

7. lecke Statisztikai mintavétel és becslés

A valószínűségi változó vizsgálatakor az lenne az ideális, ha végtelen nagy adathalmaz állna rendelkezésünkre, ugyanis ebben az esetben kapnánk csak elég pontos információkat a vizsgált sokaság különböző jellemzőiről. Végtelen sok adattal azonban nem tudunk dolgozni, mert a rendelkezésre álló erőforrások (idő, pénz, kapacitás, stb.) ennek határt szabnak. Így a vizsgált valószínűségi változó véges számú megfigyelt értékeiből álló adathalmazt ismerjük, és ezekből az adatokból kell statisztikai következtetéseket levonni a valószínűségi változó eloszlására jellemző paramétereire.

7.1. Mintavételi eljárások

A statisztikai vizsgálat tárgyát képező egyedek összességét statisztikai sokaságnak/ populációnak nevezzük. A populáció tartalmazhat véges vagy végtelen sok elemet. A statisztikai populációból vizsgálat céljából kiválasztott csoportot mintának nevezzük. A mintavétel különböző eljárásokkal a sokaság n -számú elemének kiválasztásából áll. A kiválasztott elemekhez tartozó számértékeket mintavételi változóknak nevezzük, és úgy tekintjük, mint n -számú független és egyforma eloszlású valószínűségi változót.

A kiválasztott mintának reprezentatívnak kell lennie, azaz ugyanazokkal a tulajdonságokkal kell rendelkeznie, mint az alappopulációnak, hiszen csak így tudunk a minta elemeivel végzett elemzések után következtetéseket levonni az alappopulációra vonatkozóan.

Egy populáció elemszáma lehet véges és végtelen. A véges (N) elemszámú populáció megadásának legegyszerűbb módja egyetlen ismerv szerint az alábbi:

$$Y_1, Y_2, \dots, Y_N$$

Ha a populáció végtelen számú, akkor nem adható meg ebben a formában. Ekkor két esetet különböztetünk meg:

- diszkrét ismerv esetén, ami azt jelenti, hogy az ismervértékek véges vagy megszámlálhatatlan végtelent alkotnak, akkor a valószínűségi eloszlás használható:

$$P(Y=k)=P_k$$

- folytonos ismerv esetén pedig a:

$$P(Y<y)=F(y)$$

eloszlásfüggvénnyel és ennek deriváltjával, az

$$F'(Y)=f(y)$$

sűrűségfüggvénnyel adható meg a sokaság.

Összefoglalva:

Ismerv	Populáció	
	Véges	végtelen
Diszkrét	Felsorolással, valószínűség-eloszlással	Valószínűség-eloszlás
folytonos	Felsorolással, eloszlás függvénnyel	eloszlásfüggvénnyel

A gyakorlatban többnyire véges populációból történik a mintavétel. A populáció milyenségétől függetlenül a belőle származó minta mindig véges, és elemszámát n -nel jelöljük. A minta megadása pedig az elemek felsorolásával történik: $y=(y_1, y_2, \dots, y_n)$.

Fontos kérdés, hogy hogyan válasszunk ki mintát a populációból. Ennek különböző módjai ismertek. A mintavételnél fontos követelmény a pontosság és az olcsóság. Hogy a kettő közül melyiket mennyire vesszük figyelembe, az meghatározza a mintaelemek kiválasztási módját.

A mintaelemek kiválasztása történhet visszatevéssel vagy visszatevés nélkül.

Visszatevéses mintavétel: a kiválasztott elemeket visszahelyezzük a mintába és így ugyanaz az elem többször is bekerülhet a mintába. (Ez a független, azonos eloszlású minta :FAE). Egy N elemszámú populációból n -elemet N^n -féleképpen választhatunk ki.

Visszatevés nélküli mintavétel: a kiválasztott elemet nem tesszük vissza, így minden mintaelem csak egyszer kerülhet a mintába (Egyszerű véletlen mintavétel (EV)). Egy N elemszámú

populációból n -elemet $\binom{N}{n}$ -féleképpen választhatunk ki.

Végtelen elemszámú populáció esetén mindkét eljárásnál a minta elemei, mint valószínűségi változók, minden esetben függetlenek lesznek egymástól. Véges populáció esetén csak a visszatevéses mintavétel eredményez független mintaelemeket.

Véges populáció esetén a minta jellemzője az n/N kiválasztási arány, amely azt mutatja meg, hogy a populáció elemeinek mekkora hányada kerül a mintába.

A mintavétel módja nagymértékben meghatározza a minta tulajdonságát, aminek igazi jelentősége a mintavételi hiba meghatározásánál van.

A mintával kapcsolatban fontos fogalom a kis és nagy minta. Ennek jelentőségét az adja, hogy a mintából számított jellemzők nagy részének (átlag, szórás, stb.) eloszlása nagy minta esetén közelítőleg normális eloszlásúvá válik, így egyszerűbb kezelni. Kis mintaszám esetén ez általában nem mutatható ki. Szimmetrikus vagy ahhoz közel álló populációi eloszlás esetén már viszonylag kis elemszámú minták ($n > 30$) is nagy mintának tekinthetők, míg a szimmetriktól eltérő populáció esetén csak a több százas mintanagyság tekinthető nagy mintának.

1. bemutató feladat:

Kis minta, visszatevéses mintavétellel.

Egy urnában 2 fekete és 8 fehér golyó van, 3 elemű mintát veszünk. Bármilyen golyót választunk ki elsőre, mivel visszatesszük, a 2. és 3. kiválasztáskor is ugyanolyan a populáció állapota, azaz $2/10=1/5$ a valószínűsége, hogy fehér, és $8/10=4/5$ a valószínűsége annak, hogy fekete golyót húzunk. A mintaelemek kiválasztása független egymástól.

