

1. ábra. Normális eloszlású, illetve ezer normális eloszlású ($m = 10, \sigma = 3$) valószínűségi változó átlagából álló minta hisztogramja

Valószínűségszámítás előadás, 11. hét, 2021. november 26.

Centrális határeloszlás-tétel és alkalmazásai

1. Az átlag viselkedése

A statisztikában és a valószínűségszámítás alkalmazásaiban kulcsfontosságú az átlag viselkedésének megértése. Azt már tudjuk, hogy megfelelő feltételek mellett az átlag a várható értékhez konvergál, megfelelő értelemben. Azonban az alkalmazásokhoz gyakran ennél pontosabban is meg kell érteni az átlag viselkedését, vagyis azt is kell tudnunk, hogy az átlagnak a várható értéktől való eltérése hogyan viselkedik. Először nézzünk néhány példát. Az alábbi ábrákon a bal oldalon különböző eloszlású minták hisztogramja látható, a jobb oldalon pedig az összesen 100000 megfigyelést ezresével csoportosítottuk, minden ezres csoportban kiszámítottuk az átlagot, és az így kapott száz megfigyelésből készítettünk hisztogramot.

Az első esetben a mintát a saját sűrűségfüggvényével ($N(10, 3^2)$), a második esetben pedig a 10 várható értékű és $3/\sqrt{1000}$ szórású normális eloszlás sűrűségfüggvényével hasonlítottuk össze, annak megfelelően, hogy a korábban tanultak alapján, ha X_1, \dots, X_n független azonos eloszlású, véges szórású valószínűségi változók, akkor

$$\mathbb{E}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \mathbb{E}(X_1); \quad D\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{D(X_1)}{\sqrt{n}}. \quad (1)$$

Vagyis, ha a példában $n = 1000$ darab normális eloszlású valószínűségi változót átlagolunk, melyeknek 10 a várható értéke és 3 a szórása, akkor az átlag várható értéke 10, szórása $3/\sqrt{1000}$. Ellenőrzésképpen: a jobb oldali hisztogramon szereplő minta átlaga: $\bar{x} = 9,99$, korrigált tapasztalati szórása: $s_n^* = 0,084$, míg $\sigma/\sqrt{n} = 3/\sqrt{1000} = 0,095$.

Korrigált tapasztalati szórás:

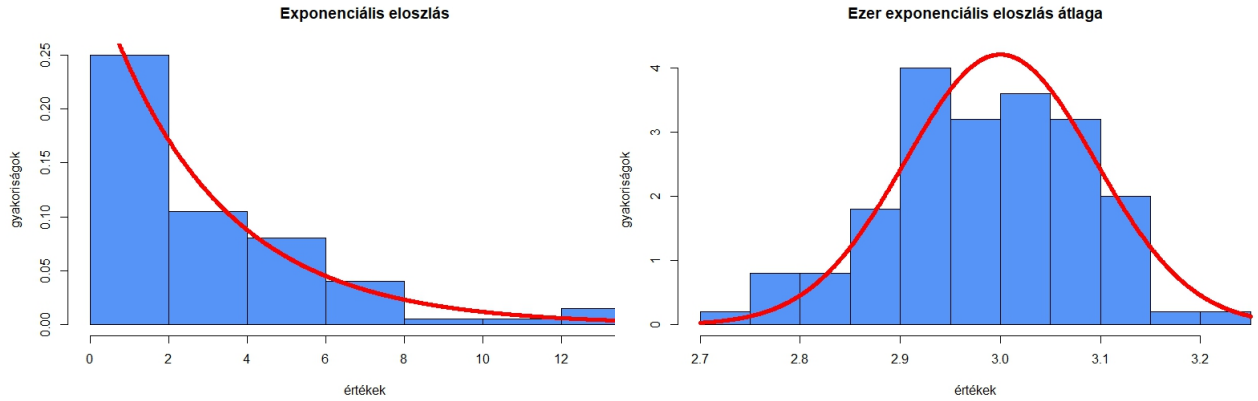
$$s_n^* = \sqrt{\frac{n}{n-1} \left(\left(\frac{1}{n} \sum_{j=1}^n X_j^2 \right) - \bar{X}^2 \right)},$$

ahol $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ a minta átlaga. Az R-ben az `sd` paranccsal kapjuk vissza.

