STATISZTIKA

Miért tanuljunk statisztikát? Mire használhatjuk?

Szakirodalom értő és kritikus olvasásához

- Mit állít egyáltalán a cikk?
- Korrektek-e a megállapítások?

Vizsgálatok (kísérletek és felmérések) tervezéséhez, kiértékeléséhez

- Mekkora mintával dolgozzunk?
- Felfedeztünk valamit, vagy csak a véletlen eredményezi azt, amit látunk?
- Mennyire megbízható az eredmény?

Az eredmények közléséhez, szemléltetéséhez

- Mit tegyünk a cikkbe? Az egész táblázatot, ábrákat, vagy csak néhány statisztikai mutatót?

A statisztika részei

Leíró statisztika (*descriptive statistics*):

Minden egyedet megvizsgálunk, az egész sokaság adatait összegezzük, többé-kevésbé részletesen

A megfigyelt adatokat tömörítjük az összegzés során, ezzel információt vesztünk.

induktív statisztika (*statistical inference*):

(indukció ~ általánosítás)

Egy, a sokaságból választott minta alapján a megfigyelt adatokból következtetünk az egész populációra jellemző adatokra.

Példa:

mintabeli selejtarány $\Rightarrow$ a sokaságban a selejt valószínűsége

Alapfogalmak

(statisztikai) populáció ~ alapsokaság (*population*)

A vizsgálandó egyedeknek vagy objektumoknak az a (teljes) köre, amelyre a vizsgálat irányul, azaz amelyre következtetéseinket vonatkoztatni szeretnénk

minta (*sample*)

A vizsgálandó egyedeknek vagy objektumoknak az a köre, amelyet ténylegesen megvizsgálunk, azaz amelynek adatain következtetéseink alapulnak

változó (*variable*)

adat, jellemző, ismérv, tulajdonság, amelyet a mintabeli egyedeken megfigyelünk, megmérünk, feljegyzünk (életkor, testtömeg, kapott kezelés típusa, időtartama, stb.). A mintán megfigyelt adatokat az *adatmátrix* tartalmazza; szokásos elrendezésében minden sor egy mintavételi egységnek és minden oszlop egy változónak felel meg.

megfigyelési egység (*observational vagy experimental unit*)

A populáció, illetve a minta egy eleme, egy egyed vagy objektum, amelynek adatait feljegyezzük (lehet egy ember vagy állat, egy élőhely, egy vérminta, egyedek egy csoportja, pl. egy család, stb.)

mintavételi egység (*sampling unit*)

Ugyanaz, mint a megfigyelési egység, ha gyakorisági adatokat számolunk. Egy egység az, amelyben számoljuk az egyedeket.

Ebben az esetben a megszámolt egyedeknek semmi közük a statisztikai populációhoz!

Megfigyelés	47.6 g	3	23
Változó	testtömeg	tojások száma	tulipánok száma
Megfigyelési egység	egy bizonyos területről származó széki lile	---	---
Mintavételi egység	---	egy fészek az adott területről	egy virágoskert az adott faluban
Minta	a területen befogott és megmért lilék	a vizsgált fészkek	a megvizsgált virágoskertek az adott faluban
Statisztikai populáció	a területen fellelhető összes lile	az összes fészek az adott területen	az összes virágoskert az adott faluban

Megfigyelés	Mintavételi egység
Orchideák száma	Meghatározott terület (kvadrát)
Tücskök száma a hálóban	A végigsöpört vegetáció térfogata
Méhek látogatási száma egy adott virágon	Meghatározott időintervallum
Gázlómadarak száma a tengerparton	A partvonal adott hosszúságú darabja
Bogarak száma egy csapdában	Adott méretű csapda
Ektoparaziták száma	Egy gazdaállat

Mintavételezés

A vizsgálatban a minta reprezentálja a populációt.

A minta **reprezentatív**, ha bármely tulajdonság előfordulási aránya megegyezik a mintában és a populációban.

A minta azonban gyakran torzított, amit számításba kell venni az eredmények interpretálásánál.

Mintavételi módszerek:

Egyszerű, véletlen mintavétel (*random sampling*):

Az alapsokaság minden egyede egyforma eséllyel kerül a mintába. A minta egyedeit egymástól függetlenül választjuk, például véletlenszám generálással.

Rétegezett mintavétel (*stratified sampling*):

Az alapsokaság valamilyen külső szempont szerint diszjunkt részekre bontható. Egyes rétegekben külön-külön véletlen mintavétel. (*A rétegek arányosan szerepeljenek a mintában?*)

Szabályos, szisztematikus mintavétel:

Ha lehetetlen a véletlen mintavétel kivitelezése. Csak az első egyedet választjuk véletlenszerűen, a többit a meghatározott mintavételi intervallumok kihagyásával (pl. minden harmadik egyedet választjuk be). Ekkor **a valószínűségszámítást nem alkalmazhatjuk** statisztikai következtetések levonására.

