Statisztika

Földtudomány szak, geológus szakirány, 2016/2017. tanév tavaszi félév

Backhausz Ágnes (ELTE TTK Valószínűségelméleti és Statisztika Tanszék)[1]

Tartalomjegyzék

[1]Kérdések, módosítási javaslatok, javítanivalók esetén: agnes@cs.elte.hu

1. Bevezetés

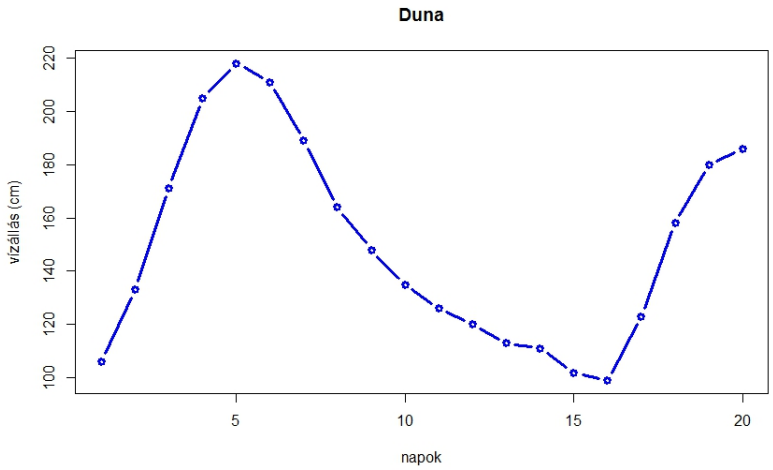
Célok: mérési eredmények, kísérletekből származó adatok alapján

- az adatok elemzése;
- a mért mennyiség vagy abból származtatott más mennyiségek becslése;
- hipotézisek ellenőrzése vagy cáfolata;
- múltbeli adatok alapján a jövőbeli folyamatok előrejelzése.

Alkalmazási területek:

- élő és élettelen természettudományok, társadalomtudományok: kísérleti eredmények értelmezése
- idősorok, véletlen folyamatok előrejelzése a természettudományokban vagy gazdaságtudományban;
- biztosítás– és pénzügyi matematika.

1.1. Példa: az adatok elemzése



1. ábra. A Duna vízállása húsz napon keresztül, éjfélkor (2016. január)

A Duna vízállása 2016. januárjában Budapestnél így alakult, húsz egymást követő napon (centiméterben mérve):

106	133	171	205	218	211	189	164	148	135
126	120	113	111	102	99	123	158	180	186

A fenti adatsort mintának nevezzük.

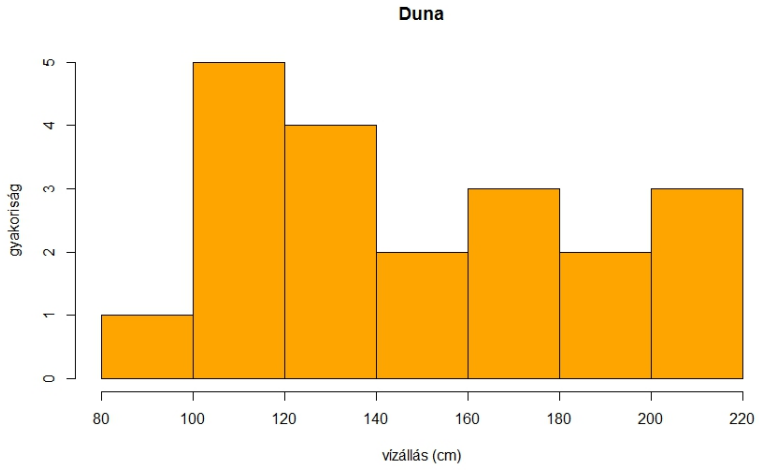
A mintaelemek száma, azaz a minta nagysága: $n = 20$. A legkisebb vízállás, vagyis a minimum 99 cm, a legnagyobb vízállás, vagyis a maxiumum 218 cm. Az átlagos vízállás 149,9 cm.

A minta mediánja (a nagyság szerinti sorrendben két középső mintaelem átlaga): 141,5 cm. A korrigált tapasztalati szórás: 38,55 (definíció később).

A vízállás 5 napon volt 115 cm-nél kevesebb (a napok egynegyedén), és 3 napon haladta meg a 2 métert (a napok 15%-án). A legnagyobb vízszintemelkedés 38 centiméter volt (a 2. és 3. nap között), a legnagyobb csökkenés 25 cm (a 7. és 8. nap között).

1.2. Példa: hisztogram

Az adatok ábrázolásának egy lehetséges módja hisztogram készítése. Választunk egy intervallumot, mely magában foglalja a mérési adatokat. Az intervallumot egyenlő nagyságú részekre osztjuk. Az így kapott kisebb intervallumok mindegyikéhez hozzárendeljük az abba eső mintaelemek számát (gyakoriságát), és ezt ábrázoljuk.



2. ábra.
A Duna vízállásáról kapott húszelemű mintából készített hisztogram

2. Alapstatisztikák

Minta (sample): $X_1, \ldots, X_n$ (ezek valószínűségi változók).

A minta **elemszáma** n (size).

Minimum: a legkisebb mintaelem, azaz $\min(X_1, X_2, \ldots, X_n)$.

Maximum: a legnagyobb mintaelem, azaz $\max(X_1, X_2, \ldots, X_n)$.

Terjedelem (range): a legnagyobb és legkisebb mintaelem különbsége, azaz

$$\max(X_1, X_2, \ldots, X_n) - \min(X_1, X_2, \ldots, X_n).$$

Módusz (mode): az a mintaelem, amelyik leggyakrabban fordul elő.

Átlag/mintaátlag (mean):

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Tapasztalati szórásnégyzet (uncorrected variance):

$$s_n^2 = \frac{1}{n}\left[\sum_{k=1}^{n}(X_k - \overline{X}_n)^2\right].$$

Tapasztalati szórás (uncorrected standard deviation):

$$s_n = \sqrt{\frac{1}{n}\left[\sum_{k=1}^{n}(X_k - \overline{X}_n)^2\right]}.$$

Korrigált tapasztalati szórásnégyzet (variance, var):

$$s_n^{*2} = \frac{1}{n-1}\left[\sum_{k=1}^{n}(X_k - \overline{X}_n)^2\right].$$

Korrigált tapasztalati szórás (standard deviation, sd):

$$s_n^* = \sqrt{\frac{1}{n-1}\left[\sum_{k=1}^{n}(X_k - \overline{X}_n)^2\right]}.$$

Szórási együttható (coefficient of variation [cv] / relative standard deviation [rsd]):

$$c_v = \frac{s_n^*}{\overline{X}_n}.$$

2.1. Példa: alapstatisztikák

Továbbra is a Duna vízállásáról kapott mintát használjuk (cm):

106	133	171	205	218	211	189	164	148	135
126	120	113	111	102	99	123	158	180	186

mintaelemszám: $n = 20$

minta: $X_1 = 106, X_2 = 133, \ldots, X_{10} = 135, \ldots, X_{20} = 186.$

átlag: $\overline{X} = 149,9$

tapasztalati szórásnégyzet: $s_n^2 = 1412,09$

tapasztalati szórás: $s_n = 37,58$

korrigált tapasztalati szórásnégyzet: $s_n^{*2} = 1486,411$

korrigált tapasztalati szórás: $s_n^* = 38,55$

szórási együttható: $c_v = 0,2571.$

2.2. Rendezett minta

Rendezett minta: a mintaelemeket nagyság szerint növekvő sorrendbe állítjuk. Jelölés:

$$(X_1^*, X_2^*, \ldots, X_n^*).$$

Vagyis $\{X_1^*, X_2^*, \ldots, X_n^*\} = \{X_1, X_2, \ldots, X_n\}$ és $X_1^* \leq X_2^* \leq \ldots \leq X_n^*$.

A minimum X_1^*, a maximum X_n^*. A k. legkisebb mintaelem X_k^*.

Példa: a vízállásról kapott húszelemű minta rendezett mintája:

99	102	106	111	113	120	123	126	133	135
148	158	164	171	180	186	189	205	211	218

$X_1^* = 99, X_2^* = 102, X_3^* = 106, \ldots, X_6^* = 120, \ldots, X_{10}^* = 135$

$X_{11}^* = 148, \ldots, X_{14}^* = 171, \ldots, X_{20}^* = 218.$

2.3. Az átlag és a szórásnégyzet tulajdonságai

2.1. állítás (A tapasztalati szórásnégyzet másik alakja). *A tapasztalati szórásnégyzet így is kiszámítható:*

$$s_n^{*2} = \frac{1}{n}\left[\sum_{k=1}^{n} X_k^2\right] - \overline{X}^2.$$

Bizonyítás. Átrendezéssel kapjuk, hogy

$$\sum_{k=1}^{n}(X_k - \overline{X})^2 = \sum_{k=1}^{n}\left[X_k^2 - 2X_k \cdot \overline{X} + \overline{X}^2\right] = \sum_{k=1}^{n} X_k^2 - 2n\overline{X} \cdot \overline{X} + n \cdot \overline{X}^2 =$$
$$= \sum_{k=1}^{n} X_k^2 - n \cdot \overline{X}^2.$$

Ebből adódik, hogy

$$s_n^2 = \frac{1}{n}\left[\sum_{k=1}^{n}(X_k - \overline{X})^2\right] = \frac{1}{n}\left[\sum_{k=1}^{n} X_k^2\right] - \overline{X}^2,$$

a tapasztalati szórásnégyzet definíciója alapján. □

2.4. Medián

Tekintsük az n elemű $(X_1, X_2, \ldots, X_n)$ mintát.

2.2. definíció. *Ha n páratlan: a rendezett minta* **középső elemét**, *azaz $X^*_{(n+1)/2}$-t a minta* **mediánjának** *nevezzük.*

Ha n páros: a rendezett minta $n/2$. és $n/2+1$. elemének átlagát, azaz a

$$\frac{X^*_{n/2} + X^*_{n/2+1}}{2}$$

mennyiséget a minta mediánjának nevezzük.

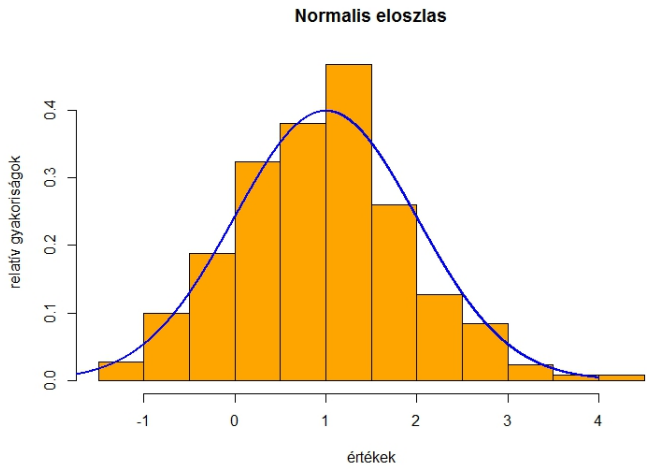
Megjegyzés: páros n esetén a teljes $\left[X^*_{n/2}, X^*_{n/2+1}\right]$ intervallumot (vagy annak bármely elemét) is a minta mediánjának lehet hívni.