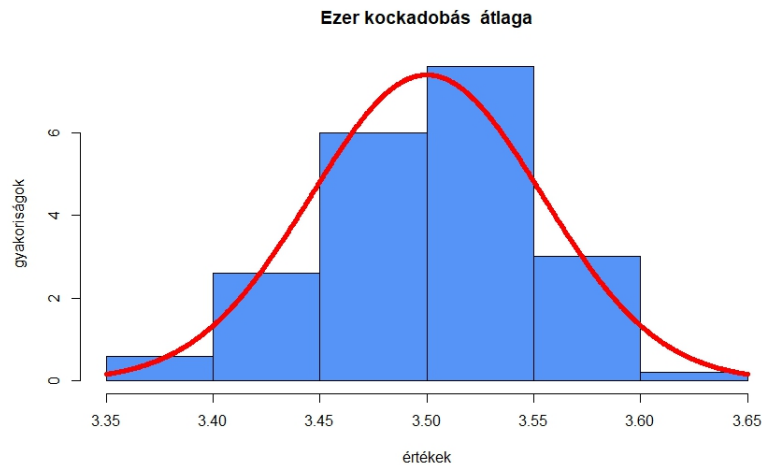
Ez összhangban van az alábbi, korábban látott állításnak az átlagra vonatkozó következményével:

1.1. Állítás. *Legyenek X, Y függetlenek, normális eloszlásúak: $X \sim N(m_1, \sigma_1^2), Y \sim N(m_2, \sigma_2^2)$, ahol a normális eloszlás második paramétere a szórásnégyzetet jelöli. Ekkor a következők igazak:*

- $X + b$ eloszlása normális, $m_1 + b$ várható értékkel és σ szórással;
- aX eloszlása normális am_1 várható értékkel és $|a|\sigma$ szórással;



2. ábra. $\lambda = 1/3$ paraméterű exponenciális eloszlású minta, illetve 1000 exponenciális eloszlású valószínűségi változó átlagaként előálló megfigyelések hisztogramja



3. ábra. Százalemeű minta az alábbi eloszlásból: $n = 1000$ független szabályos kockadobás átlaga, és az $N(3, 5, D(X_1)/\sqrt{1000})$ normális eloszlás sűrűségfüggvénye ($\bar{x} = 3,501, s_n^* = 0,098, \sigma/\sqrt{n} = 0,051$)

- $X + Y$ eloszlása normális, $m_1 + m_2$ várható értékkel és $\sqrt{\sigma_1^2 + \sigma_2^2}$ szórással.

Emlékeztető: $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, és ha X és Y függetlenek, akkor $D^2(X + Y) = D^2(X) + D^2(Y)$.

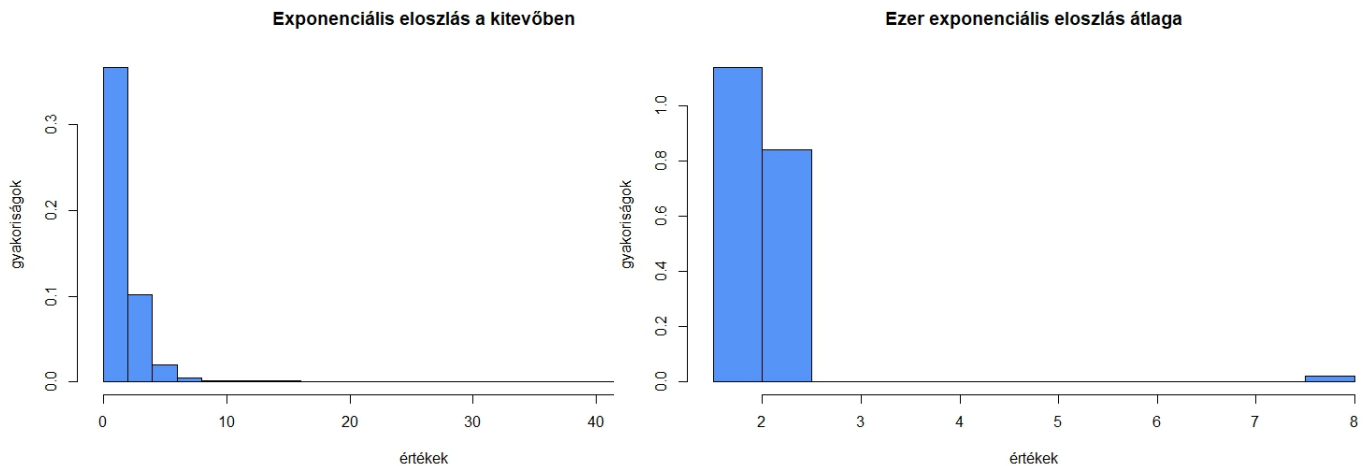
Ebből következik az (1) egyenlet alapján: ha X_1, \dots, X_n független normális eloszlásúak m várható értékkel és σ szórással, akkor