Mérési skálák (*measurement scales*)

Nominális (*nominal*)

Csak kategóriák vannak, nincs köztük rendezés, matematikai műveletek nem értelmezhetőek (hajszín, szemszín, ivar, faj)

Ordinális (*ordinal*)

A kategóriák között van rendezés, de matematikai műveletek nem értelmezhetőek („jó – közepes – rossz”, 1-5 skála az iskolai osztályozásban)

Intervallum (*interval*)

A matematikai különbségképzés már értelmes, az arány nem (°C vagy °F)

Arány vagy abszolút (*rate, absolute*)

Az arányképzés is értelmes, van abszolút 0, van fizikai jelentéstartalma annak, hogy egy mennyiség többszöröse a másiknak (testtömeg, K)

Konverzió intervallum vagy abszolút skáláról ordinálisra:

Időnként az intervallum skálán mért adatok nem alkalmasak bizonyos módszerekkel való feldolgozásra: konverzió. Pl. túl kevés adat, ismeretlen eloszlás stb..

Csoportosítás

Életkor helyett korcsoport, testtömeg helyett „kicsi-közepes-nagy“, stb.

Rangsorolás

Az adatokat sorba rendezzük és **rangszámot** (rank) adunk nekik.

Előfordulhatnak azonos megfigyelések, ekkor azzal az átlagos rangszámmal (**kapcsolt rangszám (**tied rank)) azonosítjuk, amelyet akkor kapnának, ha nem lennének azonos megfigyelések. *pl.*

Hossz:	*21.0*	*21.4*	*21.4*	*23.1*	*23.5*	*25.0*	*25.0*	*25.0*	*27.2*	*28*
rang	*1*	*2.5*	*2.5*	*4*	*5*	*7*	*7*	*7*	*9*	10

A relatív gyakoriságok közelítik az eloszlás sűrűségfüggvényét, a kumulált relatív gyakoriságok pedig az eloszlásfüggvényét.

Adatok ábrázolása

Gyakorisági táblázat (*frequency table*): megfigyelt numerikus adatok táblázatos ábrázolása → **gyakorisági eloszlás** (*frequency distribution*), **tapasztalati eloszlás** (*empirical distribution*)

Osztályok, osztályintervallumok kialakítása:

- Diszkrét: ha nincs túl sok érték, egy érték egy osztály, egyébként mint a folytonos esetben.
- Folytonos: 10-20 osztály, lehetőleg minden osztályba legalább 6 érték essen. Használjunk természetes osztályhatárokat!
  Konvenció: osztályokba az alsó határ beletartozik, a felső nem.

Abszolút vagy relatív (százalékos), esetleg kumulált gyakoriságok meghatározása

Osztály	-20	20-30	30-40	40-100	100-	Össz.
Gyakoriság	38	52	62	36	12	200
Kumulált gyak.	38	90	152	188	200	
Relatív gyakoriság	0.19	0.26	0.31	0.18	0.06	1
Kumulált rel. gyak.	0.19	0.45	0.76	0.94	1	

Hisztogram (*histogram*)

A hisztogram nem más, mint a tapasztalati sűrűségfüggvény.

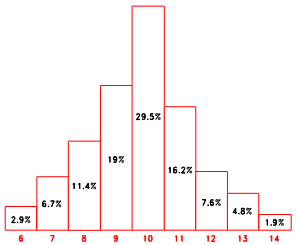
Vízszintes tengelyén: osztályintervallumok, fölötte olyan téglalapok, melyek ***területe*** megegyezik a megfelelő relatív, vagy százalékos gyakorisággal, így a hisztogram teljes területe 1, vagy 100% lesz.

Diszkrét változó esetén a változó értékei az intervallumok közepén helyezkednek el.

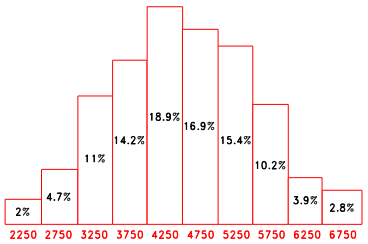
A hisztogram – ha a minta elemszámát növeljük – közelíti a valószínűségi változó elméleti sűrűségfüggvényét.

Ennek megfelelően a kumulatív hisztogram nem más, mint a tapasztalati eloszlásfüggvény

Haranggörbe alakú eloszlások

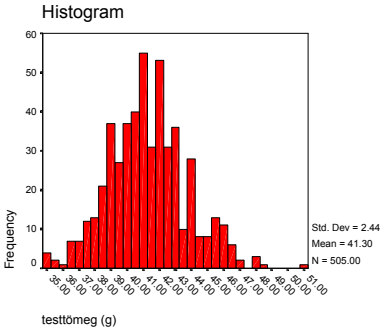


Anyakocák szaporaságának hisztogramja

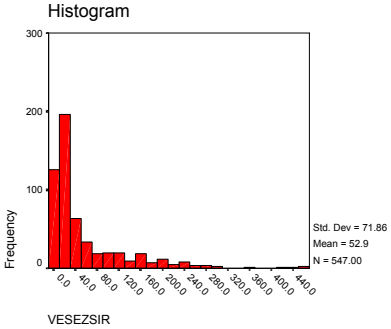


Tehenek éves tejtermelésének hisztogramja

Haranggörbe alakú eloszlások?