Kis minta visszatevés nélküli mintavétellel

1. kiválasztás: $1/5$ a fekete és $4/5$ a fehér kiválasztásának valószínűsége.
2. kiválasztás: ha elsőre feketét vettünk ki, akkor már csak $1/9$ a fekete és $8/9$ a fehér kiválasztásának valószínűsége.
3. kiválasztás: ha másodikra is feketét választottunk ki, akkor már csak fehéret tudunk választani, azaz az első két kiválasztás meghatározza a harmadikat, tehát az egyes mintaelemek nem függetlenek egymástól.

Nagy minta visszatevéses mintavétellel

Legye 200.000 fekete és 800.000fehér golyó az urnában. A fekete kiválasztásának valószínűsége itt is $1/5=0,2$

Nagy minta visszatevés nélküli mintavétellel

- 1.mintavétel: $1/5=0,2$ a fekete valószínűsége
- 2.mintavétel: $199.999/999.999=0,199999999\approx 0,2$ a fekete valószínűsége.

Azaz nagy mintaszám esetén a visszatevéses és a visszatevés nélküli mintavétel is közelítőleg egymástól független és azonos eloszlású elemekből álló mintához vezet.

A minta nagysága és a mintavétel módja mellett fontos a mintavételi eljárás megválasztása. A mintavételi eljárás során az alappopulációból meghatározott számú egyedet választunk ki. A mintavételi eljárások sokféleségét az adja, hogy a minimális ráfordítással, maximális információt elve tartalmazza az olcsóság és fontosság ellentétét. Ezért a mintavételi eljárások mindig kompromisszumot takarnak a rendelkezésre álló pénz és idő, valamint az elérhető pontosság között. A mintavételi eljárások sokfélesége végül is ennek az ésszerű kompromisszumnak az adott vizsgálati célhoz való illesztését jelentik

7.1.1. Véletlen mintavételi eljárások

A véletlen mintavétel lényege, hogy a mintát alkotó elemek a kiválasztás során egyenlő valószínűséggel kerüljenek bele a mintába. Az ember a véletlen kiválasztás végrehajtására nem megfelelő, ugyanis a szubjektív kiválasztás általában nem felel meg az egyenlő valószínűség elvének, főleg nagy populáció és minta esetén.

Ha a populáció minden tagjához egy sorsszámot rendelünk, akkor egy olyan számsort kell megadnunk, amely a véletlenszerűséget biztosítja a mintavételkor. Ilyen számsort háromféleképpen adhatunk meg:

- **Sorsolással:** papírlapokra felírjuk a sorsszámokat és egy urnába téve húzzuk ki a mintába kerülő elemek sorsszámát.
- **Véletlen számok segítségével:** véletlen számok táblázat segítségével: matematikai képlettel állították elő. Először ki kell sorsolni a táblázat valamely sorát és oszlopát, és az ott található számtól kezdve folyamatosan haladva használjuk a mintavételhez a táblázatban szereplő sorsszámokat. Az egymás mellett véletlenszerűen sorakozó négy (vagy több) jegyű számok közül azokat jegyezzük fel, melyek sorszámként szerepelnek nyilvántartásunkban, a sorszámként nem szereplő számokat pedig "átugorjuk". Ha négyjegyű táblát használunk, akkor az adataink sorszámozása is négyjegyű, azaz 0001; 0002; 0003; stb. Ezt mindaddig folytatjuk, amíg annyi sorsszámot nem jegyeztünk fel, ahány elemű mintára szükségünk van. A véletlenszám-tábla a legtöbb statisztikai könyvben megtalálható.
- **számítógépes véletlen szám generálással.**

A véletlen szám segítségével történő mintavétel csak olyan esetekben használható, ha a populáció egyenletes eloszlású.

Folytonos populáció esetében a felezési módszert alkalmazzák, addig felezik a mintát, amíg vizsgálható méretű mintanagyságot kapnak.

7.1.1.1. Egyszerű véletlen mintavétel (EV)

Homogén, véges elemszámú populáció esetén visszatevés nélkül választjuk ki a mintát, elemenként egyenlő valószínűséggel. Véletlen szám segítségével történő mintavételkor az ismételten előforduló sorsszámot átugorjuk. A mintavétel során az N elemű populációból $\binom{N}{n}$ -féle (N alatt az n) különböző összetételű mintát kapunk.

Ez a módszer főleg a természettudományi kísérleteknél, főleg a biológiai eredmények értékelésekor alkalmazható. Társadalmi-gazdasági jelenségek vizsgálatára nem használható.

7.1.1.2. Független, azonos eloszlású minta (FAE)

FAE mintát akkor kapunk, ha homogén és végtelen populációból visszatevéssel veszünk mintát. Végtelen populációból vett visszatevés nélküli minta is lehet FAE-minta, hiszen a kiválasztott elemek nem befolyásolják a megmaradó populáció eloszlását. Az egyes mintaelemek kiválasztása azonos valószínűséggel történik. Tipikus alkalmazási területe a tömegtermelés minőségi ellenőrzésének.

7.1.1.3. Rétegzett mintavétel (R)

A rétegzett mintavétel során a vizsgált ismerv szempontjából heterogén populációt több homogén (minél kisebb szórású) részpulációra bontjuk úgy, hogy a csoportok kiadják a teljes populációt, továbbá egyetlen populációi elem se tartozzon két vagy több csoportba. Az egyes rétegeken belül a minta elemének a kiválasztása egyszerű véletlen mintavétellel történik. Ez a módszer a társadalomtudományok területén nagyon gyakori.