$$\frac{X_1 + \dots + X_n}{n} \sim N\left(m, \frac{\sigma^2}{n}\right),$$

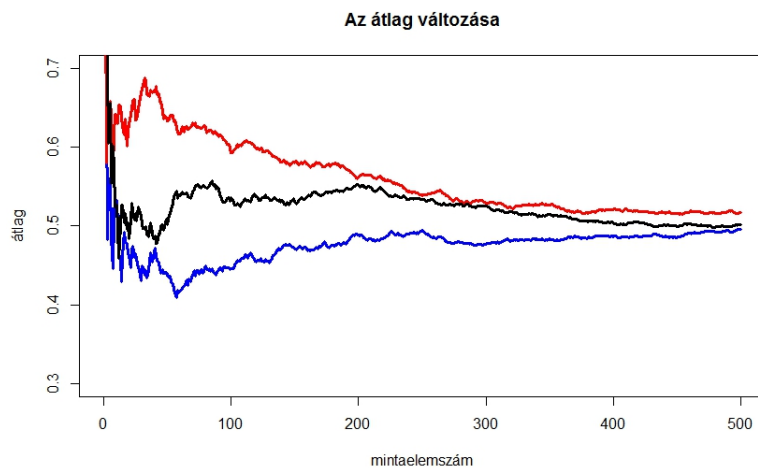
ahol a normális eloszlás második paramétere **a szórásnégyzetet** jelöli.

A 2. ábra hasonlóképpen készült, csak éppen most nem normális, hanem exponenciális eloszlásból indultunk ki. Egy megfigyelés paramétere $\lambda = 1/3$ (a sűrűségfüggvény: $\lambda e^{-\lambda x} \mathbb{I}(x > 0) = \frac{1}{3} e^{-1/3x} \mathbb{I}(x > 0)$), így a várható érték és a szórási is 3. Ezer mintaelemet átlagolva a várható érték 3 marad, a szórási az előzőhöz hasonlóan $3/\sqrt{1000}$ lesz (ellenőrzés: $\bar{x} = 2,98, s_n^* = 0,098$). A hisztogramon a bal oldalon az exponenciális eloszlás sűrűségfüggvénye, a jobb oldalon a 3 várható értékű és $3/\sqrt{1000}$ szórási normális eloszlás sűrűségfüggvénye látható még, elég jó illeszkedést figyelhetünk meg. A 3. ábrán 1000 kockadobás átlagát sorsoltuk ki 100-szor, ennek hisztogramja látható, ez is a normális eloszlás sűrűségfüggvényére illeszkedik.

A 4. ábrán exponenciális eloszlású mintát sorsolunk, azonban nem az X_1, \dots, X_{1000} , hanem az $e^{X_1}, e^{X_2}, \dots, e^{X_{1000}}$ megfigyeléseket tekintjük. Az exponenciális eloszlás eloszlás paramétere 2 volt. Itt is a bal oldalon magának a mintának, a jobb oldalon ezres csoportok átlagának hisztogramja látható.