505 lile testtömege



Őzek vese körüli zsír mennyisége:

Középértékek

Adatok gyakorisági eloszlásának grafikus ábrázolása helyett összesítő mennyiségek, (alap)**statisztikák** (*statistic*).

Átlag (*average, mean*) $\bar{x}$**:**

Minta elemei: $x_1, x_2, ..., x_n$

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Az átlag az az érték, amely a "legközelebb" van a minta elemeihez.

A mintabeli értékek és a mintaátlag közti eltérések összege mindig 0:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n \cdot \bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

Gyakorisági táblázat esetén súlyozott átlag:

$$\bar{x} = \sum_{j=1}^{N} \frac{f_j \cdot x_j}{n}, \quad \text{ahol} \quad n = \sum_{j=1}^{N} f_j.$$

ahol az osztályokat x_j-vel, az egyes osztályokban levő adatok számát f_j-vel, és az osztályok számát N-nel jelöljük.

Vigyázat! Ha van egy 80 és egy 20 fős csoportunk, akkor ha megkérdezzük a TO-t, hogy mennyi az átlagos csoportlétszám, vagy pedig megkérdezzük a hallgatókat, hogy milyen létszámú csoportba járnak, és ezt átlagoljuk, az nem ugyanaz.

Nem jellemzi jól a mintát, ha az eloszlás nem szimmetrikus, vagy kiugró értékek vannak!

Példa.

Egy éjszaka 7 csapdába esett hangyák száma egy lombhullató erdőben:

25 4 12 9 15 8 202

$$\bar{x} = \sum_{i=1}^{7} x_i / 7 = 275 / 7 = 39.3$$

Medián (*median*)

Sorba rendezzük az adatokat: $x_1 \le x_2 \le \ldots \le x_n$,

$$x_{med} = x_{k+1}, \quad \text{ha} \quad n = 2k+1,$$

$$x_{med} = \frac{x_k + x_{k+1}}{2}, \quad \text{ha} \quad n = 2k.$$

Nem érzékeny az extrém értékekre.

Ordinális adatok esetén is használható statisztika, hiszen kiszámításához elegendő a megfigyelések sorrendjének ismerete (kivéve ha két középső van).

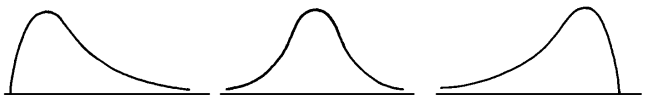
Módusz (*mode*)

A leggyakrabban előforduló érték. Nominális skálán mért adatokra csak ez a középérték alkalmazható.

A középértékek a hisztogramból is becsülhetők, bár a becslés nagyon függ az osztályokba sorolástól:

- A módusz az az érték, amely fölött a legmagasabb téglalap van.
- A mediántól balra és jobbra a hisztogram területének fele helyezkedik el.
- Az a pont az átlagérték, amelynél a hisztogram súlypontja van.
- Szimmetrikus és egy csúcsú hisztogram esetén a három középérték egybeesik (a szimmetria tengelyre).

Ferde eloszlás esetén az átlag mindig az eloszlás "farka" (tail) felé csúszik el. Biológiai eloszlásokban szinte mindig jobbra (pozitívan) ferde az eloszlás, így az átlag nagyobb mint a medián és a módusz.



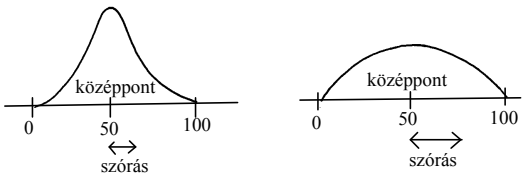
Jobbra ferde Szimmetrikus Balra ferde eloszlás

Összehasonlítás

átlag	leggyakoribb	mindig létezik	minden adatot felhasznál	extremális értékekre érzékeny	általánosan használt
medián	ritkább	mindig létezik			extremális értékek esetén jól jöhet
módusz	még ritkább				nominális skálára is jó

A szóródás mérőszámai

A középértékek nem jellemzik elég jól az eloszlást.



Kíváncsiak vagyunk arra is, hogy az adatok hogyan helyezkednek el az átlagérték körül.

