Valószínűségszámítás, 7. feladatsor, 2022. november 7-11.

(1) Legyenek az $X_1, X_2, \ldots$ valószínűségi változók függetlenek. Milyen értelemben konvergensek az alábbi sorozatok, és mi a limeszük?

(a) X_i független p paraméterű indikátorváltozó; $Y_n = (X_1^5 + \ldots + X_n^5)/n$.

(b) X_i az i. szabályos kockadobás eredménye; $Y_n = (X_1 + \ldots + X_n)/n$; illetve $Z_n = (X_1^2 + \ldots + X_n^2)/n$.

(c) X_i exponenciális eloszlású 2 paraméterrel (azaz sűrűségfüggvénye $f(x) = 2e^{-2x}\mathbb{I}(x > 0)$; $Y_n = (e^{X_1} + \ldots + e^{X_n})/n$, illetve $Z_n = \frac{X_1^2 + X_2^2 + \ldots + X_n^2}{n}$.

Megoldás

(a) Először vegyük észre, hogy mivel X_j lehetséges értékei csak 0 vagy 1, ezért $X_j^5 = X_j$ teljesül. Másrészt, mivel X_j korlátos, véges a negyedik momentuma. A függetlenség is teljesül. Így a nagy számok Cantelli-féle törvénye alapján a limesz

$$\mathbb{E}(X_1) = 1\mathbb{P}(X_1 = 1) = p$$

lesz 1 valószínűségű értelemben, amiből következik, hogy sztochasztikus értelemben is p a limesz.

(b) Az (a) részhez hasonlóan most is független, azonos eloszlású, korlátos valószínűségi változókról van szó, így a limesz az első esetben

$$\mathbb{E}(X_1) = \frac{1}{6}(1 + 2 + \ldots + 6) = 3,5$$

egy valószínűséggel, míg a második esetben az X_j^2 valószínűségi változókra alkalmazzuk a nagy számok Cantelli-féle törvényét (ezek is függetlenek, korlátosak), és a limesz

$$\mathbb{E}(X_1^2) = \frac{1}{6}(1^1 + 2^2 + \ldots + 6^2) = \frac{91}{6},$$

szintén 1 valószínűséggel, ezért sztochasztikusan is.

(c) Számítsuk ki e^{X_1} várható értékét:

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

Ez véges, ezért az e^{X_j} független, azonos eloszlású, véges várható értékű valószínűségi változókra alkalmazhatjuk a Kolmogorov-féle nagy számok erős törvényét, amiből kapjuk, hogy a limesz 2 lesz 1 valószínűséggel, ezért sztochasztikus értelemben is.

Vegyük észre ugyanakkor, hogy e^{2X_1} várható értéke nem létezik (a hasonló integrál improprius értelemben végtelen), így nem véges a szórás, sem a Bernoulli-féle, sem a Cantelli-féle nagy számok törvénye nem alkalmazható.

Mivel az exponenciális eloszlás szórása és várható értéke is $\frac{1}{\lambda}$, most

$$\mathbb{E}(X_1^2) = D^2(X_1) + \mathbb{E}(X_1)^2 = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{2}{\lambda^2} = \frac{1}{2}.$$

A Kolmogorov-féle nagy számok törvényét alkalmazva kapjuk, hogy a limesz 1 lesz, 1 valószínűségű értelemben, és így sztochasztikusan is.

(2) Az X valószínűségi változó sűrűségfüggvénye legyen x^{-5}, ha $x > c$, és 0 különben.

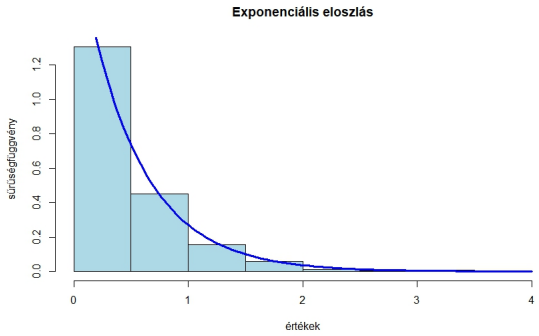
(a) Határozzuk meg c értékét.

(b) Feltéve, hogy $X > 2c$, mennyi a valószínűsége, hogy $X > 3c$?