Példa: a vízállásról kapott húszelemű minta mediánja:

$$\frac{1}{2}(X^*_{10} + X^*_{11}) = \frac{1}{2}(135 + 148) = 141,5.$$

2.5. Példa: az átlag és a medián összehasonlítása

Normális eloszlás



3. ábra. Az 500 elemű, normális eloszlású minta hisztogramja

500 elemű független minta: $X_1, X_2, \ldots, X_{500}$ függetlenek, eloszlásuk normális eloszlás $m = 1$ várható értékkel és $\sigma = 1$ szórással

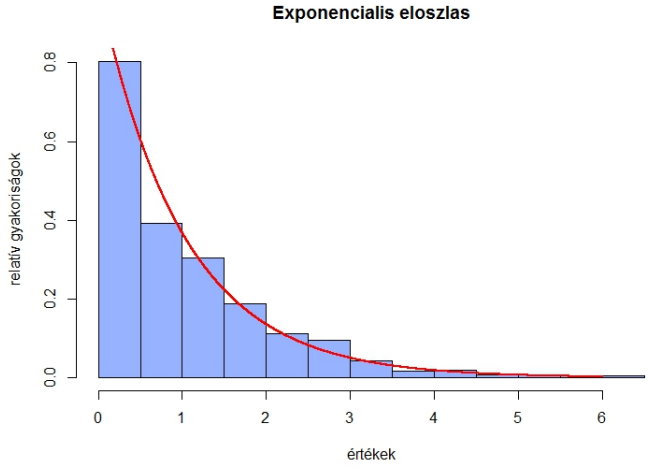
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.4870	0.3233	**0.9688**	**0.9599**	1.5320	4.4000

Exponenciális eloszlás

500 elemű független minta: $Y_1, Y_2, \ldots, Y_{500}$ függetlenek, eloszlásuk exponenciális eloszlás $b = 1$ paraméterrel. $\mathbb{E}(Y_k) = 1$ és $D(Y_k) = 1$ minden $k = 1, 2, \ldots, 500$-ra.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.001326	0.282700	**0.637300**	**0.984900**	1.349000	5.895000

A normális eloszlás esetében nincs nagy különbség az átlagra és a mediánra kapott értékek között, míg az exponenciális eloszlás esetén jelentős eltérés látszik (a várható érték és a szórás is mindkét esetben 1 volt, ebben nincs különbség).



4. ábra. Az 500 elemű, exponenciális eloszlású minta hisztogramja

Az $m = 1$ várható értékű és $\sigma = 1$ szórású normális eloszlás sűrűségfüggvénye szimmetrikus az 1 körül:

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t-1)^2}{2} \right) \qquad (t \in \mathbb{R}).$$

Az 1 paraméterű exponenciális eloszlás sűrűségfüggvénye nem ilyen:

$$g(t) = \begin{cases} \exp(-t), & \text{ha } t > 0; \\ 0, & \text{ha } t < 0. \end{cases}$$

Ha a sűrűségfüggvény szimmetrikus, akkor az átlag és a medián általában közelebb esik egymáshoz, mint ha nem érvényes a szimmetria. Ezért ha az adatok semmilyen szimmetriát nem mutatnak, gyakran a mediánt tüntetik fel. Szimmetrikus esetben inkább az átlagot használják.

2.6. Tapasztalati eloszlásfüggvény

Kérdés. Mennyi annak valószínűsége, hogy 2018. január 15-én a Duna vízállása 200 cm alatt marad? Mit tudunk erről mondani az adatok alapján?

Legyen X tetszőleges valószínűségi változó. Ennek eloszlásfüggvénye az az $F : \mathbb{R} \to [0, 1]$ függvény, melyre

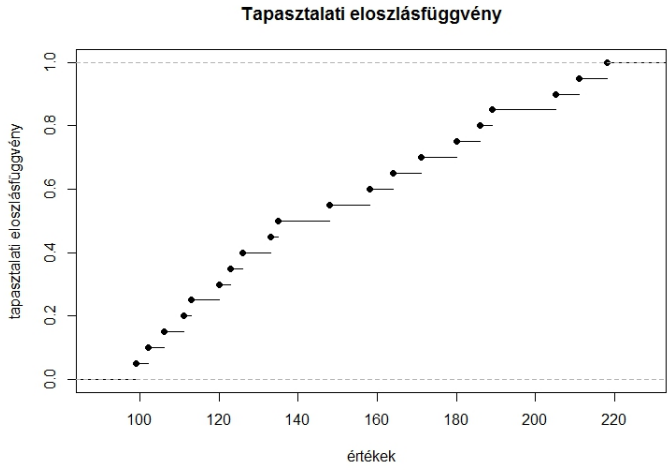
$$F(t) = \mathbb{P}(X \leq t)$$

minden $t \in \mathbb{R}$-re.

2.3. definíció (Tapasztalati eloszlásfüggvény). *Legyenek $X_1, X_2, \ldots, X_n$ valószínűségi változók. Ennek a mintának az eloszlásfüggvénye az az $\hat{F}_n : \mathbb{R} \to [0,1]$ függvény, melyre*

$$\hat{F}_n(t) = \frac{t\text{-nél kisebb mintaelemek száma}}{n} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}(X_k \leq t).$$

Itt $\mathbb{I}(X_k \leq t)$ értéke 1, ha $X_k \leq t$ teljesül (azaz a k. mintaelem legfeljebb t), és 0 különben. Tehát minden$t$-re megadjuk **a t-nél nem nagyobb mintaelemek arányát** a mintában.



5. ábra.
A Duna vízállásáról kapott húszelemű minta tapasztalati eloszlásfüggvénye

Például, a korábbi rendezett mintát tekintve a Duna vízállásáról:

99	102	106	111	113	120	123	126	133	135
148	158	164	171	180	186	189	205	211	218

A vízállás egy napon volt legfeljebb 100 cm, hat napon volt legfeljebb 120 cm, tizenkét napon volt legfeljebb 160 cm, és tizenhét napon volt legfeljebb 200 cm. Tehát:

$$\hat{F}_n(100) = 1/20 = 0,05; \qquad \hat{F}_n(120) = 6/20 = 0,3;$$

$$\hat{F}_n(160) = 12/20 = 0,6; \qquad \hat{F}_n(200) = 17/20 = 0,85.$$

2.7. Kvantilisek

Kérdés. Olyan magas gátat szeretnénk építeni, hogy nagyjából húszévente kerüljön csak sor árvízi védekezésre. Pontosabban, annak valószínűsége, hogy egy adott évben a legmagasabb vízállás legfeljebb 1/20 valószínűséggel emelkedjen a gát szintje fölé. Ha rendelkezésre állnak az egyes évek legmagasabb vízállásai, ez alapján milyen magasra kellene építenünk a gátat?

Legyen X valószínűségi változó, melynek eloszlásfüggvénye F:

$$F(t) = \mathbb{P}(X \leq t) \qquad (t \in \mathbb{R}).$$

Legyen $z \in [0,1]$ adott szám. Ekkor az F eloszlásfüggvény **z-kvantilise**:

$$q_z = \min\{t : F(t) \geq z\}.$$

Ha F szigorúan monoton növő, akkor $q_z = F^{-1}(z)$.

2.4. definíció (Tapasztalati kvantilis). *Legyen $X_1, X_2, \ldots, X_n$ minta, és $z \in [0,1]$ adott szám. Ekkor a minta tapasztalati z-kvantilise a tapasztalati eloszlásfüggvény z-kvantilise, vagyis:*

$$\hat{q}_z = \min\{t : \hat{F}_n(t) \geq z\}.$$

2.5. definíció (Tapasztalati kvartilisek). *A $z = 1/4$-hez tartozó $1/4$-kvantilist a minta első kvartilisének nevezzük, és Q_1-gyel jelöljük. A $z = 3/4$-hez tartozó $3/4$-kvantilist a minta harmadik kvartilisének nevezzük, és Q_3-mal jelöljük.*

Például, szintén a korábbi, vízállásra vonatkozó mintát tekintve legyen először $z = 0,5$. Azt a legkisebb szintet keressük, amire igaz, hogy a mintaelemek fele kisebb nála. Ez a nagyság szerinti sorrendben a 10. mintaelem lesz, tehát $q_{0,5} = 135$, a két középső mintaelem közül a kisebb.

Első kvartilis. A példában tekintsük az első kvartilist: $z = 1/4$. A legkisebb olyan szintet keressük, aminél a mintaelemek negyede kisebb vagy egyenlő. Mivel húszelemű a minta, ez a nagyság szerinti sorban az ötödik mintaelem lesz: $Q_1 = q_{1/4} = X^*_5 = 113$.

Harmadik kvartilis. Most azt a legkisebb szintet keressük, aminél a mintaelemek 3/4-e kisebb vagy egyenlő. Ez a tizenötödik lesz a nagyság szerinti sorban: $Q_3 = q_{3/4} = X^*_{15} = 180$.

További kvantilisek. Például $z = 0,2$ az, aminél az elemek egyötöde kisebb:

$$q_{0,2} = X_4^* = 111.$$

Az a szint, aminél a mintaelem $z = 0,95$ része kisebb (vagyis amit a mintaelemek 5%-a halad meg):

$$q_{0,95} = X_{19}^* = 211.$$

Kvantilisek számítása interpolációval. A fent megadott definíció helyett az alábbit is szokták használni. Ilyenkor a kvantilis nem a mintaelemek egyike, hanem a nagyság szerinti sorrendben két szomszédos mintaelem lineáris kombinációja.

1. n elemű minta z-kvantilisét szeretnénk meghatározni.
2. Legyen $m = \lfloor (n+1)z \rfloor$ az $(n+1)z$ egészrésze, $u = \{(n+1)z\}$ pedig ugyanennek a törtrésze.
3. A módosított definíció értelmében a tapasztalati z-kvantilis:

   $$q_z = X_m^* + u(X_{m+1}^* - X_m^*),$$

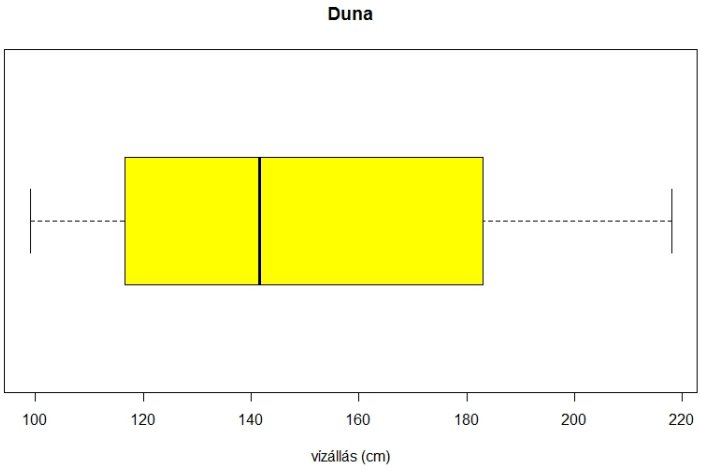
   ahol X_k^* a nagyság szerinti sorrendben a k. legkisebb mintaelem.