4. ábra. Az e^X , illetve 1000 darab e^{X_i} átlagának hisztogramja, ahol X exponenciális eloszlású 2 paraméterrel.



5. ábra. A $[0, 1]$ intervallumon egyenletes eloszlásból vett minta átlaga $n = 500$ -ig

Számítsuk ki az e^{X_1} várható értékét:

$$\mathbb{E}(e^{X_1}) = \int_{-\infty}^{\infty} e^x f(x) dx = \int_0^{\infty} e^x \cdot 2e^{-2x} dx = 2 \int_0^{\infty} e^{-x} dx = 2[-e^{-x}]_{x=0}^{\infty} = 2.$$

A szórás akkor véges, ha $(e^{X_1})^2$ várható értéke véges. Azonban ha ezt szeretnénk kiszámítani:

$$\mathbb{E}((e^{X_1})^2) = \mathbb{E}(e^{2X_1}) = \int_{-\infty}^{\infty} e^{2x} f(x) dx = \int_0^{\infty} e^{2x} \cdot 2e^{-2x} dx = 2 \int_0^{\infty} 1 dx,$$

ami nem értelmes, nem véges az integrál. Ezért az e^{X_1} szórása nem értelmezhető, nem véges:

$$D^2(e^{X_1}) = \mathbb{E}((e^{X_1})^2) - \mathbb{E}(e^{X_1})^2.$$

Itt az első tag nem véges, a második $2^2 = 4$.

2. Eloszlásbeli konvergencia és centrális határeloszlás-tétel

Emlékeztetőül:

2.1. Tétel (A nagy számok gyenge törvénye). Legyenek X_1, X_2, \dots olyan valószínűségi változók, melyek függetlenek és azonos eloszlásúak. Tegyük fel, hogy $D(X_1) < \infty$. Ekkor minden $\varepsilon > 0$ esetén

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X_1)| > \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty),$$

azaz $\bar{X}_n \rightarrow \mathbb{E}(X_1)$ sztochasztikusan.

2.2. Tétel (A nagy számok erős törvénye). Legyenek X_1, X_2, \dots valószínűségi változók, melyek függetlenek és azonos eloszlásúak. Tegyük fel még, hogy $m = \mathbb{E}(X_1) < \infty$. Ekkor

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mathbb{E}(X_1) = m$$

teljesül 1 valószínűséggel $n \rightarrow \infty$ esetén.

A második esetben gyengébb feltevésből erősebb állítás következik.

A fenti ábrákból azt a következtetést vonhatjuk le, hogy ha a tagok szórása véges, és függetlenek, azonos eloszlásúak, akkor az átlaguk normális eloszláshoz hasonlóan viselkedik. Ennek az az „oka”, amit láttunk is, hogy a normális eloszlás átlagolás után normális eloszlású marad, az átlagot részekre bontva pedig azt láthatjuk, hogy a határértékben megjelenő eloszlásnak pontosan ilyennek kell lennie. Az átlag és az összeg között csak egy konstans szorzó a különbség, így ha az átlag nagyjából normális eloszlású, akkor ez az összegre is igaz. Ezt fogalmazza meg pontosan az alábbi állítás.

Emlékeztetőül: X_i és X_j azonos eloszlásúak, ha $\mathbb{P}(X_i \leq t) = \mathbb{P}(X_j \leq t)$ minden i, j párra és t valós számra. Ilyenkor a várható értékük és a szórásuk is megegyezik. Ha van sűrűségfüggvényük, és azonos eloszlásúak, akkor a sűrűségfüggvények is megegyeznek.

2.3. Tétel (Centrális határeloszlástétel). Legyenek X_1, X_2, \dots független azonos eloszlású valószínűségi változók, melyekre $\mathbb{E}(X_1) = m$ és $D(X_1) = \sigma < \infty$, azaz **szórásuk véges**. Ekkor tetszőleges t valós számra

$$\mathbb{P}\left(\frac{X_1 + X_2 + \dots + X_n - n \cdot m}{\sigma\sqrt{n}} \leq t\right) \rightarrow \mathbb{P}(Z \leq t) \quad (n \rightarrow \infty),$$

ahol Z standard normális eloszlású, azaz

$$\mathbb{P}(Z \leq t) = \Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

Az alábbi definíció segít ennek az általános megfogalmazásában.