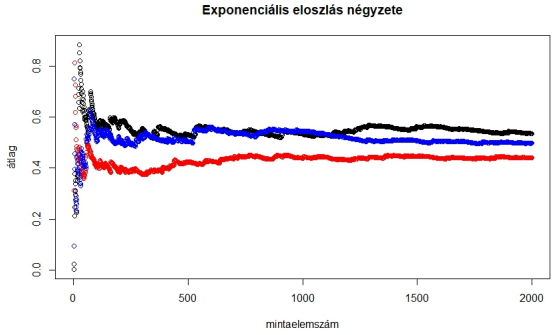
(c) Legyenek $X_1, X_2, \ldots$ az X-szel azonos eloszlású, egymástól független valószínűségi változók. Határozzuk meg az

$$\frac{X_1^3 + X_2^3 + \ldots + X_n^3}{n}$$

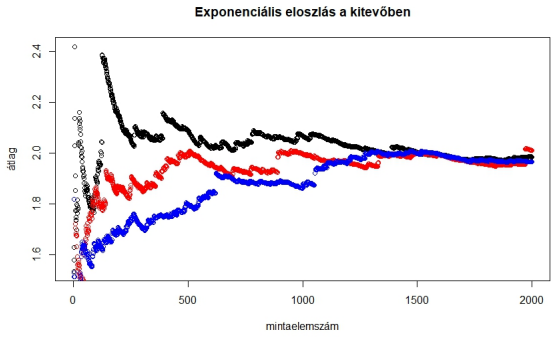
limeszét sztochasztikus, illetve 1 valószínűségű értelemben, ha ezek a limeszek léteznek $n \to \infty$ esetén.



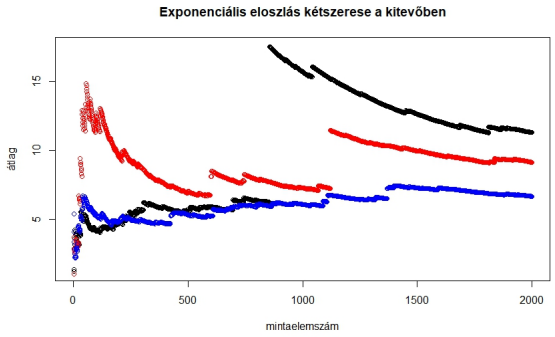
1. ábra. 2000 elemű exponenciális eloszlású minta $\lambda = 2$ paraméterrel és a sűrűségfüggvény



2. ábra. Az $(X_1^2 + \ldots + X_n^2)/n$ sorozat $n = 2000$-ig, $\lambda = 2$ paraméterű exponenciális X_j-kkel



3. ábra. Az $(e^{X_1} + \ldots + e^{X_n})/n$ sorozat $n = 2000$-ig, $\lambda = 2$ paraméterű exponenciális X_j-kkel



4. ábra. Az $(e^{2X_1} + \ldots e^{2X_n})/n$ sorozat $n = 2000$-ig, $\lambda = 2$ paraméterű exponenciális X_j-kkel

Megoldás

(a) Mivel a sűrűségfüggvény integrálja 1, annak kell teljesülnie, hogy $\int_c^\infty x^{-5} d\,x = 1$, azaz $\frac{c^{-4}}{4} = 1$, és így $c = \sqrt[4]{1/4}$.

(b) Először számítsuk ki a feltétel valószínűségét:

$$\mathbb{P}(X > 2c) = \int_{2c}^{\infty} x^{-5}\, dx = \frac{(2c)^{-4}}{4} = \frac{1}{16}.$$

Hasonlóképpen

$$\mathbb{P}(X > 3c) = \int_{3c}^{\infty} x^{-5}\, dx = \frac{(3c)^{-4}}{4} = \frac{1}{81}.$$

Ezután a feltételes valószínűség definíciója alapján

$$\mathbb{P}(X > 3c | X > 2c) = \frac{\mathbb{P}(\{X > 3c\} \cap \{X > 2c\})}{\mathbb{P}(X > 2c)} = \frac{\frac{1}{81}}{\frac{1}{16}} = \frac{16}{81} = 19,8\%.$$

Vegyük észre, hogy ez több, mint a $\mathbb{P}(X > 3c)$ valószínűség, ahogy ez általában is a Pareto-eloszlásokra jellemző (ez az eloszlás is a Pareto-eloszlások közé tartozik).

