen, random mintát veszünk. Random minta esetén a populáció minden egyes tagjának azonos esélye van a mintába kerülésre.

Bevezetés

Példa:

Tegyük fel, hogy egy adott populációt az első éves pszichológia szakos hallgatók képezik. Egy 10 elemű mintát veszünk ebből a populációból a következõképpen: a neveket felírjuk egy-egy papírra, egy dobozból húzunk 10 nevet.

Ebben az esetben a mintavétel véletlen lesz: minden egyes nevet azonos valószínûséggel húzunk ki. Az egyes események (húzások), egymástól függetlenek.

Mintavétel

Példa:

Tegyük fel, hogy egy adott populációt a telefonkönyvben szereplõ személyek alkotnak. Ebbõl a populációból szeretnénk egy mintát venni oly módon, hogy véletlenszerûen kiválasztunk egy személyt, majd aztán szisztematikusan haladva minden 100. személy választjuk ki a mintába kerülésre.

Véletlenszerû-e a mintavétel ebben az esetben?

Mintavétel

A véletlen mintavétel nagyon fontos lépés a minta reprezentativitásának biztosítására,

DE:

A véletlen mintavétel nem mindíg garantálja, hogy a minta valóban reprezentatív lesz a populációra nézve. Pl.: nõk:férfiak aránya egy populációban 43:57, de ha 100 fõs mintát veszünk, lehet, hogy a mintabeli arány 38:62 lesz, akkor is, ha a mintavétel véletlenszerû volt.

Ha ennek a változónak a reprezentativitása fontos a vizsgálat szempontjából, akkor alkalmazzák a ***rétegzett mintavétel***t, azaz, egy változó mentén felosztják a mintát, jelen esetben pl. véletlenszerûen választanak 43 nőt és véletlenszerűen választanak 57 férfit, akik a 100 fős mintát alkotják.

Statisztikai becslés

A mintavétel csak eszköz, ahhoz, hogy a minta alapján becsülni tudjuk a populáció különbözõ paramétereit, amikre az érdeklõdésünk irányul.

A mintában kiszámolhatjuk az érdeklõdésünk tárgyát képezõ paraméter értéket, amit a mintabeli, vagy tapasztalati értéknek nevezünk, és ez az érték lesz a populációbeli, vagy elméleti érték becslése. Ezeket a minta adataiból számolt értékeket ***statisztiká***nak nevezzük.

A becsült érték a gyakorlatban szinte sohasem fog pontosan megegyezni az populációbeli értékkel, hanem valamennyire szinte mindig eltér attól. Egy adott populációbeli paraméter és a mintabeli paraméter különbsége a becslés hibája.

A becslés akkor jó, ha ***torzítatlan***, azaz egyik irányba sem elfogult, azaz a becslés uo. valószínûséggel lesz kisebb, mint nagyobb, mint a becsülendő paraméter. Azt mondhatjuk tehát, hogy a becslések várható értéke (ha valós populációról van szó a becslések átlaga) megegyezik a paraméter elméleti értékével.

Statisztikák elméleti eloszlása

A populáció paramétereit a mintából számított statisztikával becsülhetjük, például a populáció átlagát a mintából számolt átlaggal. Ez a becslés nagy valószínûséggel hibával terhelt lesz.

Ha a populációból egy másik mintát veszünk, abban is kiszámolhatjuk az adott statisztika (becslés) értékét, így már 2 becslésünk lesz a populáció paraméterre. Ezek átlagát vehetjük, és azt is tekinthetjük a megfelelõ elméleti paraméter becslésének.

Aztán vehetünk még egy mintát, ebben is számolhatjuk a megfelelõ statisztikát, és így tovább…

Statisztikák elméleti eloszlása

Ily módon tehát vehetjük egy adott populáció összes n elemû mintáját és mindegyikben kiszámolhatjuk a megfelelõ statisztika értékét. Ezek az értékek mind a populáció- paraméter becslései lesznek, és ha a becslés torzítatlan, akkor a populáció-paraméter körül fognak ingadozni. Már viszonylag kis számú populációból is nagyon nagyszámú mintát vehetünk, melyekben az adott statisztikák értékei bizonyos eloszlást fognak követni.

