Központi/centrális határeloszlás tétel

– egy MATLAB® alapú megközelítés –

Róth Ágoston, Vas Orsolya

Matematika és Informatika Intézet, Babeș-Bolyai Tudományegyetem, Kolozsvár, Románia

(agoston.roth@gmail.com, vas.orsolya@yahoo.com)

8. labor / 2018. november 19–22.

- Statisztikai adatfeldolgozásban és gyakorlati alkalmazásokban a legtöbb esetben azonos eloszlású és független valószínűségi változók összegével dolgozunk.
- A továbbiakban olyan feltételeket ismertetünk, amelyek esetén az adott összeg közelítőleg standard normális eloszlású. Érvényes az alábbi tulajdonság.

1. Tétel (Központi határeloszlás azonos eloszlású és független valószínűségi változók összegére)

Tekintsük az $\{X_i\}_{i\geq 1}$ azonos eloszlású és független valószínűségi változókat, valamint jelölje μ és σ a változók létező, közös várható értékét, illetve szórását, azaz

$\mu = E(X_i)$ és $\sigma = D(X_i)$, $i \geq 1$. Ekkor az $Y_n = \sum\limits_{i=1}^{n} X_i$ valószínűségi változónak a

$$Z_n = \frac{Y_n - E(Y_n)}{D(Y_n)} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma}$$

úgynevezett standardizált alakja aszimptotikusan $\mathcal{N}(0,1)$-eloszlású, vagyis

$$\lim_{n\to\infty} F_{Z_n}(x) = \lim_{n\to\infty} P(Z_n < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt = F_{\mathcal{N}(0,1)}(x), \ \forall x \in \mathbb{R}. \quad (1)$$

Bizonyítás

Lásd a nyomtatott jegyzet 82–83., illetve az elektronikus változat 89–90. oldalait!

1. Megjegyzés

Az 1. tétel jelöléseinél és feltételeinél maradva, a következő észrevételeket tehetjük. Rögzített, de eléggé nagy n értékre az (1)-es határértéket a

$$F_{Z_n}(x) = P(Z_n < x) \approx F_{\mathcal{N}(0,1)}(x),\ x \in \mathbb{R}$$

közelítő egyenlőséggel helyettesíthetjük, vagy az ezzel ekvivalens

$$F_{Z_n}(x) = P(Z_n < x) = P\left(Y_n < x\sigma\sqrt{n} + n\mu\right) \approx F_{\mathcal{N}(0,1)}(x),\ x \in \mathbb{R}, \tag{2}$$

közelítéssel, ami azt jelenti, hogy az $Y_n = \sum_{i=1}^{n} X_i$ valószínűségi változó közelítőleg $\mathcal{N}\left(n\mu, \sigma\sqrt{n}\right)$-eloszlású. Ezért az 1. tételt az alábbi módon is megfogalmazhatjuk.

2. Tétel

Tekintsük az $\{X_i\}_{i\geq 1}$ azonos eloszlású és független valószínűségi változókat, valamint jelölje μ és σ a változók létező, közös várható értékét, illetve szórását, azaz $\mu = E(X_i)$ és $\sigma = D(X_i)$, $i \geq 1$! Ekkor az $Y_n = \sum_{i=1}^{n} X_i$ valószínűségi változó közelítőleg $E(Y_n) = n\mu$ várható értékű és $D(Y_n) = \sigma\sqrt{n}$ szórású normális eloszlást követ (ahol a közelítés nyilvánvalóan annál jobb, minél nagyobb az n értéke). Az állítást az

$$F_{Y_n}(y) = P(Y_n < y) \approx F_{\mathcal{N}(0,1)}\left(\frac{y - n\mu}{\sigma\sqrt{n}}\right) \tag{3}$$

alakban is megfogalmazhatjuk.