(c) Számítsuk ki X_1^3 várható értékét:

$$\int_{-\infty}^{\infty} x^3 f(x)\, dx = \int_c^{\infty} x^{-2}\, dx = \frac{1}{c} = \sqrt[4]{4}.$$

Mivel ez véges, és a tagok függetlenek, azonos eloszlásúak, alkalmazhatjuk a nagy számok Kolmogorov-féle erős törvényét, ebből kapjuk, hogy a határérték 1 valószínűséggel (és így sztochasztikusan is) $\sqrt[4]{4}$ lesz.

(3) Tegyük fel, hogy egy biztosító ügyfelei minden napon a többitől függetlenül 50 várható értékű Poisson-eloszlással leírható számú balesetet szenvednek. Legyen X_j a j. napon okozott károk száma.

Határozzuk meg az

$$\lim_{n\to\infty} \frac{X_1 + X_2 + \ldots + X_n}{n} \qquad \text{és} \qquad \lim_{n\to\infty} \frac{X_1^2 + X_2^2 + \ldots + X_n^2}{n}$$

határértékeket, azzal együtt, hogy milyen értelemben léteznek ezek a határértékek.

Megoldás

Az első kérdés megválaszolásához elég, hogy független, 50 várható értékű, azonos eloszlású valószínűségi változókról van szó, ebből már a Kolmogorov-féle nagy számok erős törvénye szerint következik, hogy a sorozat 1 valószínűséggel konvergál a várható értékhez, vagyis 50-hez. Az 1 valószínűségű konvergenciából a sztochasztikus is következik.

Másrészt

$$\mathbb{E}(X_1^2) = D^2(X_1) + \mathbb{E}(X_1)^2 = 50 + 50^2 = 2550,$$

hiszen Poisson-eloszlás esetén a várható érték és a szórásnégyzet is a paraméterrel egyezik meg. Mivel az X_j-k függetlenek, azonos eloszlásúak, az (X_j^2) sorozat is független, azonos eloszlású, véges várható értékű tagokból áll, így a Kolmogorov-féle nagy számok törvényéből következik, hogy az átlag a várható értékhez, vagyis 2550-hez konvergál 1 valószínűséggel, és így sztochasztikusan is.

(4) Az USA-ban a férfiak átlagos magassága 176 cm, 7 cm szórással. Mekkora az esélye, hogy valaki 2 méternél magasabb? Adjunk meg egy olyan D számot, melyre igaz, hogy a férfiak 95%-ának magassága $176 - D$ és $176 + D$ közé esik!

Feltehetjük, hogy a testmagasság normális eloszlású. Mivel normális eloszlású valószínűségi változó lineáris transzformáltja is normális eloszlású, ha X a testmagasság, akkor $(X - 176)/7$ standard normális eloszlású. Tehát, ha Φ jelöli a standard normális eloszlás eloszlásfüggvényét, akkor

$$\mathbb{P}(X > 200) = 1 - \mathbb{P}(X < 200) = 1 - \mathbb{P}\left(\frac{X - 176}{7} < \frac{200 - 176}{7}\right) = 1 - \Phi\left(\frac{23}{7}\right) = 0,0005,$$

a Φ függvény táblázata alapján.

Ezután az a kérdés, hogy milyen D számra igaz, hogy

$$\mathbb{P}(176 - D < X < 176 + D) = 0,95.$$

Az előzőhöz hasonlóan számolva

$$\mathbb{P}(176 - D < X < 176 + D) = \Phi\left(\frac{D}{7}\right) - \Phi\left(-\frac{D}{7}\right) = 2\Phi\left(\frac{D}{7}\right) - 1,$$

ahol felhasználtuk, hogy $\Phi(x) = 1 - \Phi(x)$ teljesül minden x valósz számra a standard normális eloszlás 0 körüli szimmetriája miatt. Tehát D-nek ezt kell teljesítenie:

$$\Phi\left(\frac{D}{7}\right) = 0,975 \qquad \Rightarrow \qquad D = 7 \cdot 1,96 = 13,72.$$

(5) Egy átlagos magyar háztartásban $100m^3$ víz fogy évente, $20m^3$ szórással. Ha 3×10^6 háztartás van Magyarországon, akkor mekkora az esélye, hogy a $3,3 \times 10^8 m^3$ rendelkezésre álló vízkészlet elég lesz?