- **Egyenletes elosztás:** Lényege, hogy minden rétegben azonos számú mintaelem kerül, azaz $n_j = n/M$. Egyszerű végrehajtani, de hátránya, hogy nem veszi figyelembe a teljes populációt adó részpulációk nagyságát és szórását, így nagyfokú torzítást okozhat
- **Arányos elosztás:** Lényege, hogy a mintába a populációi arányoknak megfelelően választjuk meg az arányszámot: $n_j = n \cdot N_j/N$. A mintában ugyanazok a súlyarányok szerepelnek, mint a populációban. Végrehajtása egyszerű.
- **Nem arányos elosztás:** a mintában a rétegarányok nem egyeznek meg az alappopulációi rétegarányokkal: $n_j \neq n \cdot N_j/N$
- **Neyman-féle optimális elosztás:** A nem arányos elosztás egyik fajtája. Alkalmazásának feltétele, hogy előre ismerjük vagy becsülni tudjuk az egyes rétegekbe lévő adatok szórását (σ_i). A nagyobb szórású rétegekből nagyobb számú mintát veszünk. Előnye, hogy minimális hibával számolható ki az ilyen mintából a főátlag, de nehéz végrehajtani, mivel a rétegenkénti szórás szükséges hozzá.
- **Költségoptimális elosztás:** A rétegzett mintavétel egyik módja, és feltételezi, hogy ismerjük az egyes rétegek megfigyelési egységköltségeit is. Egy elem átlagos megfigyelési költsége Π_j forint. Azonos rétegnagyság és szórás esetén minél nagyobb a mintavétel költsége, annál kisebb mintát kell venni a rétegből

7.1.1.4. Csoportos mintavétel (Cs)

A homogén populációt csoportokra bontjuk, a mintát egyszerű véletlen mintavétellel kiválasztott csoport egyedei alkotják. A csoportok meghatározása lehet természetes, azaz eleve adott, de mesterségesen is történhet. Az N elemű populációt M részre bontjuk, ahol az egyes csoportok n_i eleműek, és $\sum n_i = N$.

Főleg közvélemény-kutatáskor alkalmazzák ezt a mintavételt.

7.1.1.5. Többlépcsős mintavétel (TL)

Homogén populáció vizsgálata esetén alkalmazható. Először csoportos mintavétellel kiválasztjuk az elsődleges mintavételi egységeket. 1-1 homogén nagyobb csoportból egyszerű véletlen kiválasztással egyedeket jelölünk ki, amelyek kiscsoportokat alkotnak. Ha megfelelő a minta, akkor kétlépcsős mintavételről beszélünk. Ha nem, akkor a kiscsoportokkal tovább ismételjük az eljárást.

7.1.2. Nem véletlen mintavételi eljárások

A véletlen mintavétel esetén elkövetett hibák valószínűség-számítási ismeretek segítségével meghatározhatóak. A nem véletlen mintavétel esetén kapott minta és az eredeti populáció között azonban nehéz a mintavétel során elkövetett hibákat számszerűsíteni, a torzításokat kiszűrni.

A torzítások csak csökkenthetők, de nem szűrhetők ki teljesen. A torzítások minimalizálása érdekében célszerű, ha

- a vizsgálat alanya nem ismeri az adatfelvétel célját,
- egyértelmű kérdések vannak megfogalmazva,
- kontrollkérdéseket is beiktatnak a kérdések közé.

A nem véletlen mintavételi eljárások tipikus esetei a társadalmi vizsgálatok, sok esetben személyes megkérdezés során alakul ki az információ.

7.1.2.1. Szubjektív kiválasztás

Önkényesnek is nevezik, mivel a mintavevő a szakmai ismeretére támaszkodva az általa jellemzőnek tartott egyedeket választja ki a populációból

7.1.2.2. Kvóta szerinti kiválasztás

Előre megadjuk a minta összetételét, azaz előre rögzített megoszlási viszonyzámnak megfelelő lesz a minta. Ehhez megfelelő információ szükséges a populációról a vizsgált ismérv szerint. A véletlennel kombinált kvóta kiválasztás azonban jobb, mint a csak kvóta szerinti. A kvótás eljárás a rétegzett mintavételhez hasonló eredményt ad. A lakosság körében végzett felmérések, az adatvédelem miatt egyre inkább kvótás eljárással készülnek.

7.1.2.3. Koncentrált kiválasztás

Feltételezi, hogy a populáció vizsgált jellemzőjét döntően kevés számú egyed határozza meg. A mintavétel során ezeket, a meghatározó elemeket választjuk ki. Pl.: a nemzetgazdasági fogyasztással kapcsolatos elemzések során a fogyasztói árindex meghatározásánál a legnagyobb mértékben fogyasztott termékeket választják ki.

7.1.3. Kombinált kiválasztás

A véletlen és a tudatos kiválasztás kombinációja. Az N elemű véges populációt a vizsgálandó ismerv alapján sorba rendezzük. Az n mintaelem-szám megadása után a populáció minden k -adik eleme bekerül a mintába olyan módon, hogy

$$k = \left[\frac{N}{n} \right] \text{ (a hányados egész része)}$$

Összefoglalásul elmondható, hogy a mintavétel alapvető célja, hogy létrehozzon az alappopuláció helyett egy kevesebb költséggel és idővel vizsgálható részpopulációt, amely minta elegendő információt nyújt arra, hogy belőle az eredeti populációra levont következtetéseink valószínűségi értelemben kellően pontosak legyenek.

7.1.4. Mintavételi hiba

A statisztikai adatfelvételek és az annak eredményeit felhasználó elemzések mindig tartalmaznak hibát. A statisztikai hiba egy része a módszertan sajátosságaiból is adódik (tömörítés, közelítés, becslés), ez velejárója a statisztikai elemzéseknek. A statisztikus célja, hogy a hibát minimálisra csökkentse (mintavételi és nem mintavételi hibát együtt). A mintavételi hiba matematikai-statisztikai eszközökkel becsülhető. A nem mintavételi hiba korábbi tapasztalatok alapján becsülhető meg. A mintavétel tervezésénél a mintavételi hibával és annak vizsgálatával foglalkozunk.