2. Megjegyzés (Optimális mintavétel mérete)

Az 1. tétel jelöléseinél és feltételeinél maradva, a (2)-es képlet alapján felírhatjuk, hogy

$$P\left(\left|\frac{Y_n - n\mu}{\sigma\sqrt{n}}\right| < x\right) \approx F_{\mathcal{N}(0,1)}(x) - F_{\mathcal{N}(0,1)}(-x) = 2F_{\mathcal{N}(0,1)}(x) - 1, \tag{4}$$

amit még a szemléletesebb

$$P\left(\left|\frac{Y_n}{n} - \mu\right| < x\frac{\sigma}{\sqrt{n}}\right) \approx 2F_{\mathcal{N}(0,1)}(x) - 1 \tag{5}$$

alakra is hozhatunk. Így olyan összefüggést kaptunk, amely alkalmas arra, hogy azonos eloszlású és véges szórású független valószínűségi változókból álló sorozat esetében meghatározzuk azt az n számot, amelyre adott $\alpha \in (0,1)$ szignifikanciaszint mellett a változók számtani átlaga $1-\alpha$ valószínűséggel az $\varepsilon > 0$ küszöbhatárnál kevesebbel tér el a változók közös μ várható értékétől.

- Tekintsünk néhány példát a 2. megjegyzésbeli optimális mintavétel meghatározására!

1. feladat

Hány kísérletet kell elvégezni ahhoz, hogy valamely A véletlen esemény relatív gyakorisága $0,95$ valószínűséggel $0,05$-nél kevesebbel térjen el az ismeretlen $p = P(A) \neq 0$ valószínűségtől?

Megoldás

- Az A eseményt az
$$X\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$
$\mathcal{B}ern(p)$-eloszlású valószínűségi változóval írhatjuk le.
- Ekkor egy n hosszúságú független kísérletsorozat az $\{X_i \sim \mathcal{B}ern(p)\}_{i=1}^{n}$ független, azonos eloszlású, $\mu = p$ várható értékű és $\sigma = \sqrt{p(1-p)}$ szórású valószínűségi változókból álló mintavételt jelenti.
- Mivel a p valószínűséget nem ismerjük, ezért a $p(1-p)$ szorzatot annak legnagyobb lehetséges értékével, az $\frac{1}{4}$ számmal, helyettesítjük.
- Vegyük észre, hogy az n kísérlet során az A esemény bekövetkezéseinek számát az
$$Y_n = \sum_{i=1}^{n} X_i$$
valószínűségi változóval reprezentálhatjuk!

Megoldás – folytatás

- Ekkor az 1. tétel minden feltétele teljesül és alkalmazhatjuk a 2. megjegyzésbeli (5)-ös becslést az optimális mintavétel n méretének meghatározására.
- Ezek szerint $1-\alpha = 0,95 \approx 2F_{\mathcal{N}(0,1)}(x) - 1$, ahonnan

$$F_{\mathcal{N}(0,1)}(x) \approx 0,975 \Rightarrow x = F^{-1}_{\mathcal{N}(0,1)}(0,975) \approx 1,959963984540054;$$

a küszöbhatárra pedig az

$$\varepsilon = x\frac{\sigma}{\sqrt{n}} \approx 1,959963984540054 \cdot \frac{\frac{1}{2}}{\sqrt{n}} = \frac{0,979981992270027}{\sqrt{n}} = 0,05$$

feltételt kapjuk, ahonnan az optimális mintavétel nagyságára az

$$n \approx \left[\left(\frac{0,979981992270027}{0,05}\right)^2\right] = [384,145882069412503] = 384$$

közelítő értéket kapjuk.

Megoldás – folytatás

- A valószínűségi változók felépítése mellett a feladat fő megoldáskulcsa az $F^{-1}_{\mathcal{N}(0,1)}(0,975)$ inverz érték meghatározása. Ezt megtehetjük a 3. labor anyagában ismertetett numerikus inverziós algoritmusok valamelyikével, vagy használhatunk egy előre elkészített standard normális eloszlásfüggvény-értékeket tartalmazó táblázatot, illetve használhatjuk a MATLAB® beépített norminv parancsát is, ahogy azt a

```
% the threshold ε
epsilon = 0.05;
% the significance level α
alpha = 1 - 0.95;

% let the standard deviation σ be equal to
% the maximal value of the function √(p(1 − p)), p ∈ [0, 1]
sigma = 1 / 2;
% calculate the probability u = 1 − α/2
u = 1 - alpha / 2;

% determine the inverse value F^{-1}_{N(0,1)}(u)
x = norminv(u, 0, 1);
% calculate the optimal samplesize n = ⌈(xσ/ε)^2⌉
n = round( (x * sigma / epsilon)^2 );
```

kódrészlet mutatja.