2.1. Definíció. A Z_1, Z_2, \dots valószínűségi változókból álló sorozat **eloszlásban konvergál** a Z valószínűségi változóhoz, ha minden olyan t számra, melyre Z eloszlásfüggvénye folytonos t -ben, teljesül, hogy

$$\mathbb{P}(Z_n \leq t) \rightarrow \mathbb{P}(Z \leq t) \quad (n \rightarrow \infty).$$

Ez gyengébb, mint akár a sztochasztikus, akár az 1 valószínűségű konvergencia: ezekből következik az eloszlásbeli konvergencia, de fordítva nem állíthatjuk ezt.

Másképpen:

1 valószínűségű \Rightarrow sztochasztikus \Rightarrow eloszlásbeli

Fordítva egyik irány sem igaz.

Megjegyzés. Példa arra, hogy gyengén konvergál, de sem sztochasztikusan, sem 1 valószínűséggel: Z_1, Z_2, \dots mindegyik egymástól független standard normális eloszlású. Vagy szabályos kockadobások sorozata. Ekkor $\mathbb{P}(Z_n \leq t)$ minden n -re ugyanaz, a sorozat konvergál egy ugyanolyan kockadobáshoz eloszlásban. De például sztochasztikusan nem konvergál:

$$\mathbb{P}(|Z_n - Z| > 0,5)$$

nem tart nullához, hiszen bármilyen Z , a lehetséges értékek közül legalább három messzebb van tőle, mint 0,5, a valószínűség legalább 1/2. Az 1 valószínűségű konvergencia sem igaz, ilyesmi sorozat: 2, 5, 2, 3, 4, ... nulla valószínűséggel konvergens, nulla valószínűséggel áll egy idő után csupa azonos számból.

Megjegyzés. Példa sztochasztikusan konvergens, de nem 1 valószínűséggel konvergens sorozatra.

Legyen X egy, a $[0, 1]$ intervallumból függetlenül, egyenletes eloszlás szerint sorsolt szám. Legyen Z_1 annak indikátora, hogy $X \in [0, 1/2]$, azaz 1, ha $X \leq 1/2$, és 0 különben. Legyen Z_2 annak indikátora, hogy $X \in [1/2, 1]$. Utána Z_3 annak indikátora, hogy $X \in [0, 1/4]$, majd Z_4 annak, hogy $X \in [1/4, 1/2]$, Z_5 annak, $X \in [1/2, 3/4]$, és Z_6 annak, hogy $X \in [3/4, 1]$. A következő 8 darab Z -t ugyanígy készítjük, csak akkor 8 részre osztjuk az intervallumot, $1/8, 1/4, 3/8, \dots, 7/8$ osztópontokkal. Ezt folytatjuk a 2 hatványaival.

A kapott sorozat sztochasztikusan konvergál az azonosan 0-hoz. Ugyanis ha Z_n egy z_n hosszú intervallum indikátora (ezt pontosan meg is lehetne határozni), akkor tetszőleges $\varepsilon > 0$ -ra

$$\mathbb{P}(|Z_n| > \varepsilon) \leq \mathbb{P}(Z_n = 1) = z_n \rightarrow 0,$$

hiszen a sorozatot úgy készítettük, hogy az intervallumok hosszai nullához tartanak.

Az egy valószínűségű konvergencia azonban nem teljesül. A Z -k értékei 0-k vagy 1-ek, egy ilyen sorozat akkor konvergál csak, ha egy idő után végig vagy 0, vagy 1. Ez most nem igaz, mert akármilyen X értéke, az minden 2-hatványhoz tartozó blokkból egy (esetleg két) intervallumba beleesik, a többibe viszont nem, így minden blokkban előfordul 0 és 1 is. Ebből adódik, hogy az 1 valószínűségű konvergencia nem teljesül.