Egy adott populáció esetén egy meghatározott számú mintát nemcsak egyféleképpen lehet kiválasztani, így minden minta más és más összetételű.

Vegyünk egy példát, amelynél ismerjük a teljes populációt, de a gyakorlatban ez ritkán valósul meg, hisz a mintavételre pont azért van szükség, mert a teljes alappopulációt nem ismerjük.

2. bemutató feladat:

A Ferihegyre érkező külföldi légitársaságok adatai:

Sorszám	Utasok száma 1000 fő
1	42,5
2	31,6
3	34,9
4	32,0
5	72,6
6	48,8
7	21,3
8	57,4
9	110,5
10	17,4
Átlag	46,9
szórás	26,4

Megoldás:

Vegyünk most 2, 3 és 5 elemű mintákat, és számítsuk ki az átlagokat és a szórásokat mintánként. Az elemeket az alábbiak szerint választhatjuk ki:

Mintaelemek	átlag	Mintaelemek	átlag	Mintaelemek	átlag
1,2	37,05	1,2,3	36,33	1,2,3,4,5,	42,72
2,3	33,25	2,3,4	32,83	2,3,4,5,6	43,98
3,4	33,45	3,4,5	46,50	3,4,5,6,7	41,92
4,5	52,30	4,5,6	51,13	4,5,6,7,8	46,42
5,6	60,70	5,6,7	47,57	5,6,7,8,9	62,12
6,7	35,05	6,7,8	42,50	6,7,8,9,10	51,8
7,8	39,35	7,8,9	63,07		
8,9	83,95	8,9,10	61,77		
9,10	63,95				
Átlag	48,78		47,47		48,04
szórás	16,74		10,20		6,98

Látható, hogy az egyes mutatók értéke mintánként változó, a mintákból számított átlagok az alappopuláció átlaga körül szóródnak. Ez a szóródás a nagyobb minták esetén kisebb, azaz a nagyobb mintákból számított átlagok pontosabban jelzik a populációt. Bármelyik mintából számított átlaggal jellemezzük a populációt, hibát követünk el. A mintavételi hiba a populáció jellegén az alkalmazott mintavételi eljárástól és alapvetően a mintanagyságtól is függ.

7.2. Statisztikai becslés

A statisztikai becslés az alappopulációt alkotó valószínűségi változók eloszlásának, jellemzőinek és paramétereinek becslését jelenti az alappopulációból vett mintából számított mutatók alapján. A statisztikai becsléseket úgynevezett becslőfüggvények segítségével végezzük el. A becslőfüggvény olyan valószínűségi-változó függvény, ami valamely populációi jellemző mintából történő közelítő meghatározására szolgál.

Egy populációi jellemzőre több becslőfüggvény is készíthető. Ahogy a véletlen minta elemei valószínűségi változók, ugyanúgy a becslőfüggvény értéke is az. Egy adott n-elemű minta csak egyetlen becsléssel rendelkezik. A minta alapján az alappopulációnak többféle jellemzője is becsülhető, pl.:

- számtani átlag,
- értékösszeg,
- arány és megoszlás,
- hányados.

A becslőfüggvény akkor lesz jó, ha a különféle (véletlen) minták esetén értéke a becsléni kívánt jellemző körül ingadozik, és az ingadozás lehetőleg kicsi. A becslőfüggvénnyel szemben támasztott követelmények:

- torzítatlanság,
- hatásosság,
- koncisztencia
- robosztusság.

Torzítatlanság:

Torzítatlannak nevezünk egy becslőfüggvényt, ha annak várható értéke megegyezik a becsléni kívánt populációi jellemzővel. Azaz:

$$E(\hat{\Theta}) = \Theta$$

Két becslőfüggvény közül, ha más kritériumot nem veszünk figyelembe, a torzítatlant részesítjük előnybe. A torzítás mértékét a torzítás mérőszámával (B_s) lehet kifejezni. Torzítatlan becslőfüggvény esetén:

$$B_s(\hat{\Theta}) = E(\hat{\Theta}) = \Theta$$

Két torzított becslőfüggvény esetén azt tekintjük a jobbnak, amelyiknél a torzítás abszolút értéke kisebb, azaz az alábbi esetben:

$$|B_s(\hat{\Theta}_1)| < |B_s(\hat{\Theta}_2)|$$

az első becslőfüggvényt választjuk.

Hatásosság

Egy torzítatlan becslőfüggvénynek lehet olyan nagy a szóródása, hogy ez használhatatlanná teszi a becslésre. Ha a $\hat{\Theta}_1$ és a $\hat{\Theta}_2$ torzítatlan becslőfüggvénye θ -nak, és a $\sigma^2(\hat{\Theta}_1) < \sigma^2(\hat{\Theta}_2)$,

akkor azt mondjuk, hogy a $\hat{\Theta}_1$ hatásosabb becslőfüggvénye θ -nak, mint a $\hat{\Theta}_2$. A becslőfüggvény valamennyi lehetséges mintán felvett értékeiből számított szórásnégyzetet mintavételi szórásnégyzetnek, ennek négyzetgyökét pedig a becslőfüggvény, illetve a becslés standard hibájának nevezzük. $Se(\hat{\Theta}) = \sqrt{Var(\hat{\Theta})}$.

Két torzítatlan becslőfüggvény szórásnégyzetét hányados formában összehasonlítva relatív

hatásfoknak nevezzük. $Ef_r = \frac{Var(\hat{\Theta}_1)}{Var(\hat{\Theta}_2)}$. Ha értéke nagyobb mint 1, akkor a $\hat{\Theta}_2$ a hatásosabb.