2. feladat

Ugyanazon gyártássorozatból származó azonos típusú processzorok órajelét ellenőrizve, nagyszámú megfigyelés után a mérnökök azt tapasztalták, hogy a legyártott processzorok órajele $\mu = 2800$MHz várható értékű és $\sigma = 708$MHz szórású normális eloszlást követ. Ezzel a minőségi szinttel megelégedve, a mérnökök a processzorokat gyártó automatizált gép kalibrálását befejezték. Nyilván elengedhetetlen, hogy további gyártási folyamatok minőségét is ellenőrizzék, viszont mindezt egy optimális méretű mintavétel alapján szeretnék megtenni úgy, hogy elkerüljék a kalibrálás során végigszenvedett hatalmas munkát.

A további gyártási folyamatok során hány processzort kell tanulmányoznunk ahhoz, hogy 90%-os bizonyossággal állíthassuk azt, hogy a megvizsgált processzorok órajeleinek számtani átlaga a fenti ideális várható értékhez képest maximálisan annak 8%-val térjen el?

Megoldás

- Jelölje $\{X_i \sim \mathcal{N}(\mu, \sigma)\}_{i=1}^{n}$ a megvizsgált, egymástól függetlenül legyártott processzszorok órajelét.
- Ekkor az órajelek számtani átlagát az

$$\frac{1}{n} Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

valószínűségi változóval írhatjuk le. Erre a 2. megjegyzés feltételei teljesülenek és az (5)-ös közelítő képlet, valamint a feladat feltételei alapján a

$$\begin{aligned} 0,9 &= 1 - \alpha \\ &= 2F_{\mathcal{N}(0,1)}(x) - 1 \\ &\approx P\left(\left|\frac{Y_n}{n} - \mu\right| < x\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

összefüggésre jutunk, ahol $\mu = 2800$, $\sigma = 708$, a küszöbhatárra pedig az

$$\varepsilon = x\frac{\sigma}{\sqrt{n}} = 0,08 \cdot \mu = 224$$

megszorítást kapjuk.

Megoldás – folytatás

- Mindezek függvényében az

$$x \approx F_{\mathcal{N}(0,1)}^{-1}\left(1-\frac{\alpha}{2}\right) = F_{\mathcal{N}(0,1)}^{-1}(0,95) \approx 1,644853626951473,$$

$$n \approx \left[\left(\frac{x\sigma}{\varepsilon}\right)^2\right] = [27,028689691758689] = 27$$

számértékeket kapjuk, azaz elégséges csak 27 darab – találomra kiválasztott – processzor órajelét ellenőrizni.

1. feladat

A nyers erő (brute-force) módszerét használva, írjatok egy-egy szimulációt az elméletileg megoldott feladatok eredményeinek gyakorlati alátámasztására!

2. feladat (elméleti megoldás is szükséges az implementációhoz)

Felhasználva a centrális határeloszlás tételét, adott $\alpha \in (0, 1)$ szignifikanciaszint (vagyis $1 - \alpha$ valószínűség) mellett szerkesszetek megbízhatósági intervallumot:

- az $m = 10$ és $p \in (0, 1)$ paraméterű binomiális eloszlás ismeretlen p paraméterére;
- az $a = 3$ és $b > 0$ paraméterű Pareto-eloszlás ismeretlen b paraméterére (lásd a 3. labor 1. táblázatának 6. sorát)!

Megjegyzés. Az egyes eloszlású minták generálása során a fenti ismeretlen paramétereket is tekintsétek adottaknak, majd az egyes megbízhatósági intervallumok végpontjainak meghatározása során feltételezzétek azokat ismeretlennek és ellenőrizzétek, hogy az „elfelejtett” paraméterértékek beleesnek-e az általatok szerkesztett megbízhatósági intervallumokba. Mindkét esetben számoljatok legalább 1000-elemű mintavétellel.