2.8. Példa: boxplot

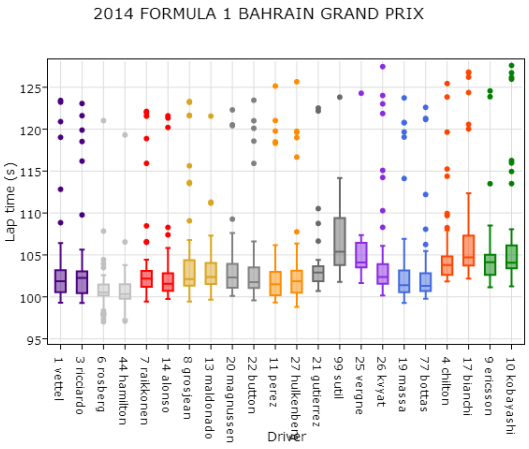
A mintaelemek ábrázolásának (és különösen más mintákkal való összehasonlításának) egy szokásos módja a boxplot készítése, melyhez a minta bizonyos kvantiliseit kell kiszámítani.

A boxplot készítéséhez szükséges adatok, és ezek értékei a vízállásra vonatkozó mintában:

- **minimum**: a legkisebb mintaelem (99);
- **első kvartilis**: a $z = 1/4$-hez tartozó kvantilis (118,2);
- **medián**: a középső mintaelem, vagy a két középső mintaelem átlaga (141,5);
- **harmadik kvartilis:** a $z = 3/4$-hez tartozó kvantilis (181,5);
- **maximum**: a legnagyobb mintaelem (218).



6. ábra. A Duna vízállásáról kapott húszelemű minta boxplotja.



7. ábra. Forrás: theansweris27.com

- **terjedelem:** a maximum és minimum különbsége.

Az egyes dobozok az első kvartilistől a harmadik kvartilisig tartanak. A középvonal helye a medián. A vonalak felölelhetik a teljes terjedelmet. Azok az adatok, melyek valamelyik irányban messzebb esnek a mediántól, mint az első és harmadik kvartilis közötti távolság másfélszerese, gyakran külön ponttal kerülnek ábrázolásra (ilyenkor a vonalak az utolsó olyan adatnál érnek véget, ami még belül van a másfélszeres távolságon).

2.9. Tapasztalati momentumok

Legyen továbbra is $X_1, X_2, \ldots, X_n$ a minta.

2.6. definíció. *Legyen $k \geq 1$ egész. Ekkor a minta k.* **tapasztalati momentuma** *(kth sample moment) a mintaelemek k. hatványainak átlaga:*

$$\frac{1}{n}\sum_{j=1}^{n} X_j^k.$$

Ekkor a minta k. **centrált tapasztalati momentuma** *(kth sample central moment):*

$$m_k = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^k.$$

2.7. definíció. *A* **tapasztalati ferdeség** *(sample skewness) két szokásos definíciója:*

$$\gamma = \frac{m_3}{s_n^{*3}} = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^3}{\left(\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}^2)\right)^{3/2}}.$$

$$\gamma_1 = \frac{n^2}{(n-1)(n-2)} \cdot \frac{m_3}{s_n^{*3}} = \frac{n}{(n-1)(n-2)} \sum_{j=1}^{n} \left(\frac{X_j - \overline{X}}{s_n^*}\right)^3.$$

Vegyük észre, hogy a definíciók csak az n-től függő szorzótényezőben különböznek. Heurisztika: ha az adatok hisztogramja nagyjából szimmetrikus (a medián körül), akkor a tapasztali ferdeség értéke a nullához közeli.

2.8. definíció. *A* **lapultság** *(sample kurtosis) egy lehetséges definíciója:*

$$\kappa = \frac{m_4}{m_2^2} - 3 = n \cdot \frac{\sum_{j=1}^{n}(X_j - \overline{X})^4}{\left(\sum_{j=1}^{n}(X_j - \overline{X}^2)\right)^2} - 3.$$

Ha Y normális eloszlású valószínűségi változó, akkor $\mathbb{E}(Y^4)/\mathbb{E}(Y^2)^2 = 3$, ezzel hasonlítják össze a mintából kapott értéket. Ha olyan eloszlásból veszünk mintát, melynek sűrűségfüggvénye közel van a normális eloszlás sűrűségfüggvényéhez, nulla közeli lapultságra számíthatunk. Pozitív lapultság "meredekebb" (abszolút értékben nagyobb deriválttal rendelkező), negatív lapultság kevésbé meredek sűrűségfüggvényre utalhat.

3. Statisztikai mező

3.1. definíció. *Az* $(\Omega, \mathcal{A}, \mathcal{P})$ *hármast* **statisztikai mezőnek** *nevezzük, ha minden* $\mathbb{P} \in \mathcal{P}$*-re* $(\Omega, \mathcal{A}, \mathbb{P})$ *Kolmogorov-féle valószínűségi mező.*

Vagyis: ugyanazon az alaphalmazon (elemi események halmazán és az események halmazán) több valószínűségi mérték adott. Frekventista megközelítés: a minta egyetlen $\mathbb{P}$-hez tartozó valószínűségi mezőből származik, és erről a $\mathbb{P}$-ről szeretnénk minél többet megtudni. (Ettől eltérő például a bayes-i módszerek alkalmazása, amiről nem fog szó esni.)

3.2. definíció. *Ha valamilyen* $\Theta \subseteq \mathbb{R}^q$ *halmazra a* $\mathcal{P}$ *halmaz felírható* $\{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ *alakban, akkor paraméteres statisztikai problémáról beszélhetünk. Ilyenkor a* Θ *halmazt paramétertérnek nevezzük.*

3.3. definíció ([1]). *Legyen* $(\Omega, \mathcal{A}, \mathcal{P})$ *statisztikai mező. Egy*

$$\underline{X} = (X_1, X_2, \ldots, X_n) : \Omega \to H \subseteq \mathbb{R}^n$$

valószínűségi vektorváltozót (n *elemű)* **mintának** *nevezünk. Itt* H *a* **mintatér**, n *a minta* **elemszáma** *vagy nagysága. Az* X_i *koordináták a minta elemei. Azt mondjuk, hogy a minta* **független**, *ha az* $X_1, X_2, \ldots, X_n$ *valószínűségi változók függetlenek.*

A mintatéren megadott $T : H \to \mathbb{R}^k$ *függvényt, illetve a* $T = T(\underline{X})$ *valószínűségi változót (*k*-dimenziós)* **statisztikának** *nevezzük.*

Példa. $X_1, X_2, \ldots, X_{20}$ a Duna vízállására fent megadott 20 elemű adatsor. Ekkor $n = 20$, a mintatér pedig legyen $H = [0, 2000]^{20} \subseteq R^{20}$, beépítve, hogy a vízállás nem lehet negatív vagy (mondjuk) 2000-nél nagyobb. Legyen $T : H \to \mathbb{R}$ az a függvény, mely H minden eleméhez hozzárendeli a koordinátáinak átlagát. Ekkor $k = 1$, és a statisztika:

$$T(\underline{X}) = \frac{X_1 + X_2 + \ldots + X_{20}}{n}.$$

Vagyis ebben az esetben a mintaátlag (mint valószínűségi változó) lesz a statisztika. (Viszont a minta nem független.)

További példák statisztikára:

- korrigált tapasztalati szórás:

$$T(X_1, \ldots, X_n) = s_n^* = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2};$$

- minimum és maximum (ilyenkor $k = 2$):

$$T(X_1, \ldots, X_n) = (\min(X_1, \ldots, X_n), \max(X_1, \ldots, X_n));$$

- terjedelem: $T(X_1, \ldots, X_n) = \min(X_1, \ldots, X_n) - \max(X_1, \ldots, X_n)$;
- medián;
- rendezett minta (ilyenkor $k = n$): $T(X_1, \ldots, X_n) = (X_1^*, X_2^*, \ldots, X_n^*)$.

4. A statisztika alaptétele

4.1. tétel (Glivenko, [1]). *Legyenek $X_1, X_2, \ldots, X_n$* **független** *azonos eloszlású valószínűségi változók, melyek közös eloszlásfüggvénye F. Ekkor az $\hat{F}_n$ tapasztalati eloszlásfüggvényekből álló sorozat 1 valószínűséggel egyenletesen tart F-hez, azaz*

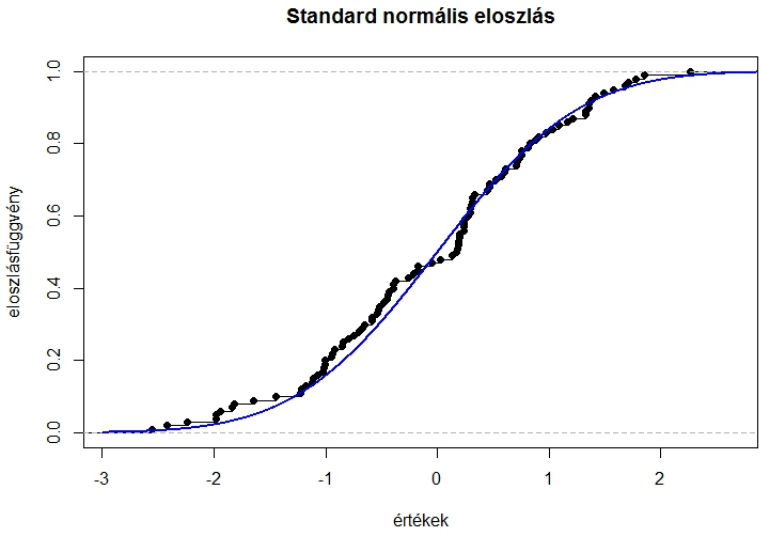
$$\mathbb{P}\left(\lim_{n \to \infty} \sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F(t) \right| = 0 \right) = 1.$$

Ennek a statisztikai mezőkre vonatkozó következményét így fogalmazhatjuk meg. Tegyük fel, hogy $X_1, X_2, \ldots$ független valószínűségi változók. Ekkor minden $n \geq 1$-re $(X_1, X_2, \ldots, X_n)$ független minta, amiből kiszámíthatjuk az $\hat{F}_n(t)$ tapasztalati eloszlásfüggvényt:

$$\hat{F}_n(t) = \frac{t\text{-nél nem nagyobb mintaelemek száma}}{n} = \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}(X_k \leq t).$$

Másrészt ha az $\mathbb{P}$ valószínűség a statisztikai mezőben az $\mathcal{P}$ egy tetszőleges eleme, akkor

$$F(t) = \mathbb{P}(X_1 \leq t) = \mathbb{P}(X_2 \leq t) = \ldots = \mathbb{P}(X_n \leq t).$$



8. ábra.
Standard normális eloszlás eloszlásfüggvénye és belőle vett 100 elemű minta tapasztalati eloszlásfüggvénye

Ilyenkor eszerint a $\mathbb{P}$ szerint egy valószínűséggel teljesül, hogy a tapasztalati eloszlásfüggvény és az "igazi" F eloszlásfüggvény közötti legnagyobb távolság nullához tart. (Tehát minden $\mathbb{P} \in \mathcal{P}$-re igaz, hogy a tapasztalati eloszlásfüggvény az ahhoz a $\mathbb{P}$-hez tartozó F-hez konvergál.)