Térjünk vissza a centrális határeloszlástételhez. Az állítást a definíció alapján úgy is fogalmazhatjuk, hogy

$$\frac{X_1 + X_2 + \dots + X_n - n \cdot m}{\sigma\sqrt{n}} \rightarrow N(0, 1)$$

teljesül $n \rightarrow \infty$ esetén eloszlásban. Másképpen

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(a \leq \frac{X_1 + X_2 + \dots + X_n - n \cdot m}{\sigma\sqrt{n}} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx = \Phi(b) - \Phi(a) = \mathbb{P}(a \leq Y \leq b),$$

ahol $Y \sim N(0, 1)$, és Φ a standard normális eloszlás eloszlásfüggvénye.

Így is átfogalmazható a tétel állítása:

$$\mathbb{P}(nm + a\sigma\sqrt{n} \leq X_1 + X_2 + \dots + X_n < nm + b\sigma\sqrt{n}) \rightarrow \Phi(b) - \Phi(a).$$

Ha n -nel osztunk, hogy az átlag jelenjen meg:

$$\mathbb{P}\left(\mathbf{m} + \mathbf{a} \frac{\sigma}{\sqrt{n}} \leq \frac{X_1 + X_2 + \dots + X_n}{n} < \mathbf{m} + \mathbf{b} \frac{\sigma}{\sqrt{n}}\right) \rightarrow \mathbb{P}(\mathbf{a} \leq Z \leq \mathbf{b}).$$

Vagyis az **átlag eloszlása** „közel van” egy \mathbf{m} várható értékű, $\frac{\sigma}{\sqrt{n}}$ szórású **normális eloszláshoz**. Ezt figyelhettük meg az ábrákon is, a véges szórású esetekben.

Példa. Legyenek X_1, X_2, \dots független, 2 várható értékű, exponenciális eloszlású valószínűségi változók. Mi a limesze a $\mathbb{P}(X_1 + \dots + X_n - 2n < 2\sqrt{n})$ mennyiségnek $n \rightarrow \infty$ esetén?

Mivel a valószínűségi változók **függetlenek**, **azonos eloszlásúak** és **véges szórásúak**, teljesülnek a centrális határeloszlástétel feltételei. Ezért

$$\mathbb{P}(X_1 + \dots + X_n - 2n < 2\sqrt{n}) = \mathbb{P}\left(\frac{X_1 + \dots + X_n - 2n}{2\sqrt{n}} < 1\right) \rightarrow \Phi(1),$$

ha $n \rightarrow \infty$, hiszen $\mathbf{m} = 2$ a várható érték, és mivel az eloszlás exponenciális, a várható érték egyenlő a szórással, így $\sigma = 2$ a szórás.

3. A centrális határeloszlástétel alkalmazása

A centrális határeloszlás-tétel egyik következménye, hogy a normális eloszlásra vonatkozó hipotézisvizsgálati eljárások (z , t , F -próba) akkor is alkalmazhatók, ha a minta eloszlása nem normális eloszlású, hanem más, véges szórású eloszlás, a mintaelemszám pedig kellőképpen nagy (ugyanakkor maguk az eljárások túl nagy mintaelemszám esetén nem feltétlenül adnak helyes végeredményt). Ugyanis ezek az eljárások az

átlagot és a korrigált tapasztalati szórást használják, és ez utóbbi is átlagként írható fel, úgy viselkedik, mintha normális eloszlású lenne a minta.

Példa. Tegyük fel, hogy egy véletlenszerűen választott ember p valószínűséggel szavaz egy adott pártra. Legalább hány embert kell megkérdeznünk (feltételezve, hogy mindenki a többiektől függetlenül válaszol és igazat mond), hogy annak valószínűsége, hogy a pártra szavazók aránya legfeljebb 0,01-gyel tér el p -től, tetszőleges p esetén legalább 95% legyen?