Konzisztencia:

Követelménye azt írja elő, hogy a becslés torzítatlan legyen, és a mintanagyság minden határon túl történő növelése esetén annak a valószínűsége, hogy a becslési kívánt paraméter és a becslőfüggvény eltérése kisebb egy ε számnál egy legyen vagy a szórásnégyzete a nullához tartson.

$$\lim_{n \rightarrow \infty} P(|\hat{\Theta} - \Theta| < \varepsilon) = 1 \text{ vagy } \lim_{n \rightarrow \infty} VAR(\hat{\Theta}(n)) = 0$$

Más szóval: nagy minta esetén a becslési érték nagy valószínűséggel közelítse meg a populációi jellemző értékét.

Robosztusság:

Akkor mondjuk, hogy egy becslőfüggvény robusztus, ha az érzéketlen a kiinduló feltételekre. Ha a populációi eloszlást nem ismerjük, akkor a becslésre a robusztus becslőfüggvényt használjuk.

7.2.1. Pontbecslés

A becslőfüggvény értékéről közismert, hogy valószínűségi változó, de egy n elemű mintához csak egyetlen konkrét értéket rendelhetünk. Az eddig tanult populációi jellemzők (átlag, szórással) becslését is elvégezhetjük pontbecsléssel. A számítás megegyezik például az átlag esetén a korábban a statisztikában tanultakkal.

Pontbecslés: A minta megfelelő statisztikáját elfogadjuk a populáció megfelelő paraméterének.

Az átlagról azonban tudjuk, hogy nem a populáció minden tagjára vonatkozó adatok ismeretében számoltuk ki, hanem a mintából így bizonytalanságot rejt magába, ezért a pontbecslés helyett a leggyakrabban az ún. intervallumbecslést alkalmazzuk.

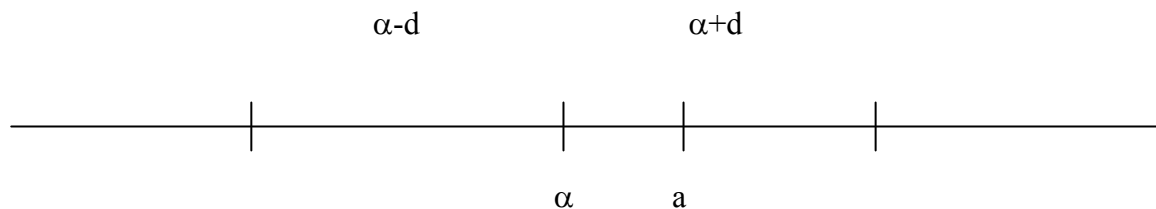
7.2.2. Intervallumbecslés

Jelölje $F(x)$ egy ξ valószínűségi változó eloszlásfüggvényét, „a” pedig a ξ valószínűségi változó eloszlásának egy jellemző paraméterét, amelynek értékét nem ismerjük. Egy n -elemű mintát veszünk; x_1, x_2, \dots, x_n -értékekkel és ebből a mintából következtetünk az a -paraméterre. A mintavételi változókból létrehozunk egy $\alpha = \alpha(x_1, x_2, \dots, x_n)$ függvényt, amelynek értéke ingadozik mintáról mintára, de következtetni lehet belőle a -értékére. Azt mondjuk, hogy α az a -paraméter statisztikai becslése. Az α -becslést torzítatlannak mondjuk, ha várható értéke a -val megegyezik. Az α általában nem egyezik meg a -val. Gyakorlatilag elegendő annyit tudnunk, hogy a valódi a -paraméter benne van egy intervallumban, amelyet a minta alapján határozunk meg. A becslésnek ezt a módját intervallumbecslésnek nevezzük, az intervallumot pedig megbízhatósági vagy konfidencia intervallumnak.

Megadunk egy α közepű intervallumot, amelybe a -értéke egyhez közelálló valószínűséggel beleesik, azaz:

$$P(\alpha-d \leq a \leq \alpha+d) = 1-p.$$

Ahol d : az intervallum szélessége, $(1-p)$ pedig azt a valószínűséget jelenti, amellyel a megadott intervallum lefedi az a -paramétert



Az $(1-p)$ -t százalékban szoktuk megadni. Leggyakrabban használt értékei: 90%; 95%; és 99%. A p a hibaszázalékot jelenti. Az intervallum szélessége a p -értéktől, a minta elemszámától és szórásától. függ

Általában szimmetrikus intervallumokat keresünk. Az intervallumbecslésnél is több esetet különböztetünk meg.

7.2.2.1. Intervallumbecslés a várható értékre: Normális eloszlású populáció esetén, ha a populáció szórása (σ) ismert

Legyen ξ valószínűségi változó normális eloszlású $M(\xi)$ várható értékkel és σ szórással. Tegyük fel, hogy σ -értékét ismerjük, de $M(\xi)$ -értéke ismeretlen. Vegyünk egy n -elemű mintát és becsüljük meg $M(\xi)$ -paramétert a minta átlagával.

A becsléshez a z -próbafüggvényt alkalmazzuk.

Keressük azt a $(z_1; z_2)$ intervallumot, melybe nagy, $(1-\alpha)$ valószínűséggel beleesik a valószínűségi változó várható értéke.

A központi határeloszlás tétel szerint független, azonos eloszlású valószínűségi változók összegének standardizáltja közelítőleg standard normális eloszlású.

Mivel $M(\xi)$ -re (várható értékre) keressük az intervallumot, ezért a konfidencia intervallum az alábbi lesz:

$$\hat{\mu} - z_p * \frac{\sigma}{\sqrt{n}} \leq M(\xi) \leq \hat{\mu} + z_p * \frac{\sigma}{\sqrt{n}}; .$$

$\hat{\mu}$: a mintából meghatározott várható érték

Konkrét minta esetén:

$$\bar{x} - z_p * \frac{\sigma}{\sqrt{n}} \leq M(\xi) \leq \bar{x} + z_p * \frac{\sigma}{\sqrt{n}}; .$$

Először ki kell számolni a standard hibát:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

A z_p értékét „A standard normális eloszlás eloszlásfüggvényeinek értékei” című táblázatból keressük ki a megfelelő valószínűségi szinten (p).