Terjedelem (*range*)

A minta legnagyobb és legkisebb értéke közötti különbség.

$$R = x_{\max} - x_{\min}$$

Interkvartilis terjedelem (*interquartile range: IQR*)

A harmadik (Q_3) és az első kvartilis (Q_1) különbsége. (középső 50% terjedelme):

$$IQR = Q_3 - Q_1$$

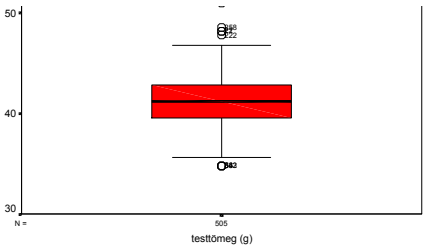
Kiugró értékek (*outlier*)

A minta olyan értékei, amelyek a többihez képest túl kicsik, vagy túl nagyok:

$$x_i < Q_1 - 1.5IQR$$
$$x_i > Q_3 + 1.5IQR$$

Grafikusan boxplot-tal ábrázolhatók: terjedelem (egyenes), medián, alsó és felső kvartilis (doboz), kiugró értékek.

Normális eloszlás esetén kiugró értékeknek tekinthetjük azokat, amelyek a szórás háromszorosánál jobban eltérnek az átlagtól.



Tapasztalati szórás és szórásnégyzet vagy variancia (*variance*)

A szórás a variancia négyzetgyöke (az alábbi s a szórás, négyzete s^2 pedig a variancia).

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}.$$ (ez a szórás "plug-in" becslése!)

A szórás azt mutatja meg, hogy az adataink átlagosan milyen távol helyezkednek el a számtani középtől.

Gyakorlatban az ún. **korrigált tapasztalati szórást** (*Standard Deviation*: SD) használjuk.

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}.$$

A nevezőben n-1 áll, ahol n a minta elemszáma. n-1 a **szabadsági fok** (*degrees of freedom*), ami a tényleges információ-tartalommal kapcsolatos. A szabadsági fok értéke attól függ, hogy egy, az adathalmazból számított mennyiséghez még hány értéket választhatunk meg szabadon úgy, hogy a már becsült értékek nem változnak. Az átlag esetén a szabadsági fok n. A szórás esetén egy becsült paramétert, az átlagot fel kell használnunk.

A szórásnak ugyanaz a mértékegysége, mint az eredeti adatainké (ezért használjuk szívesebben, mint a varianciát).

Gyakorisági táblázat esetén:

$$s = \sqrt{\frac{\sum_{j=1}^{N} f_j\left(x_i - \bar{x}\right)^2}{n-1}}, \text{ ahol } n = \sum_{j=1}^{N} f_j.$$

Eltérés négyzetösszeg: *SS* (*sum of squares of deviations*).

$$SS = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}.$$

Variációs koefficiens (*coefficient of variation*)

Különböző átlagú minták szórásának összehasonlítása esetén.

$$CV\% = \frac{s}{\bar{x}} \cdot 100\%$$

Standard hiba (*standard error, SE*)

Teljes neve “a mintaátlag standard hibája”, azaz szórása.

$$SE(\bar{x}) = \frac{SD(X)}{\sqrt{n}}$$, ahol *n* a mintaelemszám.

A mintaátlag véletlentől függő mennyiség. Ha rögzítjük a mintaelemszámot, és ugyanabból a populációból többféleképpen választunk ugyanolyan elemszámú mintát, akkor természetesen más mintaátlagot kapunk. Az így kapott értékek szórása azonban kisebb, mint a populáció szórása, hiszen a mintában általában vannak az átlagostól kisebb és nagyobb értékek is, és ezek a különbségek az átlagszámításkor kioltják egymást.

Más becsléseknek is van SE-je, ez mindig a szóban forgó becslés szórását jelenti!

Ha a mintából készített hisztogram elég jól közelíti a normális görbét, akkor a normális eloszlás táblázatából kiolvasható, hogy

az $(\bar{x} - 1s, \bar{x} + 1s)$ intervallumban van adataink kb. 68%-a (**kb 2/3-a**),

az $(\bar{x} - 2s, \bar{x} + 2s)$ intervallumban van **kb. 95%-a**,

az $(\bar{x} - 3s, \bar{x} + 3s)$ intervallumba pedig kb. 99.7%-a esik (**majdnem mind**).

A szórás eredete:

A biológiai változatosság (szórás).

A mérési hiba:

- metodikai
- véletlen hiba

Lapultság és ferdeség

Lapultság vagy csúcsosság (*Kurtosis*)

Az eloszlás lapultságára, csúcsosságára vonatkozó statisztika. Normális eloszlás esetén értéke 0, laposabb eloszlás esetén negatív, csúcsosabb eloszlás esetén pozitív.

Ferdeség (*skewness*)

Az eloszlás ferdeségére vonatkozó statisztika. Szimmetrikus esetben 0, negatív esetben az eloszlás balra ferde, pozitív esetben jobbra ferde.

A lapultság és a ferdeség standard hibája a normalitás illetve szimmetria tesztelésére szolgálhat. Ha a statisztikák értéke beleesik a $\pm 2SE$ intervallumba, akkor feltételezhetjük a normalitást, illetve a szimmetriát.

Adatok transzformálása

Sok statisztikai módszer feltételezi a normalitást.

Gyakorisági adatok esetén nagyon gyakran ferde az eloszlás (binomiális, Poisson, negatív binomiális).