A nagy számok erős törvénye szerint (ismét felhasználva a minta függetlenségére vonatkozó feltevést) az alábbi összefüggés teljesül minden rögzített $t \in \mathbb{R}$-re:

$$\mathbb{P}\Big(\lim_{n\to\infty} \big|\hat{F}_n(t) - F(t)\big| = 0 \Big) = 1.$$

A statisztika alaptétele ennél erősebbet állít: minden n-re kiválaszthatunk egy tetszőleges t pontot, ahol a különbséget kiolvassuk, és így is nullához tartó sorozatot kapunk.

5. Becslések és tulajdonságaik

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta)$ valamely Θ halmazzal (ezt paramétertérnek nevezzük). Legyen továbbá $\psi : \Theta \to \mathbb{R}$ függvény. Cél: olyan T statisztika keresése, amire a $T(X)$ valószínűségi változó és a $\psi(\vartheta)$ érték valamilyen értelemben közel esik a $\mathbb{P}_\vartheta$ valószínűség

mellett. Ezt minden $\vartheta \in \Theta$-ra szeretnénk.

5.1. Torzítatlanság és hatásosság

$\mathbb{E}_\vartheta$ azt jelenti, hogy a $(\Omega, \mathcal{A}, \mathbb{P}_\vartheta)$ valószínűségi mezőben számolunk várható értéket. A D^2_ϑ szórásnégyzetet és a D_ϑ szórást hasonlóképpen definiálhatjuk.

5.1. definíció (Torzítatlanság). *A $T : H \to \mathbb{R}$ statisztika torzítatlan becslés ψ-re, ha minden $\vartheta \in \Theta$-ra*

$$\mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) = \psi(\vartheta).$$

A T statisztika torzítása a $b_T(\vartheta) = \mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) - \psi(\vartheta)$ függvény.

5.2. állítás (A várható érték torzítatlan becslése). *Legyen $X_1, \ldots, X_n$ független azonos eloszlású minta. Legyen $\psi(\vartheta) = \mathbb{E}_\vartheta(X_1)$, azaz a mintának a $\mathbb{P}_\vartheta$ eloszlás szerinti várható értéke. Ekkor a $T(X_1, \ldots, X_n) = \overline{X}$ statisztika, vagyis a* **mintaátlag** *torzítatlan becslés ψ-re.*

Bizonyítás. A várható érték tulajdonságai alapján

$$\mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) = \mathbb{E}_\vartheta\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n}\big[\mathbb{E}_\vartheta(X_1) + \ldots + \mathbb{E}_\vartheta(X_n)\big].$$

Most felhasználva, hogy az $X_1, \ldots, X_n$ valószínűségi változók azonos eloszlásúak, vagyis a várható értékük is azonos:

$$\mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) = \frac{1}{n}[n \cdot \mathbb{E}_\vartheta(X_1)] = \mathbb{E}_\vartheta(X_1) = \psi(\vartheta).$$

Vagyis a mintaátlag torzítatlan függvénye a várható értéknek. □

5.3. állítás (A szórásnégyzet torzítatlan becslése). *$X_1, \ldots, X_n$ független azonos eloszlású minta. Legyen $\psi(\vartheta) = D^2_\vartheta(X_1)$, azaz a mintának a $\mathbb{P}_\vartheta$ eloszlás szerinti szórásnégyzete. Ekkor a $T(X_1, \ldots, X_n) = s^{*2}_n$ statisztika, vagyis a* **korrigált tapasztalati szórásnégyzet** *torzítatlan becslés ψ-re.*

Bizonyítás. A 2.1. állítás bizonyításának első egyenlősége szerint

$$s^{*2}_n = \frac{n}{n-1} s^2_n = \frac{n}{n-1}\left[\frac{1}{n}\left[\sum_{k=1}^n X_k^2\right] - \overline{X}^2\right] = \frac{1}{n-1}\left[\sum_{k=1}^n X_k^2\right] - \frac{n}{n-1}\overline{X}^2.$$

Felhasználva a szórásnégyzet definícióját, és hogy a valószínűségi változók azonos eloszlásúak:

$$\mathbb{E}_\vartheta\left(\sum_{k=1}^{n} X_k^2\right) = \sum_{k=1}^{n} \mathbb{E}_\vartheta(X_k^2) = n \cdot \mathbb{E}_\vartheta(X_1^2) = n \cdot \left[D_\vartheta^2(X_1) + \mathbb{E}_\vartheta(X_1)^2\right].$$

Másrészt, az összegre bontásnál felhasználva, hogy a valószínűségi változók függetlenek:

$$D_\vartheta^2(\overline{X}) = D_\vartheta^2\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n^2} D_\vartheta^2(X_1 + \ldots + X_n) = \frac{1}{n^2} \sum_{k=1}^{n} D_\vartheta^2(X_k) =$$

$$= \frac{1}{n^2} \cdot n \cdot D_\vartheta^2(X_1) = \frac{1}{n} D_\vartheta^2(X_1).$$

Az $\overline{X}$ mintaátlag várható értékét az előző állítás szerint ismerjük, ez $\mathbb{E}_\vartheta(X_1)$. Így, a mintaátlagra alkalmazva a szórásnégyzet definícióját:

$$\mathbb{E}_\vartheta\left(\overline{X}^2\right) = D_\vartheta^2(\overline{X}^2) + \mathbb{E}_\vartheta(\overline{X})^2 = \frac{1}{n^2} D_\vartheta^2(X_1) + \mathbb{E}_\vartheta(X_1)^2.$$

Mindezeket összerakva:

$$\mathbb{E}_\vartheta(s_n^{*2}) = \frac{n}{n-1}\left[D_\vartheta^2(X_1) + \mathbb{E}_\vartheta(X_1)^2\right] - \frac{n}{n-1}\left[\frac{1}{n} D_\vartheta^2(X_1) + \mathbb{E}_\vartheta(X_1)^2\right] = D_\vartheta^2(X_1).$$

Azaz a korrigált tapasztalati szórásnégyzet a szórásnégyzet torzítatlan becslése. □

5.4. definíció (Hatásosság). *Legyenek T_1, T_2 torzítatlan becslései a paraméter $\psi(\vartheta)$ függvényének. Azt mondjuk, hogy T_1 hatásosabb T_2-nél, ha $D_\vartheta^2(T_1) \leq D_\vartheta^2(T_2)$ teljesül minden $\vartheta \in \Theta$-ra.*

A T_1 becslés **hatásos** *$\psi(\vartheta)$-ra, ha $\psi(\vartheta)$ minden torzítatlan becslésénél hatásosabb (és ő maga is torzítatlan).*

Előfordul, hogy két torzítatlan becslés közül egyik sem hatásosabb a másiknál, azaz van két különböző ϑ, amelyiknél eltér, hogy melyiknek kisebb a szórása a $\mathbb{P}_\vartheta$ mérték szerint. Nem mindig létezik hatásos becslés, viszont ha létezik, akkor lényegében egyértelmű (pontosabban, ha T_1 és T_2 hatásos becslések $\psi(\vartheta)$-ra, akkor 1 valószínűséggel megegyeznek).

5.5. állítás. *Legyen $(X_1, \ldots, X_n)$ független azonos eloszlású minta véges szórású eloszlásból. Ekkor $\psi(\vartheta) = \mathbb{E}_\vartheta(X_i)$-re a mintaátlag hatásosabb minden $\sum_{j=1}^{n} c_j X_j$ alakú becslésnél, ahol $0 \leq c_j$ és $\sum_{j=1}^{n} c_j = 1$.*

Az állítás a számtani és négyzetes közepek közötti egyenlőtlenségből adódik. Ugyanakkor a mintaátlag nem minden esetben hatásos becslése a várható értéknek, csak a lineáris kombinációknál hatásosabb.

5.2. Aszimptotikus torzítatlanság és konzisztencia

Tekinthetjük statisztikák egy sorozatát úgy, hogy az n. statisztika az első n mérési adattól függ. Például: $X_1, X_2, \ldots$ mérési eredmények, és $T_n = \frac{1}{n}(X_1 + \ldots + X_n)$ az első n mérésből kapott adat átlaga.

5.6. definíció. *[1] A $T_n = T_n(X_1, \ldots, X_n)$* **aszimptotikusan torzítatlan** *becsléssorozat $\psi(\vartheta)$-ra, ha minden $\vartheta \in \Theta$-ra*

$$\mathbb{E}_\vartheta(T_n(X_1, \ldots, X_n)) \to \psi(\vartheta) \qquad (n \to \infty).$$

5.7. definíció. *[1] A $T_n = T_n(X_1, \ldots, X_n)$* **konzisztens** *becsléssorozat $\psi(\vartheta)$-ra, ha minden $\vartheta \in \Theta$-ra*

$$(T_n(X_1, \ldots, X_n)) \to \psi(\vartheta)$$

$n \to \infty$ esetén sztochasztikusan, azaz minden $\vartheta \in \Theta$ és $\varepsilon > 0$-ra teljesül, hogy

$$\mathbb{P}_\vartheta\big(|T_n - \psi(\vartheta)| > \varepsilon\big) \to 0 \qquad (n \to \infty).$$

A nagy számok gyenge törvénye alapján a $\psi(\vartheta) = \mathbb{E}_\vartheta(X_1)$ függvényre a $T_n = \frac{X_1 + \ldots + X_n}{n}$ becsléssorozat konzisztens. Sőt a nagy számok erős törvénye alapján $T_n \to \psi(\vartheta)$ 1 valószínűséggel is teljesül minden $\vartheta \in \Theta$-ra $n \to \infty$ esetén.