Az intervallumbecslés általában kétoldali, mivel a becslt érték köré szimmetrikus intervallumot szerkesztünk, azaz az intervallum alsó és felső határát pontosan meghatározzuk. Ebben az esetben a táblázatból nem az $(1-\alpha)$ valószínűségi szinthez tartozó értéket keressük, hanem az $(1-\alpha/2)$ -hez tartozót.

Ha $(1-\alpha)=95\%=0,95$, akkor $\alpha=0,05$, tehát $\alpha/2=0,025$, így $(1-\alpha/2)=0,975$. A „ z ”-értékek táblázatból a 0,975-höz tartozó értéket keressük ki, amely 1,96, azaz $z_{0,975}=1,96$.

Egyoldali intervallum esetén, az $(1-\alpha)$ valószínűségi szinthez tartozó értéket keressük, az intervallumnak csak a felső határát tudjuk megállapítani, az alsó határ a negatív végtelen lesz. A „ z ”-értékek táblázatból a 0,95-höz tartozó értéket keressük ki, amely 1,65, azaz $z_{0,95}=1,65$.

A $z_p * \frac{\sigma}{\sqrt{n}}$ mennyiséget hibahatárnak vagy maximális hibának nevezzük, és Δ -val jelöljük.

$$\Delta = z_p * \frac{\sigma}{\sqrt{n}}$$

A várható érték tehát:

$$\bar{x} \pm \Delta$$

A becslési hibahatárt többféleképpen csökkenthetjük:

- Csökkentjük a standard hibát: a standard hiba, a mintaátlag szórása csak a minta elemszámától függ, mégpedig annak gyökével fordítottan arányos. Így az elemszám növelésével csökken a standard hiba, és ezáltal a konfidencia intervallum is.
- Csökkentjük a megbízhatósági szintet, így azaz z_p -értéke is kisebb lesz és szintén kisebb lesz a hibahatár.

3. bemutató feladat:

Egy $\sigma^2 = 0,81$ varianciájú normál eloszlású populációból a vizsgálat céljából vett 50 elemű minta átlaga 6,2. 95%-os megbízhatósági szinten állapítsuk meg a populáció átlagának konfidencia intervallumát!

a) Először ki kell számolni a standard hibát:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0,9}{\sqrt{50}} = 0,13$$

A „z”-értékek táblázatból a 0,975-höz tartozó értéket keressük ki:

$$z_{0,975}=1,96$$

$$\Delta = z_p * \sigma_{\bar{x}} = 1,96 * 0,13 = 0,25$$

(6,2 - 0,25; 6,2 + 0,25) A populáció átlaga 5,95 és 6,45 között van.

b/ Nézzük meg, ha a minta elemszáma 100 lett volna, hogyan alakul a konfidencia intervallum.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0,9}{\sqrt{100}} = 0,09 \quad \Delta = z_p * \sigma_{\bar{x}} = 1,96 * 0,09 = 0,1764 = 0,18$$

$$(6,2 - 0,18 ; 6,2 + 0,18)$$

A populáció átlaga 6,02 és 6,38.

c/ Legyen a megbízhatósági szint 90%, az elemszám az eredetei, azaz n=50

$$z_{0,95}=1,65$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0,9}{\sqrt{50}} = 0,13 \quad \Delta = z_p * \sigma_{\bar{x}} = 1,65 * 0,13 = 0,2145 = 0,21$$

(6,2 - 0,21; 6,2 + 0,21) A populáció átlaga 5,99 és 6,41 között van.

4. bemutató feladat:

Legyen ξ egy ismeretlen várható értékű és ismert $\sigma = 3$ szórású valószínűségi változó. A várható értéket egy 56 elemű mintából becsüljük, a minta átlaga 18,3. Határozzuk meg a várható érték mintaátlaggal történő becslésének 98%-os megbízhatósági szintű konfidencia intervallumát!

Megoldás:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{56}} = 0,4$$

A „z”-értékek táblázatból a 0,99-hez tartozó értéket keressük ki:

$$z_{0,975}=2,33$$

$$\Delta = z_p * \sigma_{\bar{x}} = 2,33 * 0,4 = 0,932$$

A várható érték tehát $18,3 \pm 0,932$, azaz $17,368 \leq M(\xi) \leq 19,232$

7.2.2.2. Intervallumbecslés a várható értékre: Normális eloszlású populáció esetén, ha a populáció szórása nem ismert

Illetve nem normális, de ismert eloszlású populáció esetén, ha nagy mintát vettünk.

Legyen ξ valószínűségi változó normális eloszlású, $M(\xi)$ várható értékkel és σ szórással. Tegyük fel, hogy σ -értéke és $M(\xi)$ -értéke ismeretlen. Vegyünk egy n -elemű mintát és becsüljük meg $M(\xi)$ -paramétert a minta átlagával.

Mivel az alappopuláció szórása nem ismert, így azt is becsülni kell, azaz a mintából számítjuk ki. Ilyenkor z -valószínűségi változó helyett t -valószínűségi változót (Student-féle eloszlás) használjuk.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

A t -eloszlású valószínűségi változó szabadságfoka : $szf=n-1$.