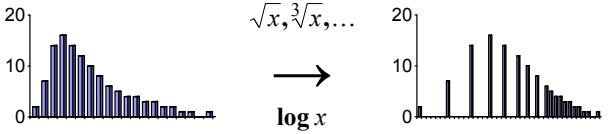
Ha nagyon ferde az eloszlás, az adatokat a paraméteres módszerek alkalmazhatósága érdekében lehet **normalizálni** (=normálissá transzformálni).

A paraméteres statisztikai módszerek, – amelyek két vagy több átlagot hasonlítanak össze – általában feltételezik, hogy a variancia a mintákban közel ugyanakkora. Poisson, binomiális és negatív binomiális eloszlás esetén a variancia függ az átlagértéktől.

A transzformációs technikák stabilizálják a varianciát, azaz megszüntetik az átlagtól való függést.

Transzformáció: $x_i \rightarrow f(x_i)$

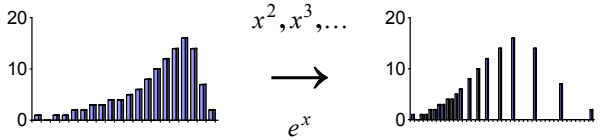
Például gyakorisági adatok esetén, ha $s^2 > \bar{x}$ a gyök- vagy a logaritmus-transzformáció segít:



Nem tökéletesen normális az új eloszlás, de normalizált, azaz a paraméteres módszerek használhatóak.

Ha vannak 0 értékek, akkor $\log x$ helyett $\log(x+1)$ használandó, ugyanis $\log 0$ nincs értelmezve…

A másik irányú ferdeség esetén a hatvány- vagy exponenciális transzformáció segíthet:



A négyzetgyök transzformáció

Poisson eloszlás vagy ha $s^2 \approx \bar{x}$ esetén használatos.

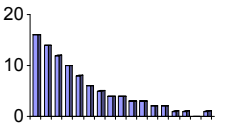
$$x \rightarrow \sqrt{x}$$

Az arcsin transzformáció

Megfigyelt arányok esetén használható. Az eloszlás mindkét "farka" le van vágva, hiszen minden érték 0 és 1 közé esik.

$$x \rightarrow \mathbf{arcsin}\sqrt{x}$$

Az adatok transzformálása segíthet, ha a vizsgálni kívánt változó nem normális eloszlású, de *a sikerre nincs garancia*, van olyan eset is, amikor az eloszlást semmilyen transzformáció sem képes normálissá tenni, mint például a következő ábrán:



Transzformációra szükség lehet más miatt is, például ha az értékek szóródása az értékek nagyságától függ (szóráskiegyenlítés), vagy ha két változó között a kapcsolat nem lineáris (linearizálás).

Figyelem! Előfordulhat, hogy az eredeti adatok biológiailag jól interpretálhatók, a transzformált adatoknak viszont már nem tudunk biológiai jelentést tulajdonítani. Ilyenkor inkább ne transzformáljunk.

Becslés (*estimation*)

A minta megfigyelései alapján a populációban valamely ismeretlen mennyiség vagy hatás mérése

Pontbecslés (*point estimate*)

A válasz egy szám. Mivel a mintából számítjuk, ez a szám a véletlentől is függ (az ebből adódó bizonytalanság mértékét leggyakrabban a becslés standard hibájával fejezzük ki)

Példák:

- minta átlag $\bar{x}$ → pop. átlag (E(X))
- minta variancia (korrigálatlan ill. korrigált)(s^2) → pop. variancia (var(X))
- mintabeli arány (relatív gyakoriság) → pop. arány (valószínűség)
- minta maximum → pop. maximum

A pontbecslés torzítatlansága

Általánosan: Egy α paraméterre egy $\hat{\alpha}(x_1, x_2, ..., x_n)$ **becslés**t adhatunk, amely

- a minta függvénye
- véletlen változó.

Vannak olyan becslések, amelyek a tapasztalatok alapján nem használhatóak. Például tendenciózusan alábecsülnek a következők:

- minta maximum → pop. maximum
- minta variancia (korrigálatlan) → pop. variancia (var(X))

Definíció:

$\hat{\alpha}(x_1, x_2, ..., x_n)$ **torzítatlan becslése** α-nak, ha

$$E(\hat{\alpha}(x_1, x_2, ..., x_n)) = \alpha.$$

Példa:

A mintaátlag torzítatlan becslése a populáció átlagnak: $E(\bar{x}) = E(X)$, mert

$$E\left(\frac{x_1 + x_2 + ... + x_n}{n}\right) = \frac{E(X) + ... + E(X)}{n} = E(X).$$

Definíció:

$\hat{\alpha}(x_1, x_2, ..., x_n)$ **aszimptotikusan torzítatlan becslése** α-nak, ha $n \to \infty$-re $E(\hat{\alpha}(x_1, x_2, ..., x_n)) \to \alpha$ (minél nagyobb a minta, annál kisebb a torzítás, sőt a mintaelemszám növelésével tetszőlegesen kicsivé tehető).