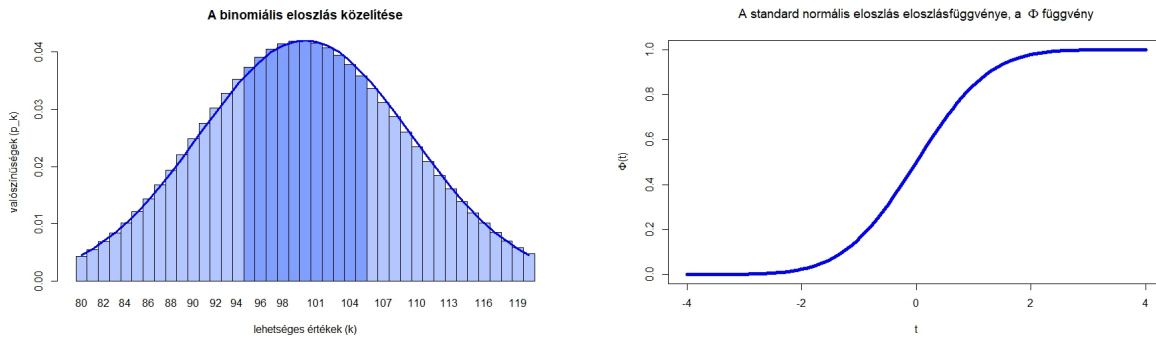
n megkérdezett, mindenki p valószínűséggel támogatja a pártot

X : a pártot támogatók száma a megkérdezettek között, ez binomiális eloszlású.

Kell:

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \leq 0,01\right) \geq 0,95$$

teljesüljön minden $0 \leq p \leq 1$ -re.



6. ábra. A binomiális eloszlás közelítése a normális eloszlással, és a Φ függvény

Itt tehát $X = \sum_{j=1}^n X_j$, az X_j -k függetlenek, és

$$\mathbb{P}(X_j = 1) = 1 - \mathbb{P}(X_j = 0) = p; \quad \mathbb{E}(X_j) = p; \quad D(X_j) = \sqrt{p(1-p)}.$$

A kérdéses valószínűséget így közelíthetjük:

$$\begin{aligned} \mathbb{P}\left(\left|\frac{X}{n} - p\right| \leq 0,01\right) &= \mathbb{P}\left(\left|\frac{X - np}{\sqrt{n}}\right| \leq 0,01\sqrt{n}\right) = \\ &= \mathbb{P}\left(\left|\frac{\sum_{j=1}^n X_j - np}{\sqrt{np(1-p)}}\right| \leq \frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) \approx 2\Phi\left(\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1, \end{aligned}$$

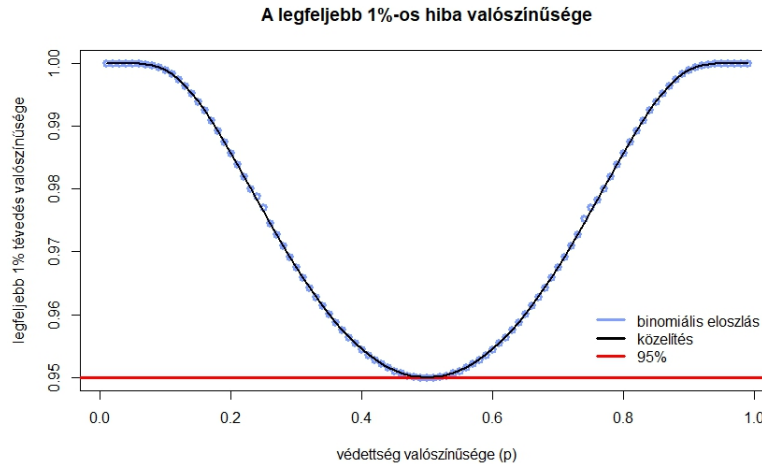
ugyanis a centrális határeloszlás-tétel alapján

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(-t \leq \frac{\sum_{j=1}^n X_j - np}{\sqrt{np(1-p)}} \leq t\right) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1,$$

és ugyan most t helyén egy n -től függő érték szerepel, még erősebb tételek (Berry–Esséen-tétel) segítségével az így elkövetett hiba is megbecsülhető lenne, ha $np(1-p) \geq 10$, akkor ez nem egy rossz közelítés. A $\Phi(-t) = 1 - \Phi(t)$ azonosság a standard normális eloszlás szimmetriájából következik.

