Becslések

Induktív statisztika versus leíró statisztika
Az előző félévben statisztikai eszköztárként **leíró statisztikával** foglalkoztunk. Ide tartoznak a táblázatok, diagramok, középértékek, szóródási mérőszámok, eloszlás jellemzése, viszonyszámok**. A leíró statisztika célja a megfigyeltek tömör jellemzése. Ennek eredménye teljes körű megfigyelés esetén az alapsokaságra vonatkozik, míg mintavétel esetén kizárólag a mintára vonatkozik.**

A kérdés az, hogy minták alapján, hogyan tudjuk jellemezni az alapsokaságot. Erre alkalmazhatjuk az induktív statisztikát**. Ezen belül becsléseket és a hipotézisvizsgálatot fogjuk taglalni.** Induktív statisztika csak véletlen mintavétel esetén alkalmazható!

Becslések során valamely sokasági jellemző mintából történő közelítő meghatározása a cél, míg **Hipotézisvizsgálat** során egy a sokaságra vonatkozó előzetes állítást, feltételezést vizsgálunk minta(k) alapján.

Induktív statisztika versus leíró statisztika

Statisztika	Teljes körű megfigyelés	Reprezentatív minta	További részleges megfigyelések
Leíró statisztika	Eredménye az **alapsokaságra** vonatkozik	Eredménye a **mintára** vonatkozik	Eredménye a **megfigyeltekre** vonatkozik
Induktív statisztika	✘	Eredménye az **alapsokaságra** vonatkozik	✘

Becslések
Amikor egy véletlen minta alapján az alapsokaság valamely paraméterét, jellemzőjét szeretnék becsülni, két lehetőség van:

- **Pontbecslés** esetén a becslés eredménye egy konkrét szám, mely a mintából kiszámítható. Ezt tekintjük a becsülni kívánt sokasági paraméter becsült értékének.
- **Intervallumbecslés** esetén egy olyan intervallumot határozunk meg, amely előre adott valószínűséggel tartalmazza a becsülni kívánt paramétert. Ezt az intervallumot **konfidenciaintervallumnak, vagy más szóval megbízhatósági intervallumnak** nevezzük.

Pontbecslések
Pontbecslések esetében a sokaság valamely jellemzőjét (például átlag, szórás, valamilyen tulajdonságú elemek aránya) a minta alapján szeretnénk megbecsülni. Mivel véletlen mintákat vizsgálunk, így mindegyik mintaelem, illetve ezek átlaga, szórása is valószínűségi változónak

tekinthető, azaz a sokasági jellemzők becslésekor a valószínűségszámításból tanultakat kellene alkalmazni. Kérdés ezt hogyan tehetjük meg. Ha a mintában kiszámítom a mintaelemek átlagát, szórását, valaminek az arányát, az tekinthető-e a sokasági átlag, szórás, arány becslésének? A becslésekkel az egyik legalapvetőbb elvárás a **torzítatlanság**. Mit jelent ez? Amikor egy sokasági jellemzőt egy mintából becslünk (azaz a mintából kiszámítunk, meghatározunk) valamit, akkor mintánként eltérő eredményre juthatunk. A torzítatlanság ekkor azt jelenti, hogy tekintve az összes adott elemszámú mintát, akkor a kiszámított/meghatározott jellemzők várható értéke (~átlaga) meg kell, hogy egyezzen a sokasági jellemző értékével. Például, amikor a sokasági átlagot szeretnénk megbecsülni egy minta alapján, akkor mintánként más lehet a mintaátlag. A torzítatlanság azt jelentené, hogy az összes adott elemszámú mintaátlag átlaga megegyezik az alapsokaság átlagával.
Bizonyítani lehet, hogy a mintaátlag a sokasági átlag torzítatlan becslésének tekinthető, ugyanakkor a korábban tanult szórásnégyzet, illetve szórás képlet nem torzítatlan becslése a sokasági szórásnégyzetnek, szórásnak. E helyett becslésként, az úgynevezett korrigált tapasztalati szórással