Általában, a statisztikában egy tulajdonságra akkor mondjuk, hogy "aszimptotikus", ha nagyon nagy ($n \to \infty$) minták esetén igaz.

Definíció:

$\hat{\alpha}(x_1, x_2, ..., x_n)$ **konzisztens becslése** α-nak, ha bármely ε>0-ra $P(|\hat{\alpha}(x_1, x_2, ..., x_n) - \alpha| \geq \varepsilon) \to 0$, ha $n \to \infty$. (azaz $\hat{\alpha}$-nak α-tól való "nagy" eltérésének valószínűsége 0-hoz tart, ha $n \to \infty$.)

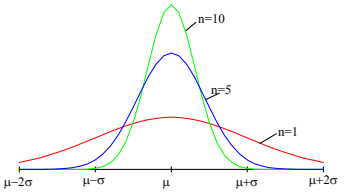
A populációátlag becslése a mintaátlaggal

A mintaátlagok nem egyenlők, és nem is egyeznek meg a populáció átlaggal. Mekkora a mintaátlag szórása vagy hibája (standard error: SE)?

A mintaátlag is egy valószínűségi változó: $\overline{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$

$SE = \frac{\sigma}{\sqrt{n}}$ a **mintaátlag szórása, vagy standard hibája**.

Ha n nő akkor a standard hiba csökken. Matematikailag bizonyítható (Centrális határeloszlás tétel), hogy függetlenül a mintaelemek eloszlásától, a mintaátlag eloszlása mindig a normális eloszláshoz tart, várható értéke a populáció várható értékével egyezik meg.

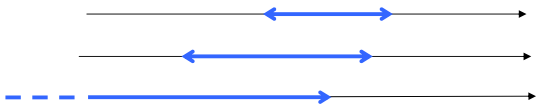


n>30 esetén feltételezhetjük a mintaátlag normalitását.

Intervallumbecslés (*interval estimate*)

Konfidencia-intervallum (*confidence interval*) esetén a válasz egy értéktartomány, amelybe az ismeretlen mennyiség 95% (esetleg 90% vagy 99%) valószínűséggel beleesik. A választott valószínűség a **megbízhatósági szint** (*confidence level*).

Általában szimmetrikus konfidencia-intervallumot keresünk (de nem mindig).



A konfidencia-intervallum konstrukciója nagyon egyszerű azokban az esetekben, amikor a szokásos pontbecslés – legalábbis közelítőleg – normális eloszlást követ (a $\overline{p}$, az $\overline{x}$, a $\overline{p}_2 - \overline{p}_1$, az $\overline{x}_2 - \overline{x}_1$ ilyenek), mert ekkor a normális eloszlásra érvényes képlettel számolhatunk:

95%-os intervallum: a pontbecslés ± 1.96 SE

Definíció:

Az eloszlás ismeretlen a paraméterének becslésekor a ***p* szintű konfidencia (megbízhatósági) intervallum** egy olyan (α_1, α_2) intervallum, amely p valószínűséggel tartalmazza a-t, azaz $P(\alpha_1 < a < \alpha_2) = p$.

Konfidencia-intervallum normális eloszlású változó átlagára

Tudjuk, hogy a mintaátlag eloszlása $\overline{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$, tehát a mintaátlag p valószínűséggel benne van a $\left(\mu - z_{\frac{1-p}{2}} \frac{\sigma}{\sqrt{n}}, \mu + z_{\frac{1-p}{2}} \frac{\sigma}{\sqrt{n}}\right)$ intervallumban.

Ez azt jelenti, hogy a mintaátlag p% valószínűséggel nem esik távolabb a populáció-átlagtól, mint $z_{\frac{1-p}{2}} \frac{\sigma}{\sqrt{n}}$. Ha a populáció-átlagot nem ismerjük, de egy mintaátlagot igen, akkor ebből visszakövetkeztethetünk a populáció-átlagra, így kapjuk a konfidencia-intervallumot.

Ha nem ismerjük a populáció szórását, σ-t, akkor megbecsülhetjük azt is ugyanabból a mintából, mint az $\bar{x}$-t, de ekkor a normális eloszlás kritikus értékei helyett a *t*-eloszláséit kell használnunk, így a konfidencia-intervallum:

$$\left(\bar{x} - t_{\frac{1-p}{2}} \cdot \frac{s}{\sqrt{n}}; \bar{x} + t_{\frac{1-p}{2}} \cdot \frac{s}{\sqrt{n}}\right).$$

A *t*-eloszlás szabadsági foka: n - 1.

n>50 esetén a *t*-eloszlás és a normális eloszlás már nem tér el nagyon, ezért közelítésként a normális eloszlás kritikus értékei is használhatók.