Ez tehát, a közelítést elfogadva, elégséges feltétel, felhasználva, hogy Φ szigorúan monoton növekvő (hiszen a standard normális eloszlás eloszlásfüggvénye), ezért az inverzét alkalmazva megmarad az egyenlőtlenség (a szigorú monoton növekedés miatt injektív is, azaz különböző értékekhez különbözőt rendel):

$$\begin{aligned} 2\Phi\left(\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 &\geq 0,95; \\ \Phi\left(\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) &\geq 0,975; \\ \frac{0,01\sqrt{n}}{\sqrt{p(1-p)}} &\geq \Phi^{-1}(0,975) = \text{qnorm}(0,975) = 1,96; \\ n &\geq p(1-p) \cdot 1,96^2 \cdot \frac{1}{0,01^2}. \end{aligned}$$



7. ábra. $n = 9607$ megkérdezett esetén annak valószínűsége, hogy 0,01-nél kevesebbet tévedünk, p függvényében

A feltételt minden n -re teljesítenünk kell. Mivel $p(1 - p) \leq 1/4$ (azaz a feltétel $p = 1/2$ esetén a legerősebb), az elég, hogy:

$$n \geq \frac{1}{4} \cdot 1,96^2 \cdot \frac{1}{0,01^2} = 9607.$$

A példát korábban a Csebisev-egyenlőtlenséggel is megoldottuk, ezzel és a pontos értékkel is összehasonlíthatjuk az eredményt:

- Csebisev-egyenlőtlenséggel: $n \geq 50000$ biztosan elég
- centrális határeloszlástétellel közelítve: $n \geq 9607$ elég
- valójában: $n = 9607, p = 1/2$ esetén 0,94987 adódik a 0,95 helyett
- valójában $n \geq 9650$ kell (pontos számolással)

A kapott képletből azt is láthatjuk, hogy

- ha 0,01 helyett ε a megengedett tévedés, akkor a szükséges mintaelemszám $1/\varepsilon^2$ -tel arányos, azaz kétszer akkora pontossághoz négyszer akkora minta kell;
- ha 0,95 helyett $1 - \alpha$ valószínűséggel kell ε -nál kevesebbet tévednünk, akkor $\Phi^{-1}(0,975)^2$ szerepel a becslésben, erre kevésbé érzékeny a szükséges mintaelemszám;
- ha p -ről van valamilyen előzetes információnk, akkor $p(1-p)$ -re jobb felső becslés, ennek megfelelően kisebb n is adható;
- ugyanakkor a fenti következtetések csak akkor érvényesek, ha np nem túlságosan kicsi, itt az $np(1-p) \geq 10$ feltételt szokták megfogalmazni, enélkül a normális eloszlással való közelítés elromolhat. Ebből az is következik, hogy a kis gyakoriságokat nehéz pontosan megbecsülni, főleg, ha a hibát relatív, és nem additív értelemben értjük.

Fontos, hogy a fenti számolás csak "ideális körülmények" között érvényes: mindenki válaszolt, igazat mondott, a válaszok függetlenek voltak. Minden más esetben ez inkább egy alsó becslésnek tekinthető.

Házi feladat december 3., péntek, 12:15-ig Sorsoljunk 100000 elemű mintát valamilyen $[a, b]$ intervallumon egyenletes eloszlásból (itt $[a, b]$ tetszőlegesen választható. A mintát osszuk szét ezres csoportokra, mindegyiknek számítsuk ki az átlagát, és tekintsük ezt a száz megfigyelést.

a) Készítsünk hisztogramot a kapott mintából.

b) A hisztogramon válasszuk ki egy oszlopot, a határai legyenek c és d . A centrális határeloszlástétel alapján hogyan közelíthető annak valószínűsége, hogy az 1000 egyenletes eloszlás átlaga c és d közé esik? Ez mennyire egyezik meg a hisztogramról leolvasható értékkel?