6. Elégséges statisztikák

6.1. definíció (Diszkrét eset, [1]). *Legyen $\underline{X} = (X_1, X_2, \ldots, X_n)$ diszkrét minta (azaz tegyük fel, hogy a H mintatér véges vagy megszámlálhatóan végtelen). A $T(\underline{X})$ statisztika* **elégséges**, *ha minden $\underline{x} \in H, t \in T(H)$ párra igaz, hogy a $\mathbb{P}_\vartheta(\underline{X} = \underline{x} | T(\underline{X}) = t))$ feltételes valószínűség nem függ ϑ-tól.*

6.2. definíció (Abszolút folytonos eset, [1]). *Legyen $\underline{X}$ független minta. Tegyük fel, hogy az $\underline{X} = (X_1, \ldots, X_n)$ minta eloszlása abszolút folytonos, együttes sűrűségfüggvénye $f_{n,\vartheta}$. A $T : H \to \mathbb{R}$ statisztika* **elégséges**, *ha az együttes sűrűségfüggvény felírható*

$$f_{n,\vartheta}(y_1, \ldots, y_n) = h(y_1, \ldots, y_n) \cdot g_\vartheta(T(y_1, \ldots, y_n))$$

alakban minden $\vartheta \in \Theta$-ra, valamely h és g_ϑ függvényekre.

Független azonos eloszlású minta esetén a rendezett minta (az adatok sorbarendezésével kapott adatsor) elégséges statisztika.

7. Maximumlikelihood-módszer

7.1. definíció (Likelihood-függvény). *Legyen $Y_1, \dots, Y_n$ minta. Ha ezek abszolút folytonosak, és Y_j sűrűségfüggvénye (a $\mathbb{P}_\vartheta$-re vonatkozóan) $f_{j,\vartheta}$, akkor a minta likelihood-függvénye:*

$$L_{n,\vartheta}(t_1, \dots, t_n) = \prod_{j=1}^{n} f_{j,\vartheta}(t_j) \qquad (t_1, \dots, t_n \in \mathbb{R}).$$

Ha a minta diszkrét, akkor a minta likelihood-függvénye:

$$L_{n,\vartheta}(k_1, \dots, k_n) = \prod_{j=1}^{n} \mathbb{P}_{j,\vartheta}(Y_j = k_j) \qquad ((k_1, \dots, k_n) \in H).$$

7.2. definíció (Maximum-likelihood becslés). *A ϑ maximumlikelihood-becslése (ML-becslése) az $X_1, \dots, X_n$ mintából $\hat{\vartheta}$, ha $\hat{\vartheta}$ maximalizálja a $\vartheta \mapsto L_{n,\vartheta}(X_1, \dots, X_n)$ függvényt, ahol $L_{n,\vartheta}$ a minta likelihood-függvénye. Azaz, ha*

$$L_{n,\hat{\vartheta}}(X_1, \dots, X_n) \geq L_{n,\vartheta}(X_1, \dots, X_n) \text{ minden } \vartheta \in \Theta\text{-ra.}$$

A maximumlikelihood-becslés tulajdonságai

- Nem minden statisztikai mezőn létezik ML-becslés.
- Az ML-becslés nem feltétlenül egyértelmű.
- Ha létezik ML-becslés, T pedig elégséges statisztika, akkor az ML-becslés felírható $h(T(X_1, \dots, X_n))$ alakban valamely h függvényre.
- A $\psi(\vartheta)$ függvény ML-becslése $\psi(\hat{\vartheta})$, ahol $\hat{\vartheta}$ ML-becslés ϑ-ra.
- Megfelelő feltételek (erős regularitási feltételek mellett) az ML-becslés aszimpotikusan torzítatlan, és aszimptotikusan normális eloszlású, azaz $\sqrt{n}(\hat{\vartheta}_n - \vartheta)$ normális eloszláshoz konvergál eloszlásban $n \to \infty$ esetén (a $\mathbb{P}_\vartheta$ valószínűségre vonatkozóan).
- Az alábbi egyenlet a maximumlikelihood-egyenlet:

  $$\frac{\partial}{\partial \vartheta} \ln L_{n,\vartheta}(X_1, \dots, X_n) = 0.$$

  Megfelelő feltételek mellett az ML-becslés a maximumlikelihood-egyenlet megoldása (ha az ML-becslés nem számítható ki, de az egyenlet megoldható, gyakran az egyenlet megoldásával helyettesítik az ML-becslést).

8. Momentummódszer

Legyen $X_1, \dots, X_n$ független azonos eloszlású minta, $(\Omega, \mathcal{A}, \mathcal{P})$ pedig statisztikai mező, $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$. Bizonyos esetekben alkalmazható az alábbi eljárás.

1. Az eloszlás k. momentuma: $\mu_{k,\vartheta} = \mathbb{E}_\vartheta(X_1^k)$.
2. Legyen $\hat{\mu}_k = \frac{1}{n}\sum_{j=1}^n X_j^k$ az eloszlás k. tapasztalati momentuma.
3. Írjuk fel az alábbi egyenleteket a legkisebb olyan k-ig, amire igaz, hogy az egyenletrendszer egyértelműen meghatározza ϑ-t:
$$\mathbb{E}_\vartheta(X_1) = \frac{1}{n}\sum_{j=1}^n X_j;$$
$$\mathbb{E}_\vartheta(X_1^2) = \frac{1}{n}\sum_{j=1}^n X_j^2;$$
$$\dots$$
$$\mathbb{E}_\vartheta(X_1^k) = \frac{1}{n}\sum_{j=1}^n X_j^k.$$
4. A ϑ momentummódszerrel kapott becslése az a $\hat{\vartheta}$, ami megoldása a fenti egyenletrendszernek.

A momentummódszerrel kapott becslés nem biztos, hogy létezik, és nem biztos, hogy egyértelmű.

9. Konfidenciaintervallumok

Legyen $\underline{X} = (X_1, \dots, X_n)$ független azonos eloszlású minta, $(\Omega, \mathcal{A}, \mathcal{P})$ pedig statisztikai mező, $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$, és tegyük fel, hogy ϑ valós paraméter, vagyis $\Theta \subseteq \mathbb{R}$.

9.1. definíció. *Azt mondjuk, hogy a $(T_1(\underline{X}), T_2(\underline{X}))$ intervallum legalább $1-\alpha$ megbízhatósági szintű konfidenciaintervallum ϑ-ra, ha minden $\vartheta \in \mathbb{R}$ esetén teljesül, hogy*
$$\mathbb{P}_\vartheta(T_1(\underline{X}) < \vartheta < T_2(\underline{X})) \geq 1 - \alpha.$$

A konfidenciaintervallum megbízhatósági szintje: $\inf_{\vartheta\in\Theta}\{\mathbb{P}_\vartheta(\vartheta \in (T_1, T_2))\}$.

A várható értékre normális eloszlás esetén tudunk könnyen konfidenciaintervallumot adni. (A centrális határeloszlástétel alapján nagy mintaelemszám esetén alkalmazható lehet a normális eloszlással való közelítés.)

A következő jelölést fogjuk használni: ha $q \in [0, 1]$, akkor $u_q = \Phi^{-1}(q)$, ahol Φ a standard normális eloszlás eloszlásfüggvénye. Vagyis, ha Z standard normális eloszlású valószínűségi változó, akkor

$$q = \mathbb{P}(Z \leq u_q) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_q} e^{-s^2/2} ds.$$

9.2. állítás (Konfidenciaintervallum a várható értékre, ismert szórás). *Tegyük fel, hogy $X_1, \ldots, X_n$ független azonos eloszlású normális eloszlású valószínűségi változók, melyek szórása, σ ismert.*

Kétoldali konfidenciaintervallum: *Ekkor a*

$$(T_1, T_2) = \left(\overline{X} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \overline{X} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

intervallum $1 - \alpha$ megbízhatósági szintű konfidenciaintervallum az eloszlás várható értékére.

Egyoldali konfidenciaintervallumok $1 - \alpha$ *megbízhatósági szinttel, jobbról, illetve balról:*

$$\left(-\infty,\ \overline{X} + u_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right); \qquad \left(\overline{X} - u_{1-\alpha} \frac{\sigma}{\sqrt{n}}, \infty\right).$$

9.3. definíció (t-eloszlás). *Legyenek $Z_0, Z_1, \ldots, Z_n$ független standard normális eloszlású valószínűségi változók. Ekkor a*

$$Y = \frac{Z_0}{\sqrt{Z_1^2 + \ldots + Z_n^2}}$$

valószínűségi változó eloszlását n szabadsági fokú t-eloszlásnak nevezzük. Legyen $t_n(q)$ a q-kvantilise, vagyis az a szám, melyre az alábbi teljesül:

$$q = \mathbb{P}(Y \leq t_n(q)) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \ldots + Z_n^2}} \leq t_n(q)\right).$$

9.4. állítás (Konfidenciaintervallum a várható értékre, ismeretlen szórás). *Tegyük fel, hogy $X_1, \ldots, X_n$ független azonos eloszlású normális eloszlású valószínűségi változók (sem a várható értékük, sem a szórásuk nem ismert).*

Kétoldali konfidenciaintervallum: *Ekkor a*

$$(T_1, T_2) = \left(\overline{X} - t_{n-1}\Big(1 - \frac{\alpha}{2}\Big) \cdot \frac{s_n^*}{\sqrt{n}},\ \overline{X} + t_{n-1}\Big(1 - \frac{\alpha}{2}\Big) \cdot \frac{s_n^*}{\sqrt{n}} \right)$$

intervallum $1 - \alpha$ megbízhatósági szintű konfidenciaintervallum az eloszlás várható értékére.

Egyoldali konfidenciaintervallumok $1 - \alpha$ *megbízhatósági szinttel, jobbról, illetve balról:*

$$\left(-\infty,\ \overline{X} + t_{n-1}(1-\alpha) \cdot \frac{s_n^*}{\sqrt{n}} \right); \qquad \left(\overline{X} - t_{n-1}(1-\alpha) \cdot \frac{s_n^*}{\sqrt{n}},\ \infty \right).$$

10. Hipotézisvizsgálat

A hipotézisvizsgálat fő kérdései: lehet-e egy előzetes feltételezést (nullhipotézist) *cáfolni* az adatok alapján? Mennyire tér el a minta a nullhipotézis esetén várható tapasztalati eloszlástól?

10.1. definíció. *Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, azaz $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ valamilyen Θ paramétertérrel. A paraméterteret bontsuk fel két diszjunkt halmaz uniójára: $\Theta = \Theta_0 \cup \Theta_1$, ahol tehát $\Theta_0 \cap \Theta_1 = \emptyset$.*

Nullhipotézis. $H_0 : \vartheta \in \Theta_0$.

Ellenhipotézis. $H_1 : \vartheta \in \Theta_1$.