Bár általában azt mondjuk, hogy a populációátlag 95% valószínűséggel benne van a konfidencia-intervallumban, a szóhasználat helytelen. A populációátlag ugyanis egy pontosan adott, bár általunk nem ismert szám. Ha a konfidencia-intervallumot meghatároztuk, az vagy tartalmazza ezt az értéket, vagy nem, de az már nem véletlenszerű. A helyes szóhasználat az lenne, hogy az adott mintaelemszám mellett 95% valószínűséggel tudunk választani olyan mintát, amelyből számított konfidencia-intervallum ténylegesen tartalmazza a populációátlagot.

Konfidencia-intervallum két normális eloszlású változó átlaga közötti különbségre (független mintákon)

Ismert szórások esetén:

$$\left((\bar{x}_1 - \bar{x}_2) - z_{\frac{1-p}{2}}\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} \;,\; (\bar{x}_1 - \bar{x}_2) + z_{\frac{1-p}{2}}\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}\right),$$

ahol $\bar{x}_1$ és $\bar{x}_2$ a mintaátlagok, σ_1 és σ_2 az ismert szórások, n_1 és n_2 a mintaelemszámok, $z_{\frac{1-p}{2}}$ pedig a normális eloszlás megfelelő értéke.

Ismeretlen szórások esetén:

Ha van okunk feltételezni, hogy a szórások egyenlők:

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\frac{1-p}{2}}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1){s_1}^2 + (n_2-1){s_2}^2}{n_1 + n_2 - 2}} \;,\; (\bar{x}_1 - \bar{x}_2) + t_{\frac{1-p}{2}}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1){s_1}^2 + (n_2-1){s_2}^2}{n_1 + n_2 - 2}}\right),$$

ahol $\bar{x}_1$ és $\bar{x}_2$ a mintaátlagok, s_1 és s_2 a mintákból szokásos módon becsült szórások, n_1 és n_2 a mintaelemszámok, $t_{\frac{1-p}{2}}$ pedig az $n_1 + n_2 - 2$ szabadsági fokú *t*-eloszlás megfelelő értéke.

Ha a szórások egyenlőségét máshonnan nem tudjuk, *F*-próbával szokás ellenőrizni.

Ha a szórások egyenlősége nem feltételezhető (nem tudjuk előre, és az *F*-próba alapján is el kell vetni), nagy mintára (n_1 , $n_2 \geq 30$) közelítő érvénnyel az ismert szórások esetére megadott képlet is használható, egyszerűen a σ-k helyére a becsült szórásokat írva. Kis mintára a Welch-féle korrekció alkalmazható, amit most nem ismertetünk.

A statisztikusok egy része úgy véli, hogy a fentieknek nincs értelme. Általános esetben nem feltételezhető a szórások egyezősége, az F-próba alkalmazásával pedig felesleges bizonytalanság kerül a rendszerbe, ezért mindig úgy kell tekinteni, hogy a szórások különbözőek.

A vita a mai napig nincs eldöntve, ezért ebben az esetben úgy kell számolni, ahogy az adott tudományterületen (adott folyóiratban) szokás.

Konfidencia-intervallum két normális eloszlású változó átlaga közötti különbségre (ugyanazon egyedeken)

Ha mindkét változót ugyanazokon az egyedeken mértük, akkor először minden egyedre kiszámítjuk a két mért érték különbségét (d), majd ezekből a konfidencia-intervallumot az alábbi módon:

$$\left(\overline{d} - t_{\frac{1-p}{2}} \cdot \frac{s_d}{\sqrt{n}}, \overline{d} + t_{\frac{1-p}{2}} \cdot \frac{s_d}{\sqrt{n}}\right),$$

ahol $\overline{d}$ a különbségek átlaga, s_d a különbségek becsült szórása, n a mintaelemszám (úgy értve, hogy mindkét minta n elemű!), $t_{\frac{1-p}{2}}$ pedig az n-1 szabadsági fokú t-eloszlás megfelelő értéke.

Megjegyzések

Ugyanígy számolhatunk akkor is, ha a mérések nem ugyanazokon az egyedeken történtek, de a két minta elemei párosíthatók (pl. ikerpárok adatai).

Nem szükséges az, hogy mindkét változó normális eloszlású legyen, elegendő, ha a különbségek normális eloszlást követnek.

Nagy minták esetén ($n \geq 30$) közelítőleg érvényes akkor is, ha a különbség nem normális eloszlású.

Nagy minták esetén ($n \geq 50$) a t-eloszlás kritikus értékei helyett itt is használhatjuk a normális eloszlás kritikus értékeit.