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(d_i)^2}{n-1}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} f_i(d_i)^2}{n-1}}$$

becsülhetjük a sokasági szórás értékét. A képlet a sokasági szórás képletéhez képest annyiban más, hogy a nevezőben nem a megfigyelt elemszám, hanem n-1 szerepel. Az Excelben a szórásp() függvény az sokasági szórását, míg a szórás() függvény a korrigált tapasztalati szórást számolja.
Az eddig leírtak szemléltetését tartalmazza a mintavételieloszlás.xls fájl. Ebben egy ötelemű alapsokaság átlagát, szórásnégyzetét, szórását láthatjuk. Az FAE munkalapon az összes lehetséges 2 elemű mintát előállítottuk (mivel ez visszatevéses mintavétel, így ezek száma: 5x5=25). Ebben láthatjuk, hogy a 25 darab mintaátlag átlaga megegyezik a sokasági átlaggal, míg a 25 darab szórásnégyzet átlaga nem egyezik meg az alapsokaság szórásnégyzetével. Ugyanakkor a 25 darab korrigált tapasztalati szórásnégyzet átlaga megegyezik az alapsokaság szórásnégyzetével. Az EV munkalap esetében szintén láthatjuk a sokasági átlag becslését.
Ez a mintapélda a valóságban nem játszható el, ugyanis nem ismerjük a sokaság paramétereit (ezért akarjuk becsülni), másrészt a 25 tényleges minta közül csak 1 áll rendelkezésünkre, amely alapján becslést kell adnunk, ugyanakkor fontos szerepe van a becslések során fellépő hibák kiszámításának megértésekor.

Sokasági jellemzők pontbecslése mintákból

Sokasági jellemző	Ennek pontbecslése a mintából
Sokasági átlag (μ)	mintaátlag (xátlag)
Sokasági szórás (σ)	Korrigált tapasztalati szórás (s)
Sokasági arány (P)	Mintabeli arány (p)

Amikor minták alapján a sokaság valamely paraméterének pontbecslését végezzük el, akkor a becslés során hiba léphet el, amely a paraméter tényleges és becsült értékének különbsége. Az előző teoretikus példában bármely esetben meg tudjuk mondani a becslés hibáját (mintaátlag-

sokasági átlag), ugyanakkor a gyakorlatban ez nem lehetséges, mivel a sokasági átlag értékét nem ismerjük. A becslés során fellépő hiba gyakorlati jellemzésére két lehetőségünk van:

1. **Átlagos hiba (becslés standard hibája; angolul: standard error of estimation): a**dott elemszámú minták esetében a mintából számított pontbecslések **átlagosan** mennyivel térnek el a sokasági paramétertől. Ez az előző példában azt jelenti, hogy a mintaátlagok sokasági átlagtól való átlagos eltérését, azaz a mintaátlagok szórását kell kiszámítani. **Ez egy közönséges szórás,** így a standard hiba jelölésére a szórás jelölését használjuk, alsó indexben pedig azt tüntetjük fel, hogy mely mintajellemzővel dolgozunk. A becslés standard hibájának kiszámításához ismételten szükségünk lenne a 25 minta és a sokaság átlagának ismeretére is. Ugyanakkor bizonyítani lehet, hogy ezt anélkül is megtehetjük. Például:

Minta	**Sokasági átlag becslésének standard hibája**	
	ha az alapsokaság szórása ismert	**ha az alapsokaság szórása nem ismert**
FAE-minta	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$s_{\bar{x}} = \frac{s}{\sqrt{n}}$
EV-minta	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$	$s_{\bar{x}} = \frac{s}{\sqrt{n}} \cdot \sqrt{1-\frac{n}{N}}$
Arányosan rétegzett R-minta	$\sigma_{\bar{x}} = \frac{\sqrt{\sum_{j=1}^{k} n_j . \sigma_j^2}}{n}$	$s_{\bar{x}} = \frac{\sqrt{\sum_{j=1}^{k} n_j . s_j^2}}{n}$

A sokasági arány becslésének standard hibája

Minta	**Standard hiba**
FAE-minta	$s_p = \sqrt{\frac{pq}{n}}$, ahol q=1-p
EV-minta	$s_p = \sqrt{\frac{pq}{n} \cdot \frac{N-n}{N-1}}$

A becslések standard hibájáról elmondható, hogy minél nagyobb elemszámú mintákkal dolgozunk, annál kisebb lesz a standard hiba. FAE-minták esetében, ha feleakkora standard hibát szeretnénk, akkor ahhoz négyszer akkora mintát kell vennünk.