A minta $\underline{X} = (X_1, \ldots, X_n)$, a mintatér legyen B (vagyis $(X_1, \ldots, X_n)$ a $B \subseteq \mathbb{R}^n$ halmaz egy véletlen eleme). A mintateret is felbontjuk két diszjunkt halmaz uniójára: $B = B_0 \cup B_1$, ahol $B_0 \cap B_1 = \emptyset$.

Elfogadási tartomány: B_0. Ha $(X_1, \ldots, X_n) \in B_0$, akkor H_0-t elfogadjuk.

Elutasítási (kritikus) tartomány: B_1. Ha $(X_1, \ldots, X_n) \in B_1$, akkor H_0-t elutasítjuk.

A döntés értelmezése: ha H_0-t elutasítottuk, az adatok statisztikai bizonyítékot szolgáltattak arra, hogy H_0 nem igaz. Ha H_0-t elfogadjuk: az adatok alapján nem tudjuk H_0-t cáfolni, de arra sincs bizonyíték, hogy igaz lenne.

10.2. definíció.

- *Elsőfajú hibát vétünk, ha H_0 igaz, és elutasítjuk.*

- *A próba terjedelme:*

$$\alpha = \sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta(\underline{X} \in B_1).$$

- *Másodfajú hibát vétünk, ha H_0 nem igaz, és elfogadjuk.*
- *A próba erőfüggvénye az alábbi $\beta : \Theta_1 \to [0,1]$ függvény:*

$$\beta(\vartheta) = \mathbb{P}_\vartheta(\underline{X} \in B_1) \qquad (\vartheta \in \Theta_1).$$

- *p-érték: a legnagyobb olyan terjedelem, ami mellett H_0-t elfogadjuk.*

10.1. A próbák jósága

10.3. definíció. *A próba* torzítatlan, *ha erőfüggvénye legalább akkora, mint a terjedelme:*

$$\beta(\vartheta) \geq \alpha \qquad \textit{minden } \vartheta \in \Theta_1\textit{-re.}$$

A (B_0, B_1) próba egyenletesen erősebb, *mint a (B_0', B_1') próba, ha*

$$\mathbb{P}_\vartheta(\underline{X} \in B_1) \geq \mathbb{P}_\vartheta(\underline{X} \in B_1') \qquad \textit{minden } \vartheta \in \Theta_1\textit{-re.}$$

A $\left(B_0^{(n)}, B_1^{(n)}\right)$ konzisztens próbasorozat, ha

$$\alpha_n \leq \alpha \textit{ minden } n\textit{-re és } \lim_{n\to\infty} \beta_n(\vartheta) = 1 \textit{ minden } \vartheta \in \Theta_1\textit{-re.}$$

Itt α_n az n. próbához tartozó terjedelmet, β_n pedig a hozzá tartozó erőfüggvényt jelenti.

10.2. Neyman–Pearson-lemma

Tegyük fel, hogy a nullhipotézis és az ellenhipotézis is egyetlen paraméterhez tartozik, vagyis: $H_0 : \vartheta = \vartheta_0$; $H_1 : \vartheta = \vartheta_1$.

Legyen ϑ_0 mellett a minta likelihood-függvénye $L_n(0, \underline{x})$, míg ϑ_1 mellett $L_n(1, \underline{x})$. Rögzítsünk egy c pozitív számot és $\gamma \in [0,1]$-t, és végezzük a következő eljárást (egy véletlenített próbát):

- ha $\frac{L_n(1,\underline{X})}{L_n(0,\underline{X})} > c$, akkor elutasítjuk H_0-t;

- ha $\frac{L_n(1,\underline{X})}{L_n(0,\underline{X})} = c$, akkor sorsolást végzünk (a mintától függetlenül), és γ valószínűséggel elutasítjuk H_0-t, különben elfogadjuk;
- ha $\frac{L_n(1,\underline{X})}{L_n(0,\underline{X})} > c$, akkor elfogadjuk H_0-t.

10.4. tétel (Neyman–Pearson-lemma). *(i) Ha adott $0 < \alpha \leq 1$ és a fenti H_0 és H_1 egyszerű hipotézisek, akkor létezik olyan c és γ, hogy a fenti véletlenített próba terjedelme pontosan α.*

(ii) Ha adott c és γ: a fenti véletlenített próba egyenletesen erősebb minden olyan próbánál, melynek terjedelme nem nagyobb a fenti véletlenített próba terjedelménél.

11. A normális eloszlásra vonatkozó próbák

Az alábbi próbák egyenletesen legerősebb próbák a megegyező terjedelmű próbák közül az adott feladatokban.

11.1. Egymintás u-próba

Az u-próba a normális eloszlás várható értékére vonatkozik, ha az eloszlás szórása ismert. Legyenek tehát $X_1, X_2, \ldots, X_n$ független normális eloszlású valószínűségi változók m várható értékkel és σ szórással, ahol m ismeretlen paraméter, σ ismert. Nullhipotézisre több lehetőség van (az m_0 érték adott): $H_0 : m = m_0$, vagy $H_0 : m \leq m_0$, vagy $H_0 : m \geq m_0$.

A próbastatisztika, ami alapján a döntést hozzuk:

$$u = \frac{\overline{X} - m_0}{\sigma} \cdot \sqrt{n}.$$

Ezt egy úgynevezett *kritikus értékkel* hasonlítjuk össze, és ez alapján fogadjuk el vagy utasítjuk el a nullhipotézist. A H_0 hipotézis mellett az u statisztika standard normális eloszlású. Emlékeztetőül: ha $q \in [0,1]$, akkor $u_q = \Phi^{-1}(q)$, ahol Φ a standard normális eloszlás eloszlásfüggvénye.

- Kétoldali ellenhipotézis: $H_0 : m = m_0$; $\quad H_1 : m \neq m_0$.

  Ha $|u| > u_{1-\alpha/2}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

  A p-érték ilyenkor $2 - 2\Phi(|u|)$.

- Egyoldali ellenhipotézis, balról:

  $H_0 : m \leq m_0$; $\quad H_1 : m > m_0$.

  Ha $u > u_{1-\alpha}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

  A p-érték ilyenkor $1 - \Phi(u)$.

- Egyoldali ellenhipotézis, jobbról:

  $H_0 : m \geq m_0$; $\quad H_1 : m < m_0$.

  Ha $u < -u_{1-\alpha}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

  A p-érték ilyenkor $\Phi(u)$.

11.2. Kétmintás u-próba

Legyenek most $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma_1^2)$, $Y_i \sim N(m_2, \sigma_2^2)$. Itt m_1, m_2 ismeretlen paraméterek, σ_1, σ_2 ismertek.

A próbastatisztika, ami alapján a döntést hozzuk:

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

A $H_0 : m_1 = m_2$ hipotézis mellett az u statisztika standard normális eloszlású.

- Kétoldali ellenhipotézis: $H_0 : m_1 = m_2$; $\quad H_1 : m_1 \neq m_2$.

  Ha $|u| > u_{1-\alpha/2}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

- Egyoldali ellenhipotézis, balról:

  $H_0 : m_1 \leq m_2$; $\quad H_1 : m_1 > m_2$.

  Ha $u > u_{1-\alpha}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

- Egyoldali ellenhipotézis, jobbról:

  $H_0 : m_1 \geq m_2$; $\quad H_1 : m_1 < m_2$.

  Ha $u < -u_{1-\alpha}$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

11.3. Egymintás t-próba

A t-próba a normális eloszlás várható értékére vonatkozik, ha az eloszlás szórása ismeretlen. Legyenek tehát $X_1, X_2, \ldots, X_n$ független normális eloszlású valószínűségi változók m várható értékkel és σ szórással, ahol m és σ is ismeretlen paraméter. Nullhipotézisre több lehetőség van (az m_0 érték adott): $H_0 : m = m_0$, vagy $H_0 : m \leq m_0$, vagy $H_0 : m \geq m_0$.

A próbastatisztika, ami alapján a döntést hozzuk:

$$t = \frac{\overline{X} - m_0}{s_n^*} \cdot \sqrt{n},$$

ahol $s_n^* = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2}$. A $H_0 : m = m_0$ hipotézis mellett a t statisztika $n-1$ szabadsági fokú t-eloszlású. Emlékeztetőül: legyen $t_n(q)$ a q-kvantilise, vagyis az a szám, melyre az alábbi teljesül:

$$q = \mathbb{P}(Y \leq t_n(q)) = \mathbb{P}\left(\frac{Z_0}{\sqrt{Z_1^2 + \ldots + Z_n^2}} \leq t_n(q) \right),$$

ahol $Z_0, Z_1, \ldots, Z_n$ független standard normális eloszlásúak.

- Kétoldali ellenhipotézis: $H_0 : m = m_0$; $\quad H_1 : m \neq m_0$.

  Ha $|t| > t_{n-1}(1 - \alpha/2)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

- Egyoldali ellenhipotézis, balról:

  $H_0 : m \leq m_0$; $\quad H_1 : m > m_0$.

  Ha $t > t_{m-1}(1 - \alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

- Egyoldali ellenhipotézis, jobbról:

  $H_0 : m \geq m_0$; $\quad H_1 : m < m_0$.

  Ha $t < -t_{n-1}(1-\alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

11.4. Kétmintás t-próba

Legyenek most $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású, azonos szórású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma^2)$, $Y_i \sim N(m_2, \sigma^2)$. Itt m_1, m_2, σ ismeretlen paraméterek.

A próbastatisztika, ami alapján a döntést hozzuk:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1 - 1)s^{*2}_{n_1}(X) + (n_2 - 1)s^{*2}_{n_2}(Y)}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}.$$

A $H_0 : m_1 = m_2$ hipotézis mellett a t statisztika $n_1 + n_2 - 2$ szabadsági fokú t-eloszlású.

- Kétoldali ellenhipotézis: $H_0 : m_1 = m_2$; $\quad H_1 : m_1 \neq m_2$.

  Ha $|t| > t_{n_1+n_2-2}(1 - \alpha/2)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

- Egyoldali ellenhipotézis, balról:

  $H_0 : m_1 \leq m_2$; $\quad H_1 : m_1 > m_2$.

  Ha $t > t_{n_1+n_2-2}(1 - \alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

- Egyoldali ellenhipotézis, jobbról:

  $H_0 : m_1 \geq m_2$; $\quad H_1 : m_1 < m_2$.

  Ha $t < -t_{n_1+n_2-2}(1 - \alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

Feltételeztük, hogy a két minta szórása megegyezik. Ezt (a kétmintás t-próba elvégzése előtt) gyakran az alábbi F-próbával ellenőrzik. Ha a két szórás szignifikánsan eltér, más módszerekre lehet szükség.