Konfidencia intervallum populációbeli arányra (vagy esemény valószínűségére)

(binomiális eloszlás paraméterére)

Durva közelítés (a binomiálist normálissal közelítve):

$$\left(\hat{p} - z_{\frac{1-p}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{1-p}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

ahol $\hat{p}$ - a mintából becsült érték

Feltétel: $5 \leq n\hat{p} \leq n-5$

Finomabb közelítés:

$$\left(\frac{n\hat{p} + \frac{z^2_{\frac{1-p}{2}}}{2} - z_{\frac{1-p}{2}}\sqrt{\frac{z^2_{\frac{1-p}{2}}}{4} + n\hat{p}(1-\hat{p})}}{n + z^2_{\frac{1-p}{2}}}, \frac{n\hat{p} + \frac{z^2_{\frac{1-p}{2}}}{2} + z_{\frac{1-p}{2}}\sqrt{\frac{z^2_{\frac{1-p}{2}}}{4} + n\hat{p}(1-\hat{p})}}{n + z^2_{\frac{1-p}{2}}}\right)$$

Feltétel: $5 \leq n\hat{p} \leq n-5$

Példa:

Egy antigén 100 megvizsgált egyed közül 10 vérében volt kimutatható. Adjunk 95%-os konfidencia-intervallumot az antigénnel rendelkezők populációbeli arányára!

$n = 100$

$\hat{p} = 10/100 = 0.1 \Rightarrow n\hat{p} = 100 \cdot 0.1 = 10$

$z_{\frac{1-p}{2}} = z_{2.5\%} = 1.96$

A feltétel fennáll. Számoljunk a durva közelítéssel:

$$\left(\hat{p} - z_{\frac{1-p}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{1-p}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) =$$

$$\left(0.1 - 1.96\sqrt{\frac{0.1 \cdot 0.9}{100}}, 0.1 + 1.96\sqrt{\frac{0.1 \cdot 0.9}{100}}\right) = (0.041, 0.159)$$

A számítások követhetősége kedvéért most használjuk a konfidencia-intervallum konstrukciójára a legegyszerűbb eljárást. Ezzel a 95%-os intervallum:

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

ahol $\hat{p}$ a mintabeli arányt, n pedig a mintaelemszámot jelöli. Az intervallum szélessége innen a gyök alatti kifejezés szorozva 3.92-vel. Azt szeretnénk, hogy ez legfeljebb 10% legyen, azaz

$$3.92\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.1$$

A $\hat{p}$-t megsaccolva, majd az egyenlőtlenséget n-re megoldva kapjuk a mintaelemszámot. Például ha $\hat{p}$=0.3 körüli értékre számítunk, akkor $n \geq 325$ adódik.

Mindig – legyen szó akár átlagértékről, akár populáció arányól, vagy bármi másról – ugyanígy, a szóban forgó konfidencia-intervallum számítási képletéből kiindulva határozhatjuk meg a szükséges mintaelemszámot.

Persze mindig lesz olyan paraméter, amelyet ehhez meg kell saccolni, mert tőle is függ az intervallum szélessége.

A szükséges mintaelemszám meghatározása populációbeli arány becsléséhez

Számítsuk ki, mekkora minta szükséges ahhoz, hogy egy tulajdonság populációbeli előfordulási arányára adott 95%-os intervallum szélessége a 10%-ot ne haladja meg (mint például 26% - 36%).

Az hogy milyen széles konfidencia-intervallummal lehetünk elégedettek, az adott vizsgálat pontossági követelményei szabják meg.

A konfidencia-intervallum szélességét több dolog befolyásolja. Annál keskenyebb lesz az intervallum,

- minél kisebb megbízhatósági szintet követelünk meg (90% alá ne menjünk…)
- minél jobb, pontosabb eljárást alkalmazunk a konfidencia-intervallum konstrukciójára,
- minél nagyobb mintával dolgozunk,
- minél távolabb esik az arány az 50%-tól (bármelyik irányban)

A szükséges mintaelemszám meghatározása átlag becsléséhez

A konfidencia-intervallum fél-hossza: $h = z_{\frac{1-p}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

Ebből kifejezve a szükséges elemszámot: $n = \left(\frac{z_{\frac{1-p}{2}}\sigma}{h}\right)^2$

Ha nem ismerjük a populáció szórását, akkor előzetes mintából becsüljük a szórást:

$n = \left(\frac{t_{\frac{1-p}{2}}s}{h}\right)^2$, a $t_{\frac{1-p}{2}}$ szabadsági foka az előzetes minta elemszáma - 1.

Ha a kapott mintaelemszám nem nagyobb, mint az előzetes, akkor a meglévő minta már elegendő a kívánt pontossághoz.

Konfidencia-intervallum a populációbeli varianciára, ill. szórásra

A $\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$ statisztika χ^2 eloszlású, n-1 szabadsági fokú valószínűségi változó, ezért létezik olyan χ_1^2, χ_2^2, hogy $P\left(\chi_2^2 \le \chi^2 = \frac{(n-1)s^2}{\sigma^2} \le \chi_1^2\right) = p$

Az egyenlőtlenséget átrendezve: $P\left(\frac{(n-1)s^2}{\chi_1^2} \le \sigma^2 \le \frac{(n-1)s^2}{\chi_2^2}\right) = p$

χ_1^2 - $\frac{1-p}{2}$-höz tartozó χ^2 érték, (p=95% esetén a 0.025-höz tartozó kritikus érték)

χ_2^2 - $\frac{1+p}{2}$-höz tartozó χ^2 érték, (p=95% esetén a 0.975-höz tartozó kritikus érték)