A statisztikai szoftverek többsége alapértelmezés szerint azt feltételezi, hogy FAE-mintával dolgozunk és a sokaság szórásáról nem tudunk semmit.

Példa

Valamely azonnal oldódó kávékivonatot automata tölti az üvegekbe. Előző adat-felvételekből ismeretes, hogy a gép által töltött tömeg jó megközelítéssel normális eloszlású valószínűségi változónak tekinthető, 1 g szórással. A gép pontosságának ellenőrzésére vett 22 elemű mintában (FAE-minta) az üvegekben lévő kávé granulátum tömege (gramm):

55, 54, 54, 56, 57, 56, 55, 57, 54, 56, 55, 54, 57, 54, 56, 50,54,56,54,56,50,60.

A) Mekkora a sokaság elemszáma, mennyi a sokaság átlaga?

B) Mekkora a minta elemszáma?
C) Készítsen pontbecslést a várható átlagos töltőtömegre! Mekkora a becslés standard hibája?
D) Készítsen pontbecslést az 55 grammnál nehezebb üvegek arányára! Mekkora a becslés standard hibája?

Megoldás:
A feladat szövegkörnyezete alapján egy 22 elemű mintánk van (n=22), a sokaság nagyságáról és így a sokaság átlagáról semmit sem tudunk. A sokaság átlagát a mintaátlaggal tudjuk becsülni.

$$\bar{x} = \frac{55+54+54+56+57+56+55+57+54+56+55+54+ 57+54+56+50+54+56+54+56+50+60}{22} = 55 \text{ g.}$$

Tehát a sokasági átlag pontbecslése 55 gramm. Mivel az alapsokaság szórása ismert (előző vizsgálatokból ismert a szórás), illetve FAE mintával van dolgunk, így a becslés standard hibája: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 1/4 = 0{,}21$ gramm.

A mintában 10 üveg nehezebb 55 grammnál, így az 55 grammnál nehezebb üvegek arányának pontbecslése: $10/22 = 0{,}455 \Rightarrow 45{,}5\%$. A becslés standard hibája: $s_p = \sqrt{\frac{0{,}455 \cdot (1-0{,}455)}{22}} = 0{,}106$. azaz az 55 grammnál nehezebb üvegek arányának becslése 45,5%, a becslés standard hibája pedig 10,6 százalékpont.

2. **Maximális hiba (hibahatár):** Adott elemszámú minták esetében a mintából számított pontbecslések **maximálisan** mennyivel térnek el a sokasági paramétertől, **adott valószínűség mellett**. A hibahatár jele: Δ. Mivel több olyan valószínűségi tétel is létezik (például: Csebisev-egyenlőtlenség), mely egy jellemző (esetünkben pld. mintaátlag, arány) és a várható értékének (esetünkben pld. mintaátlag, arány) a maximális eltérését, a jellemző szórásának (ez esetünkben pont a standard hiba) valahányszorosaként (ez pont a valószínűség) jellemzi, így a hibahatárt az alábbi szerkezetben keressük:

$$\Delta = k \bullet \text{becslés standard hibája.}$$

k értéke a vizsgált mintajellemző valószínűségi eloszlása alapján számítható ki.

- Mintaátlagok esetében alkalmazható a centrális határeloszlás, mely szerint minél nagyobb a minta elemszáma, a mintaátlagok annál inkább normális eloszláshoz fognak tartani, azaz kellően nagy elemszámú minták esetén a mintaátlagok megközelítőleg normális eloszlásúak. Ehhez azonban ismerni kell az alapsokaság szórását, ugyanakkor ez a valóságban nem teljesül. Ilyenkor a szórást a mintából a korrigált tapasztalati szórás segítségével számítjuk ki, a mintaátlagok úgynevezett pedig t-eloszlást fognak követni.