11.5. F-próba

Az F-próba független normális eloszlású minták szórását hasonlítja össze. Legyenek most $X_1, X_2, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ független normális eloszlású valószínűségi változók, ahol $X_i \sim N(m_1, \sigma_1^2)$, $Y_i \sim N(m_2, \sigma_2^2)$. Itt $m_1, m_2, \sigma_1, \sigma_2$ ismeretlen paraméterek.

A próbastatisztika, ami alapján a döntést hozzuk:

$$F = \frac{s^{*2}_{n_1}}{s^{*2}_{n_2}}.$$

A $H_0 : m_1 = m_2$ hipotézis mellett a F statisztika $d_1 = n_1 - 1$ és $d_2 = n_2 - 1$ szabadsági fokokkal. Az F-eloszlás definíciója: ha $U_1, \ldots, U_{d_1}, V_1, \ldots, V_{d_2}$

független standard normális eloszlású valószínűségi változók, akkor az alábbi hányados F-eloszlású d_1 és d_2 szabadsági fokokkal:

$$\frac{d_2(U_1^2 + U_2^2 + \ldots + U_{d_1}^2)}{d_1(V_1^2 + V_2^2 + \ldots + V_{d_2}^2)}.$$

Legyen $F_{d_1,d_2}(q)$ az F-eloszlás q-kvantilise, vagyis az a szám, melyre $q = \mathbb{P}(W \leq F_{d_1,d_2}(q))$ teljesül, ha a W valószínűségi változó eloszlása F-eloszlás d_1 és d_2 szabadsági fokokkal.

- Kétoldali ellenhipotézis: $H_0 : \sigma_1 = \sigma_2$; $H_1 : \sigma_1 \neq \sigma_2$.

  Ha $F > F_{d_1,d_2}(1-\alpha/2)$ vagy $F < F_{d_1,d_2}(\alpha/2)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

- Egyoldali ellenhipotézis, balról:

  $H_0 : \sigma_1 \leq \sigma_2$; $H_1 : \sigma_1 > \sigma_2$.

  Ha $F > F_{d_1,d_2}(1-\alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

- Egyoldali ellenhipotézis, jobbról:

  $H_0 : \sigma_1 \geq \sigma_2$; $H_1 : \sigma_1 < \sigma_2$.

  Ha $F < F_{d_1,d_2}(\alpha)$, akkor elvetjük a nullhipotézist, különben elfogadjuk.

12. χ^2-próbák

12.1. Illeszkedésvizsgálat

Legyen $A_1, A_2, \ldots, A_r$ teljes eseményrendszer, $p_1, p_2, \ldots, p_r$ pedig olyan nemnegatív számok, melyek összege 1.

$H_0 : \mathbb{P}(A_i) = p_i$ minden $i = 1, 2, \ldots, r$-re.

$H_1 : \mathbb{P}(A_i) \neq p_i$ valamelyik $i = 1, 2, \ldots, r$-re.

n független megfigyelést végzünk, jelölje N_i, hogy hányszor következett be A_i. Ha van olyan N_i, mely 4-nél kevesebb: néhány eseményt össze kell vonnunk, hogy a próbát alkalmazhassuk (vagyis A_i és A_j helyett $A_i \cup A_j$-t és $p_1 + p_2$-t tekintjük). Számítsuk ki az alábbi mennyiséget:

$$T = \sum_{i=1}^{r} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}.$$

χ^2-próba: H_0-t elfogadjuk, ha T kisebb az $f = r-1$ szabadsági fokú, α terjedelmű χ^2-próba c kritikus értékénél. A c kritikus értéket így definiálhatjuk:

$$\mathbb{P}(Z_1^2 + Z_2^2 + \ldots + Z_f^2 < c) = 1 - \alpha),$$

ahol $Z_1, \ldots, Z_f$ független standard normális eloszlású valószínűségi változók.

Példa: $r = 6$, dobókockával dobunk, A_i: a dobás értéke i. Legyen $p_1 = p_2 = \ldots = p_6 = 1/6$, vagyis a nullhipotézis az, hogy szabályos a dobókocka. A próba terjedelmének $\alpha = 0,05$-öt választjuk. $n = 100$ dobásból az alábbi értékek adódtak:

érték	1	2	3	4	5	6
gyakoriság	21	11	20	22	11	15

```
Chi-squared test for given probabilities

data:  kocka1

X-squared = 7.52, df = 5, p-value = 0.1847
```

Ekkor $T = 7,52 < c = 11,1$, tehát elfogadjuk azt a nullhipotézist, hogy a dobókocka szabályos. A p-érték $0,1847 > 0,05$, tehát nincs szignifikáns eltérés a szabályossághoz képest. (Minden szám legalább 4-szer előfordult, nem kell a beosztáson módosítani.)

Ha ezerszer dobunk, és az alábbi eredmények adódnak:

érték	1	2	3	4	5	6
gyakoriság	191	154	140	184	156	175

```
Chi-squared test for given probabilities

data:  kocka2

X-squared = 11.684, df = 5, p-value = 0.03938
```

Továbbra is $\alpha = 0,05$ terjedelem mellett számolva: $T = 11,684 > c = 11,1$, tehát elutasítjuk a nullhipotézist, statisztikai bizonyítékunk van arra, hogy a dobókocka nem szabályos. A p-érték $0,03938 < 0,05$, szignifikáns eltérés van a szabályossághoz képest.

12.2. Becsléses illeszkedésvizsgálat

Továbbra is $A_1, A_2, \ldots, A_r$ teljes eseményrendszer, n elemű független mintánk van, és N_i jelöli, hogy a hányszor következik be A_i. Minden $s \in S \subseteq \mathbb{R}^d$-re adottak $p_1(s), p_2(s), \ldots, p_r(s)$ nemnegatív számok, melyek összege 1.

H_0: van olyan $s \in S$, melyre $\mathbb{P}(A_i) = p_i(s)$ minden $r = 1, 2, \ldots, r$-re.

H_1: nincs olyan $s \in S$, melyre $\mathbb{P}(A_i) = p_i(s)$ minden $r = 1, 2, \ldots, r$-re teljesülne.

Az s paramétervektor (d dimenziós) maximumlikelihood-becslése legyen $\hat{s}$, és legyen $\hat{p}_i = p_i(\hat{s})$. Számítsuk ki az alábbi mennyiséget:

$$T = \sum_{i=1}^{r} \frac{(N_i - n \cdot \hat{p}_i)^2}{n \cdot \hat{p}_i}.$$

Legyen $f = r - d - 1$. A H_0-t α terjedelem mellett elfogadjuk, ha $T < c$, ahol c az f szabadsági fokú kritikus értéke α terjedelem mellett. H_0-t elutasítjuk, ha $T > c$, ilyenkor a minta szignifikánsan eltér az S által megadott eloszláscsaládtól.

Példa. Az egy futballmérkőzésen lőtt gólok száma a világbajnokság 95 mérkőzésén:

gólok száma	0	1	2	3	4	5	6	7	8
mérkőzések száma	23	37	20	11	2	1	0	0	1

Poisson-esetben az s paraméter maximumlikelihood-becslése:

$$\hat{s} = \overline{X} = \frac{0 \cdot 23 + 1 \cdot 37 + 2 \cdot 20 + 3 \cdot 11 + 4 \cdot 2 + 5 \cdot 1 + 8 \cdot 1}{95} = 1,379.$$

Mivel vannak olyan osztályok, ahova 4-nél kevesebb megfigyelés esik, a beosztást módosítjuk:

gólok száma	0	1	2	3	≥ 4
mérkőzések száma	23	37	20	11	4
Poisson($\hat{p}$)-eloszlás	23,92	32,99	22,75	10,46	4,88

H_0: az eloszlás Poisson-eloszlásból származik, valamely $s > 0$ paraméterrel (most $d = 1$).

H_1: az eloszlás nem Poisson-eloszlás.

Ebben az esetben $T = 1,04$, $f = 5 - 1 - 1 = 3$, a kritikus érték $7,81$. Tehát $T < c$, elfogadjuk, hogy a minta Poisson-eloszlásból származik.

12.3. Függetlenségvizsgálat

Két szempont szerint soroljuk osztályokba a megfigyeléseket. Az első szempont szerint r osztály van: $A_1, \ldots, A_r$. A második szempont szerint s osztály van: $B_1, \ldots, B_s$.

H_0: a két szempont független egymástól, azaz $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(B_j)$ minden i, j-re.

H_1: a nullhipotézis nem igaz, a két szempont összefügg.

Jelölje N_{ij} azt, hogy hány olyan megfigyelés van, melyre A_i és B_j teljesül. Legyen továbbá $N_{i\cdot} = \sum_{j=1}^{s} N_{ij}$ (azaz az A_i gyakorisága); $N_{\cdot j} = \sum_{i=1}^{r} N_{ij}$ (azaz B_j gyakorisága); n pedig az összes megfigyelés száma. Ekkor a próbastatisztika:

$$T = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(N_{ij} - \frac{N_{i\cdot} N_{\cdot j}}{n}\right)^2}{\frac{N_{i\cdot} N_{\cdot j}}{n}}.$$

A szabadsági fok $f = (r-1)(s-1)$. Legyen c az f szabadsági fokú χ^2-próba kritikus értéke α terjedelem mellett. A próba: ha $T < c$ (azaz a p-érték nagyobb a terjedelmél), akkor elfogadjuk H_0-t, nem találtunk szignifikáns összefüggést a szempontok között. Ha $T > c$ (azaz a p-érték kisebb a terjedelemnél), akkor elutasítjuk H_0-t, az adatok szignifikáns összefüggést mutatnak.

Ha $r = s = 2$, a próbastatisztika az alábbi egyszerűbb alakra hozható:

$$T = \frac{n\left(N_{11}N_{22} - N_{12}N_{21}\right)^2}{N_{1\cdot} N_{2\cdot} N_{\cdot 1} N_{\cdot 2}}.$$

12.4. Homogenitásvizsgálat

Legyenek X, Y valószínűségi változók. A valós számok halmazát bontsuk fel diszjunkt halmazok uniójára: $A_1, \ldots, A_r$.

H_0: az X és Y valószínűségi változók eloszlása megegyezik, azaz $\mathbb{P}(X \in A_i) = \mathbb{P}(Y \in A_i)$ minden $i = 1, 2, \ldots, r$-re.

H_1: az X és Y valószínűségi változók eloszlás eltérő, azaz van legalább egy i, melyre $\mathbb{P}(X \in A_i) \neq \mathbb{P}(Y \in A_i)$.