Ha feltesszük a függetlenséget, akkor még a Csebisev-egyenlőtlenség alapján is igen kicsi valósznűség adódik arra, hogy ne legyen elég a vízkészlet:

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - nm\right| \geq 0,3 \times 10^8\right) \leq \frac{D^2(\sum X_i)}{0,09 \times 10^{16}} = \frac{6 \times 10^9}{9 \times 10^{14}} = 6,67 \times 10^{-6}.$$

Normális közelítéssel számolva: tegyük fel, hogy a j. háztartás fogyasztása X_j, és ezek a valószínűségi változók egymástól függetlenek, azonos eloszlásúak. Azt tudjuk, hogy a szórásuk véges. Alkalmazzunk közelítést a centrális határeloszlástétel alapján:

$$\mathbb{P}\left(\sum_{j=1}^{n} X_j < 3,3 \cdot 10^8\right) = \mathbb{P}\left(\frac{\sum_{j=1}^{n} X_j - 3 \cdot 10^8}{20 \cdot \sqrt{3 \cdot 10^6}} < \frac{3 \cdot 10^7}{20 \cdot \sqrt{3 \cdot 10^6}}\right) \approx \Phi\left(\frac{3 \cdot 10^7}{20 \cdot \sqrt{3 \cdot 10^6}}\right) = \Phi(273,8),$$

ami nagyon közel van 1-hez.

A valóságban azonban korántsem várható, hogy függetlenek legyenek egymástól a vízfogyasztások (például szárazság idején minden kertes házban többet locsolnak). Ha teljes összefüggés lenne ($X_i = X_1$ minden i-re), akkor a teljes fogyasztás szórása $60 \times 10^6 m^3$, ami összemérhető a tartalék mennyiségével, azaz még normális eloszlás feltételezésével is bőven lehet vízhiány. Ez jól mutatja, hogy mennyire fontos a függetlenség - illetve a kovariancia, aminek segítségével számolható az általános esetben az összeg szórása.

(6) Egy gyárban 2000 lámpa világít. Évente mindegyikben a többitől függetlenül $p = 1/4$ valószínűséggel ég ki az izzó. Mekkora az esélye, hogy az idénre megvásárolt 540 tartalék izzó elegendő lesz a pótlásra?

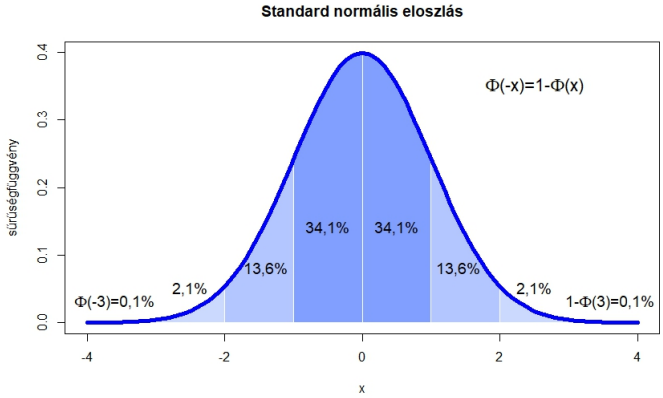
Legyen X_j annak indikátora, hogy a j. lámpa kiég. Ezek független, azonos eloszlású, véges szórású valószínűségi változók, nevezetesen, a várható értékük $1/4$, a szórásnégyzetük $3/16$. Alkalmazhatjuk tehát a centrális határeloszlástételen alapuló közelítést:

$$\mathbb{P}\left(\sum_{j=1}^{2000} X_j \leq 540\right) = \mathbb{P}\left(\frac{\sum_{j=1}^{2000} X_j - 500}{\sqrt{2000 \cdot 3/16}} < \frac{40}{\sqrt{2000 \cdot 3/16}}\right) \approx \Phi\left(\frac{40}{\sqrt{375}}\right) = \Phi(2,07) = 98,1\%.$$