Mintaátlagok eloszlása

Mintanagyság	Eloszlás
Nagyminta (n>=100)	Megközelítőleg normális eloszlás (Centrális határeloszlás tétel alapján), tehát k értékét - ha ismerjük az alapsokaság szórását normális eloszlás alapján - ha nem ismerjük az alapsokaság szórását t-eloszlás alapján számítjuk ki.
30<=n<100	- Ha nincs erős baloldali aszimmetria (mutatók értéke <1), akkor megközelítőleg normális eloszlás, tehát k értékét ha ismerjük az alapsokaság szórását normális eloszlás alapján - ha nem ismerjük az alapsokaság szórását t-eloszlás alapján számítjuk ki.
30<n	Megegyezik a vizsgált változó eloszlásával. k értékét az eloszlásnak megfelelően számítjuk ki.

- Arányok esetében amennyiben a vizsgált tulajdonságú elemekből és a komplementerből is van legalább 10-10 elem, akkor a k értékét binomiális eloszlás helyett normális eloszlással határozhatjuk meg.

A k értékét meghatározásához a megfelelő eloszlás eloszlásfüggvényének táblázatát szoktuk használni. Ugyanakkor társadalomtudományi elemzések során, általában 95 százalékos valószínűséget alkalmazunk, de előfordul 90%, 99% is. **Minél nagyobb a valószínűség a hibahatár annál nagyobb lesz.** Normális eloszlás alkalmazhatósága esetében a valószínűség ismeretében k értéke egy konkrét szám lesz a mintanagyságától függetlenül. Ezzel szemben t-eloszlás esetében k-értéke nem csak a valószínűségtől, hanem a mintanagyságától is függ, így ennek értéke különböző mintanagyságok esetében különböző, ezt számítógép segítségével fogjuk kiszámítani.

.

Hibahatár kiszámítása

Minta	**valószínűség**		
	90%	**95%**	**99%**
sokasági átlag becslése, ha az alapsokaság szórása ismert	$\Delta = 1{,}65 \bullet \sigma_{\bar{x}}$	$\Delta = 1{,}96 \bullet \sigma_{\bar{x}}$	$\Delta = 2{,}58 \bullet \sigma_{\bar{x}}$
sokasági átlag becslése, ha az alapsokaság szórása nem ismert	$\Delta = t_{0,95}(n-1) \bullet s_{\bar{x}}$	$\Delta = t_{0,975}(n-1) \bullet s_{\bar{x}}$	$\Delta = t_{0,995}(n-1) \bullet s_{\bar{x}}$
sokasági arány becslése	$\Delta = 1{,}65 \bullet s_p$	$\Delta = 1{,}96 \bullet s_p$	$\Delta = 2{,}58 \bullet s_p$

Példa

Valamely azonnal oldódó kávékivonatot automata tölti az üvegekbe. Előző adat-felvételekből ismeretes, hogy a gép által töltött tömeg jó megközelítéssel normális eloszlású valószínűségi változónak tekinthető, 1 g szórással. A gép pontosságának ellenőrzésére vett 22 elemű mintában (FAE-minta) az üvegekben lévő kávé granulátum tömege (gramm):

55, 54, 54, 56, 57, 56, 55, 57, 54, 56, 55, 54, 57, 54, 56, 50,54,56,54,56,50,60.

95 százalékos valószínűség mellett adja meg a hibahatárt

A) a várható átlagos töltőtömeg becslésekor,

B) az 55 grammnál nehezebb üvegek arányának becslésekor!

Megoldás:

Mivel kismintánk van, és a változó normális eloszlású, a várható átlagos töltőtömeg becslésekor $\Delta = 1{,}96 \bullet \sigma_{\bar{x}} = 1{,}96 \bullet 0{,}21 = 0{,}41$ gramm

az 55 grammnál nehezebb üvegek arányának becslésekor, mivel van legalább tíz 55 grammnál nehezebb és legalább tíz nem nehezebb üveg is, így $\Delta = 1{,}96 \bullet 0{,}106 = 0{,}208 \Rightarrow 20{,}8$ százalékpont.