Legyen $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ független minta úgy, hogy $X_1, \ldots, X_n$ eloszlása X eloszlása, $Y_1, \ldots, Y_n$ eloszlása Y eloszlása. Legyen N_i az A_i gyakorisága az $\underline{X}$ mintában (azaz hányszor fordul elő, hogy X_k az A_i-be esik, és M_i az A_i gyakorisága az $\underline{Y}$ mintában. A próbastatisztika:

$$T = \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \cdot n \cdot m.$$

A szabadsági fok: $f = r - 1$. Legyen c az f szabadsági fokú χ^2-próba kritikus értéke α terjedelem mellett. A próba: ha $T < c$ (azaz a p-érték nagyobb

a terjedelmél), akkor elfogadjuk H_0-t, nem találtunk szignifikáns eltérést az eloszlások között. Ha $T > c$ (azaz a p-érték kisebb a terjedelemnél), akkor elutasítjuk H_0-t, az adatok szignifikáns eltérést mutatnak az eloszlások között.

13. Lineáris modell

13.1. állítás (Lineáris regresszió). *Legyenek* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *adott számpárok. Azokat az* a *és* b *együtthatókat keressük, melyre a*

$$h^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

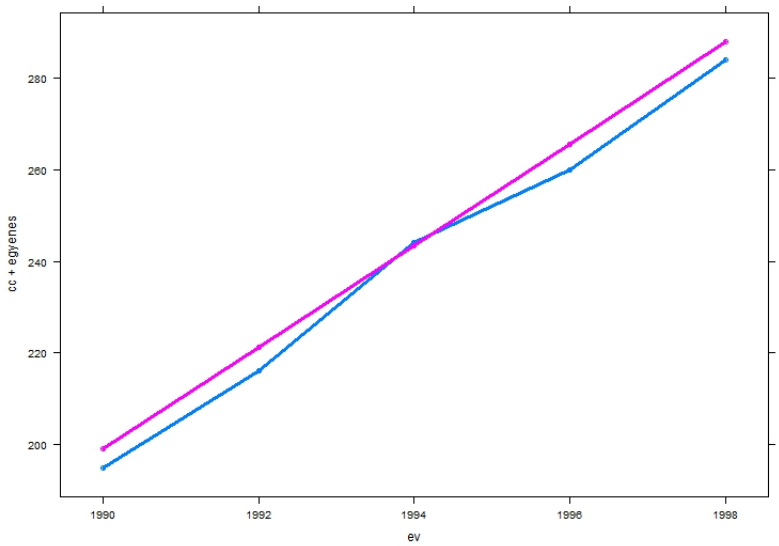
mennyiség minimális. Ennek megoldása:

$$\hat{a} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{k=1}^{n} (x_k - \overline{x})^2}; \qquad \hat{b} = \overline{y} - \hat{a}\overline{x}.$$

Példa: a CFC-12 gáz koncentrációja az Antarktiszon (a gáz gyártását 1996-ban tiltották be).

év	1990	1992	1994	1996	1998
koncentráció (ppm)	195	216	244	260	284

```
Call:

lm(formula = cc   ev, data = f12)

Residuals:
  1     2     3     4     5
−0.4  −1.6   4.2  −2.0  −0.2

Coefficients:
              Estimate   Std.  Error  t value    Pr(>|t|)
(Intercept)  −2.189e+04   8.991e+02   −24.35   0.000152 ***
ev            1.110e+01   4.509e−01    24.62   0.000147 ***
---

Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error:  2.852 on 3 degrees of freedom

Multiple R-squared:  0.9951, Adjusted R-squared:  0.9934

F-statistic:  606 on 1 and 3 DF, p-value:  0.000147
```



9. ábra.
A CFC-12 (freon) gáz koncentrációja az Antarktiszon és az adatokra illesztett egyenes

13.2. definíció (Lineáris modell). *Legyenek $X_1, X_2, \ldots, X_n, Y_1, \ldots, Y_n$ valószínűségi változók, és tegyük fel, hogy valamely a, b valós számokra*

$$Y_i = aX_i + b + \varepsilon_i,$$

ahol $\varepsilon_1, \ldots, \varepsilon_n$ független $N(0, \sigma^2)$ eloszlású valószínűségi változók. Az így kapott (X_i, Y_i) párok együttes eloszlását lineáris modellnek nevezzük. Az X_i valószínűségi változókat magyarázó változóknak, az ε_i valószínűségi változókat hibának szokták nevezni.

13.3. állítás (Becslések a lineáris modellben). *A lineáris modellben az a, b együtthatók ML-likelihood becslése a következőképpen írható:*

$$\hat{a} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{k=1}^n (X_k - \overline{X})^2}; \qquad \hat{b} = \overline{Y} - \hat{a}\overline{X}.$$

Továbbá, ezek a becslések torzítatlan becslései az a és b paramétereknek. A hiba szórásának becslése (ez torzítatlan becslés σ-ra):

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{j=1}^n (Y_i - \hat{a}X_i - \hat{b})^2.$$

A becslések szórása:

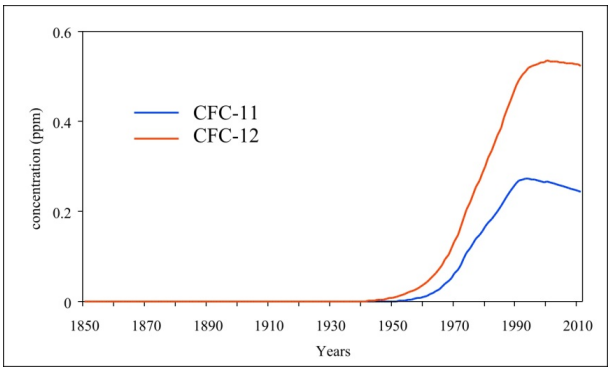
$$D(\hat{a}) = \frac{\sigma}{\sum_{j=1}^{n}(X_j - \overline{X})^2}; \qquad D(\hat{b}) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}}.$$

13.4. állítás (Előrejelzés a lineáris modellben). *Legyen x^* adott szám. A lineáris modellből kapott előrejelzés az Y véletlen folyamat x^* pontban felvett értékére:*

$$\hat{a}x^* + \hat{b}.$$

Az előrejelzés szórása:

$$D(\hat{a}x^* + \hat{b}) = \sigma\sqrt{\frac{1}{n} + \frac{(x^* - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}}.$$



10. ábra.
A CFC-11 és CFC-12 (freon) gáz koncentrációja (forrás: elte.promt.hu)

Az előrejelzés szórásának becslésekor a σ értéket gyakran $\hat{\sigma}$-val helyettesítik.

A teljes ingadozás (total sum of squares): $\sum_{j=1}^{n}(Y_j - \overline{Y})^2$.

Reziduális négyzetösszeg (residual sum of squares):

$$\sum_{j=1}^{n}(Y_j - \hat{a}X_j - \hat{b})^2 = \frac{\left[\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})\right]^2}{\sum_{k=1}^{n}(X_k - \overline{X})^2}.$$

13.5. definíció. *A megmagyarázott ingadozás részaránya (coefficient of determination):*

$$R^2 = \frac{\left[\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})\right]^2}{\left[\sum_{k=1}^{n}(X_k - \overline{X})^2\right]\left[\sum_{k=1}^{n}(Y_k - \overline{Y})^2\right]}.$$

Az R^2 értéke 0 és 1 közé esik. Értelmezés: minél közelebb van 1-hez, annál inkább jó közelítést ad a lineáris modell. Ugyanakkor R érzékeny a kiugró értékekre.

13.1. Az egyenes meredeksége

A lineáris tag együtthatójára vonatkozó hipotézisvizsgálati feladat a következő:

$H_0 : a = 0$

$H_1 : a \neq 0$, vagy $H_1 : a > 0$ vagy $H_1 : a < 0$.

A nullhipotézis mellett az alábbi mennyiség $n-2$ szabadsági fokú t-eloszlású :

$$t = \hat{a}\frac{\sqrt{(n-2)\sum_{i=1}^{n}(X_i - \overline{X})^2}}{\sqrt{\sum_{i=1}^{n}(Y_i - \hat{a}X_i - \hat{b})^2}}.$$

Tehát α terjedelem mellett az alábbi próbát végezhetjük (a definíciók a 11.3. részben szerepeltek).

- Kétoldali ellenhipotézis, $H_1 : a \neq 0$. Ha $|t| > t_{n-2}(1-\alpha/2)$, akkor elutasítjuk H_0-t (az együttható szignifikánsan eltér 0-tól), különben elfogadjuk.
- Egyoldali ellenhipotézis, $H_1 : a > 0$. Ha $t > t_{n-2}(1-\alpha)$, akkor elutasítjuk H_0-t (az együttható szignifikánsan nagyobb 0-nál), különben elfogadjuk.
- Kétoldali ellenhipotézis, $H_1 : a < 0$. Ha $t < t_{n-2}(\alpha)$, akkor elutasítjuk H_0-t (az együttható szignifikánsan kisebb 0-nál), különben elfogadjuk.

$1-\alpha$ megbízhatósági szintű konfidenciaintervallum a-ra:

$$\left(\hat{a} - t_{n-2}(1-\alpha)\frac{\hat{\sigma}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \hat{a} + t_{n-2}(1-\alpha)\frac{\hat{\sigma}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right).$$

13.2. Előrejelzés

Ahogyan korábban láttuk, az x^* pontban az előrejelzett érték becslése $\hat{a}\cdot x^* + \hat{b}$.

$1-\alpha$ megbízhatósági szintű konfidenciaintervallum ax^*+b-re, azaz az x^*-ban felvett érték várható értékére:

$$\left(\hat{a}x^* + \hat{b} \pm t_{n-2}(1-\alpha)\cdot\hat{\sigma}\cdot\sqrt{\frac{1}{n}+\frac{(x^*-\overline{X})^2}{\sum_{i=1}^n (X_i-\overline{X})^2}}\right).$$

$1-\alpha$ megbízhatósági szintű konfidenciaintervallum $ax^*+b+\epsilon(x^*)$-ra, azaz az x^*-ban felvett értékre:

$$\left(\hat{a}x^* + \hat{b} \pm t_{n-2}(1-\alpha)\cdot\hat{\sigma}\cdot\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{X})^2}{\sum_{i=1}^n (X_i-\overline{X})^2}}\right).$$

A konstans tagról azt tudhatjuk, hogy a $b=0$ nullhipotézis esetén a

$$t = \hat{b}\frac{\sqrt{n\sum_{i=1}^n (X_i-\overline{X})^2}}{\hat{\sigma}\sqrt{\sum_{j=1}^n X_j^2}}.$$

Ez alapján szintén lehet hipotézisvizsgálatot végezni az a együttható esetéhez hasonlóan.

Hivatkozások

[1] Csiszár Villő: Statisztika jegyzet. 2009.
`http://www.cs.elte.hu/~villo/esti/stat.pdf`

[2] Móri-Szeidl-Zempléni: Matematikai statisztika példatár. ELTE Eötvös Kiadó, 1997.

[3] John C. Davis: Statistics and data analysis in geology. Wiley, 2002.

[4] E. H. Isaaks and R. M. Srivastava: Applied geostatistics. Oxford University Press, 1989.