Először ki kell számolni a standard hibát:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

A hibahatár: $\Delta = t_p^{(szf)} * \frac{s}{\sqrt{n}},$

az intervallum pedig: $(\bar{x} - t_p^{(szf)} * \frac{s}{\sqrt{n}}; \bar{x} + t_p^{(szf)} * \frac{s}{\sqrt{n}})$

A t -eloszlás is szimmetrikus eloszlás, azaz a z -eloszláshoz hasonlóan alakul az egy- és kétoldali intervallum. A szabadságfok növelésével a t -eloszlás egyre inkább közelít a normális eloszláshoz, száznál nagyobb mintaszám esetén a két eloszlás eltérése minimális.

5. bemutató feladat:

Egy konzervüzemben az egyik műszakban elkészült konzervek töltési súlyának átlagát szeretnénk meghatározni. 500 minta alapján az átlagos töltőtömeg 497 g, a töltőtömeg szórása 19,49g. 95%-os valószínűséggel mennyi a konzervek töltősúlya?

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{19,49}{\sqrt{500}} = 0,8716$$

$$t_{0,975}^{(499)} = 1,96$$

$$\Delta = t_p * s_{\bar{x}} = 1,96 * 0,8716 = 1,7083 \approx 1,71$$

A konfidencia intervalluma $(497 - 1,71; 497 + 1,71)$ g, azaz 495,29 és 498,71g között várható a konzervek súlya.

6. bemutató feladat:

Egy ebédszállító cégnél 100 napig figyelve az adagokat azt kapjuk, hogy az egytálételek tömegének átlaga 43,2 dkg, 2 dkg-os szórással. Adjunk meg 98%-os megbízhatósági szinten a konfidencia intervallumot az étel tömegének várható értékére!

Megoldás:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0,2$$

$$t_{0,99}^{(99)} = 2,36$$

$$\Delta = t_p \cdot s_{\bar{x}} = 2,36 \cdot 0,2 = 0,472$$

A konfidencia intervalluma $(43,2 - 0,472; 43,2 + 0,472)$ g, azaz 42,728 és 43,672g között van az étel tömegének várható értékére.

7.2.2.3. Normális eloszlású populáció szórásnégyzetének és szórásának konfidencia intervalluma

A populációi szórásnégyzet (σ^2) becslésére a torzítatlan becslést eredményező korrigált tapasztalati szórásnégyzetet (s^2) használjuk.

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

Ha Y normális eloszlású, akkor bizonyítható, hogy az $\chi^2 = \frac{(n-1) \cdot s^2}{\sigma^2}$ változó (n-1) szabadságfokú χ^2 -eloszlást követ. A χ^2 -eloszlás aszimmetrikus, így a konfidencia intervallum az alábbi:

$$P(\chi_{\alpha/2}^2(szf) < \frac{(n-1) \cdot s^2}{\sigma^2} < \chi_{1-\alpha/2}^2(szf)) = 1 - \alpha$$

Átrendezve:

$$P\left(\frac{(n-1) \cdot s^2}{\chi_{1-\alpha/2}^2(szf)} < \sigma^2 < \frac{(n-1) \cdot s^2}{\chi_{\alpha/2}^2(szf)}\right) = 1 - \alpha$$

$\chi_{1-\alpha/2}^2(szf)$ és a $\chi_{\alpha/2}^2(szf)$ értékét táblázatból kell kikeresni a megfelelő szabadságfoknál és valószínűségi szintnél.

Konkrét minta esetén a szórás intervalluma a következő:

$$\sqrt{\frac{(n-1) \cdot s^2}{\chi_{1-\alpha/2}^2(szf)}} < \sigma < \sqrt{\frac{(n-1) \cdot s^2}{\chi_{\alpha/2}^2(szf)}}$$

7. bemutató feladat:

250 g-os kávé csomagoló gép működését vizsgálva 100 elemű mintát veszünk. A töltési tömeg normális eloszlású. Határozzuk meg, hogy milyen határok között lesz a kávécsomagok töltési tömegének szórása 95%-os valószínűségen. Az átlag töltési súly 248g, a minta szórása 5,53g.

Megoldás:

$$Szf=100-1=99; \quad \alpha=0,05, \quad \alpha/2=0,025, \quad 1-\alpha/2=0,975,$$

$$\chi_{0,025(99)}^2=74,2; \quad \chi_{0,975(99)}^2=129,6$$

$$\Delta_1 = \sqrt{\frac{(n-1) * s^2}{\chi_{0,975}^2}} = \sqrt{\frac{99 * 5,53^2}{129,6}} = 4,83g$$

$$\Delta_2 = \sqrt{\frac{(n-1) * s^2}{\chi_{0,025}^2}} = \sqrt{\frac{99 * 5,53^2}{74,2}} = 6,39g$$

A nettó töltési tömeg szóródása (ingadozása) 95%-os valószínűségi szinten 4,83g és 6,39g között van.

7.2.2.4. Adott intervallumszélességhez tartozó elemszám illetve valószínűségi szint meghatározása

Eddig adott elemszám és valószínűségi szint mellett határoztuk meg a konfidencia intervallumot. A becsléskor azonban előre rögzíthetjük a hibahatárt és ehhez kell meghatározni a szükséges minta elemszámát adott valószínűségi szinten. Ha az elemszám és a hibahatár is adott, akkor meg tudjuk mondani, hogy hány százalékos valószínűséggel kerül a populációi jellemző az adott elemszám esetén az előre meghatározott intervallumba.

Elemszám meghatározása: adott az intervallum és a valószínűség.