Statisztikák elméleti eloszlása

Az átlagot például véve tekintsünk egy populációt, amelyben egy bizonyos

X véletlen változó normál eloszlást követ ***μ*** átlaggal, és ***σ*** szórással (X ~ N(μ, σ)). Ha ebbõl a populációból 30 elemû mintákat veszünk, akkor a minták átlaga normál eloszlást fog követni melynek várható értéke (átlaga) megegyezik az eredeti véletlen változó X átlagával:

$$E(\bar{x}_{30}) = \mu$$

szórása pedig egyenlõ lesz az X szórásának és a mintaelemszám gyökének hányadosával.

$$\sigma_{\bar{x}_{30}} = \frac{\sigma_X}{\sqrt{n}}$$

Azt, hogy a mintaátlagok, a populációátlag becslései mennyire ingadoznak a populációátlag (ami egyenlõ a mintaátlagok elméleti eloszlásának átlagával) körül, a mintaátlagok elméleti eloszlásának szórása fejezi ki.

A becslés standard hibája

A becslések hibájának nagyságát fejezi ki, a ***becslés standard hibá***jának (SE) nevezzük. Tehát a becslés standard hibája megegyezik a becslõ statisztika elméleti eloszlásának szórásával.

$$SE = \frac{\sigma_X}{\sqrt{n}}$$

Ha a populáció szórása nem ismert, akkor a minta szórásával becsülhetjük, így a standard hiba becslése:

$$\hat{S}E = \frac{s_{x_{30}}}{\sqrt{n}}$$

A becslés standard hibája

Mint a standard hiba képletéből is látható, az elemszám növelésével a standard hiba csökken, azaz a becslő statisztikák átlagosan egyre kisebb hibával becslik a populáció átlagát.

A becslés standard hibájával kapcsolatos a becslés két további fontos ismérve, a ***hatékonyság***, ami arra utal, hogy a becslés kis hibával közelíti a becsülendő paramétert, és a ***konzisztencia***, ami arra utal, hogy a mintanagyság növelésével a becslő statisztika elméleti eloszlásának szórása, vagyis standard hibája, egyre kisebb lesz.

Intervallumbecslés

Az eddigiekben tárgyalt becslést pontbecslésnek nevezzük, mert a becsülendõ populáció-paraméter becsléseként egyetlen értéket, egyetlen pontot adunk meg. Láttuk, hogy a minta elemszámának növelésével a becslés egyre pontosabb lesz, de pontosan ritkán fog megegyezni a populáció-paraméterrel.

A statisztikai becslés egy másik típusát az ún. Intervallumbecslés jelenti, amikor nem pontos értéket adunk meg a populációparaméter becsléseként, hanem egy tartományt, intervallumot, amibe a becsülendõ paraméter bizonyos valószínûséggel bele fog esni. Ezt a valószínûséget a becslés megbízhatóságának, konfidenciájának nevezzük, az intervallumot pedig megbízhatósági tartománynak, vagy konfidencia intervallumnak.

Intervallumbecslés

Az átlagok példájánál maradva: láttuk, hogy a mintaátlagok normál eloszlást követnek μ várható értékkel és $\frac{\sigma}{\sqrt{n}}$ szórással.

Ha 95%-os konfidencia intervallumot szeretnénk az átlagra, akkor meg kell adnunk azt a tartományt, ami a populációátlagot 95%-os valószínûséggel tartalmazza. Standard normál eloszlás esetén az értékek 95%-os valószínûséggel a -1.96 és 1.96 szórásnyi tartományba esnek, tehát az átlag esetén a $-1.96 * \frac{\sigma}{\sqrt{n}}, 1.96 * \frac{\sigma}{\sqrt{n}}$ tartomány lesz a 95%-os konfidencia intervallum.