Sokasági átlag és arány intervallumbecslése

Az intervallumbecslés eredménye egy olyan tartomány, mely adott valószínűséggel tartalmazza a becsülni kívánt sokasági jellemzőt. Például az alapsokaság átlagának becslésekor a 95 százalékos konfidenciaintervallum azt jelenti, hogy az összes lehetséges – adott elemszámú – mintát véve, átlagosan az esetek 95 százalékában a sokasági átlag bele esik a konfidenciaintervallumba.

Mivel a hibahatár megadja azt, hogy adott valószínűség és mintanagyság mellett mekkora a maximális eltérés a sokasági jellemző és a mintajellemző, így a konfidenciaintervallum szerkezete az alábbi: (*pontbecslés – hibahatár*; *pontbecslés + hibahatár*)

Példa

Valamely azonnal oldódó kávékivonatot automata tölti az üvegekbe. Előző adat-felvételekből ismeretes, hogy a gép által töltött tömeg jó megközelítéssel normális eloszlású valószínűségi változónak tekinthető, 1 g szórással. A gép pontosságának ellenőrzésére vett 20 elemű mintában (FAE-minta) az üvegekben lévő kávé granulátum tömege (gramm):

55, 54, 54, 56, 57, 56, 55, 57, 54, 56, 55, 54, 57, 54, 56, 50,54,56,54,56,50,60.

95 százalékos valószínűség mellett becsülje meg

a) a várható átlagos töltőtömeget,

b) az 55 grammnál nehezebb üvegek arányáT

Megoldás

(*pontbecslés – hibahatár*; *pontbecslés + hibahatár*)

sokasági átlag esetében: (*55 – 0,41*; *55 + 0,41*)=(54,59;55,41). Tehát 95 százalékos megbízhatósággal a töltőtömegek átlaga (a sokaságban) 54,59 g és 55,41 gramm közé esik.

sokasági arány esetében: (*0,455 – 0,208*; 0,455 + *0,208*)=(0,247;0,663). Tehát 95 százalékos megbízhatósággal az 55 grammnál nehezebb üvegek aránya 24,7% és 66,3% közé esik.

Mintaelemszámának meghatározása

A valóságban sokasági jellemzők becslésekor sokszor előre adott, hogy milyen maximális becslési hibával, azaz milyen hibahatárral szeretnénk dolgozni. A hibahatár és a standard hiba közti összefüggések segítségével egy becslést nyerhetünk a minimális mintanagyságra. Például sokasági átlag 95 százalékos valószínűség melletti becslésekor: $\boldsymbol{n = \left(\frac{1{,}96 \cdot \sigma}{\Delta}\right)^2}$.

Önellenőrző kérdések

1. Mi a különbség induktív statisztika és leíró statisztika között?
2. Milyen becslési eljárásokat ismer?
3. Mi a pontbecslés lényege? alapsokaság átlagát, szórását mivel lehet a mintából torzítatlanul becsülni?
4. Mit fejez ki a becslés standard hibája?
5. Mit fejez ki a hibahatár?
6. Mi az intervallumbecslés lényege?
7. Egy város a háztartások vízfogyasztásának átlagát szeretné megbecsülni 300 kiválasztott háztartás adatai alapján. Ekkor mi lesz a pontbecslés eredménye?
8. Becslést szeretnénk adni egy adott pártot preferálók arányáról 1000 fős minta alapján. Mi lehet ekkor a sokasági arány torzítatlan pontbecslése?

SZEGEDI TUDOMÁNYEGYETEM
GAZDASÁGTUDOMÁNYI KAR
KÖZGAZDÁSZ KÉPZÉS
TÁVOKTATÁSI TAGOZAT
LECKESOROZAT

JELEN TANANYAG
A SZEGEDI TUDOMÁNYEGYETEMEN KÉSZÜLT
AZ EURÓPAI UNIÓ TÁMOGATÁSÁVAL.
PROJEKT AZONOSÍTÓ: EFOP-3.4.3-16-2016-00014

SZÉCHENYI 2020

MAGYARORSZÁG KORMÁNYA

Európai Unió
Európai Szociális Alap

BEFEKTETÉS A JÖVŐBE