Mivel: $\pm \Delta = z_p * \frac{\sigma}{\sqrt{n}}$ illetve $\pm \Delta = t_p^{(szf)} * \frac{s}{\sqrt{n}}$, átalakítás után az alábbiakat kapjuk:

$$n = \left(\frac{z_p * \sigma}{\Delta} \right)^2 \quad \text{illetve} \quad n = \left(\frac{t_p^{(szf)} * s}{\Delta} \right)^2$$

Valószínűségi szint meghatározása:

$\pm \Delta = z_p * \frac{\sigma}{\sqrt{n}}$ és a $\pm \Delta = t_p^{(szf)} * \frac{s}{\sqrt{n}}$ képletek az átalakítás után az alábbiak:

$$z_p = \frac{\Delta * \sqrt{n}}{\sigma} \quad \text{illetve} \quad t_p^{(szf)} = \frac{\Delta * \sqrt{n}}{s}.$$

A kiszámított z_p -értékhez tartozó táblázatbeli $\Phi(z)$ -értékből tudjuk a valószínűséget kiszámítani:

$$P = \Psi(z) = \Phi(z) - (1 - \Phi(z)),$$

ahol $P = \Psi(z)$: a keresett valószínűség,

$\Phi(z)$: a kiszámított z -értékhez tartozó táblázatbeli érték.

A t-eloszlás esetében a kiszámított értéket megkeressük a táblázatban az adott szabadságfoknál, és leolvassuk a hozzá tartozó valószínűséget.

8. bemutató feladat:

Egy vizsgálat során, megállapították, hogy a hallgatók átlagos testmagassága 169 cm és 173 cm között van 95%-os valószínűséggel.

A populáció szórása 9,99 cm.

- Hány hallgatót vontak be a vizsgálatba?
- Mekkora valószínűségi szinten végezték a vizsgálatot, ha 200 hallgató esetében is ugyanezt a konfidencia intervallumot kapták?

Megoldás:

a.)

$$\Delta = \frac{173 - 169}{2} = 2$$

$$n = \left(\frac{z_p * \sigma}{\Delta} \right)^2 = \left(\frac{1,96 * 9,99}{2} \right)^2 = 95,84, \text{ azaz } 96 \text{ hallgató testmagasságát mérték le.}$$

b.)

$$z_p = \frac{\Delta * \sqrt{n}}{\sigma} = \frac{2 * \sqrt{200}}{9,99} = 2,83$$

A táblázatbeli $\Phi(z)=0,9977$

$$P=0,9977-(1-0,9977)=0,9977-0,0023=0,9954=99,54\%.$$

Tehát a vizsgálatot 99,54%-os megbízhatósági szinten végezték.

7.2.2.5. Értékösszegsor becslése

Nagy populáció esetén általában nem az átlagra, hanem a sokasági értékösszegre keresünk konfidencia intervallumot (mint az előző bemutató feladat végén). A feladat visszavezethető az átlagbecslésre.

Ha a várható érték: $\mu = \bar{x}$, akkor a sokasági értékösszeg

$$S_i = N * \bar{x}$$

A konfidencia intervallum alsó és felső határa a várhatóértékre meghatározzak N-szerese lesz.

$$N * (\bar{x} - z_p * \frac{\sigma}{\sqrt{n}}); N * (\bar{x} + z_p * \frac{\sigma}{\sqrt{n}});$$

Illetve:

$$N * (\bar{x} - t_p^{(szf)} * \frac{s}{\sqrt{n}}); N * (\bar{x} + t_p^{(szf)} * \frac{s}{\sqrt{n}})$$

9. bemutató feladat:

50.000 üvegből vett 500 minta alapján az átlagos töltőtömeg 497 g, a töltőtömeg szórása 19,49g. 95%-os valószínűséggel mennyi lesz az 50000 üveg töltősúlya?

$$s_x = \frac{s}{\sqrt{n}} = \frac{19,49}{\sqrt{500}} = 0,8716$$

$$t_{0,975}^{(499)} = 1,96$$

$$\Delta = t_p * s_x = 1,96 * 0,8716 = 1,7083$$

Az értékösszegsor konfidencia intervalluma $50000 * (497 - 1,7083; 497 + 1,7083)$ g, azaz 24.765kg és 24935 kg között várható az 50.000 db üveg osztó töltő súlya.

7.2.2.6. A populáció aránybecslése

A nem homogén populációt valamilyen minőségi vagy mennyiségi ismerv alapján két csoportba soroljuk és az egyes csoportokba esés valószínűségét kívánjuk meghatározni, akkor aránybecslést végzünk.

A populáción belül egy adott tulajdonságú egyedek aránya p , ez azt jelenti, hogy egy egyedet kiválasztva p a valószínűsége annak, hogy az egyed rendelkezik az adott tulajdonsággal. Továbbra is feltételezzük, hogy független, azonos eloszlású minta áll rendelkezésünkre.

A standard hibát az arányokból számoljuk ki.

$$\sigma_p = \sqrt{\frac{p^*(1-p)}{n}} \text{ vagy } s_p = \sqrt{\frac{p^*(1-p)}{n}}$$

A becslés a továbbiakban a z-próba függvényével történő becsléssel megegyező.

10. bemutató feladat:

Gazdálkodókat kérdeztek meg, hogy két pohánka vetőmag közül melyiket választják. 2000 megkérdezett közül 700 gazdálkodó a „Hajnalka” elnevezésűt választaná, ha vetésre kerül a sor. Becsüljük meg 99%-os valószínűséggel, hogy milyen határok között lenne az ezt a fajtát választók aránya, ha az összes pohánkát vető gazdát megkérdezhettük volna.

$$p = 700/2000 = 0,35$$

$$\sigma_p = \sqrt{\frac{p^*(1-p)}{n}} = \sqrt{\frac{0,35^*(1-0,35)}{2000}} = 0,011$$

$$1 - \frac{\alpha}{2} = 1 - 0,005 = 0,995 \quad z_{0,995} = 2,58; \quad \Delta = 2,58 * 0,011 = 0,028 = 2,8\%$$

$$(35 - 2,8; 35 + 2,8) = (32,2; 37,8)\%$$

99%-os valószínűséggel arra számíthatunk, hogy gazdálkodók 32,2-37,8%-a a „Hajnal” elnevezésű pohánkát vetné.