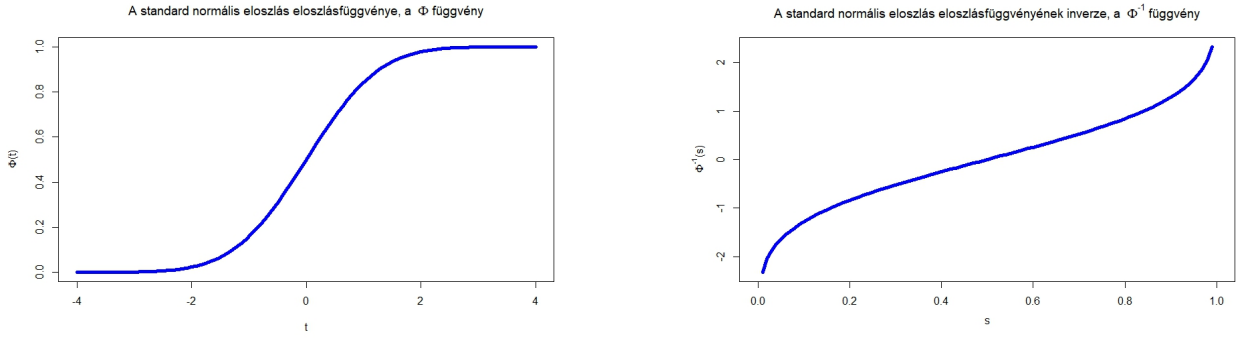
(7) Szeretnénk megállapítani, hogy hány dohányos él Budapesten. Ezért megkérdezünk n véletlenszerűen kiválasztott budapesti lakost arról, hogy dohányoznak-e (visszatevéses mintavétellel). A de Moivre–Laplace (centrális határeloszlás) tétel alapján milyen nagyra kell n-et választani, ha azt szeretnénk, hogy a kapott relatív gyakoriság legfeljebb 1 százalékot tévedjen legalább 95%-os megbízhatósággal? Mit kapunk a Csebisev-egyenlőtlenségből?

Tegyük fel, hogy minden megkérdezett egymástól függetlenül p valószínűséggel dohányzik ($p \in [0,1]$ ismeretlen). Itt feltettük, hogy mindenki válaszol, és igazat mond (vagyis nem megbízható válaszok esetén az eredmény sem megbízható), és mindenkit azonos valószínűséggel találunk meg. A függetlenséget viszonylag könnyű biztosítani, ha függetlenül választunk, és a válaszadók nem látják egymás válaszát.

A dohányosok száma binomiális eloszlású n renddel és p paraméterrel. Legyen X_i ($i = 1, \ldots, n$) annak indikátora, hogy az i. ember dohányzik, azaz X_i értéke 1, ha az i. megkérdezett dohányzik, 0 különben. A p-re adott becslés a dohányzók száma osztva az összes megkérdezett számával, vagyis $\sum_{i=1}^{n} X_i/n$. A relatív gyakoriság: Y_n/n, azaz



5. ábra. A φ függvény



6. ábra. A standard normális eloszlásfüggvény, azaz $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$ és Φ^{-1}

Y_n ember mondta, hogy dohányzik. Itt Y_n binomiális eloszlás, az Y_n/n várható értéke p. Vagyis a feltétel, amit teljesíteni kell: minden $p \in [0,1]$-re

$$\mathbb{P}\left(\left|\frac{Y_n}{n} - p\right| < 0,01\right) \geq 0,95 \qquad \Leftrightarrow \qquad \mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - p\right| \geq 0,01\right) \leq 0,05. \tag{1}$$

A centrális határeloszlástételből a következőt tudjuk (hiszen $\sum_{j=1}^n X_j$ eloszlása binomiális eloszlás n renddel és p paraméterrel):

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{\sum_{j=1}^n X_j - np}{\sqrt{np(1-p)}} \leq b\right) = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\,dx.$$

$$\mathbb{P}\left(\left|\frac{Y_n}{n} - p\right| < 0,01\right) = \mathbb{P}\left(-0,01 \leq \frac{Y_n - np}{n} \leq 0,01\right) =$$
$$= \mathbb{P}\left(-\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}} \leq \frac{Y_n - np}{\sqrt{np(1-p)}} \leq \frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right).$$

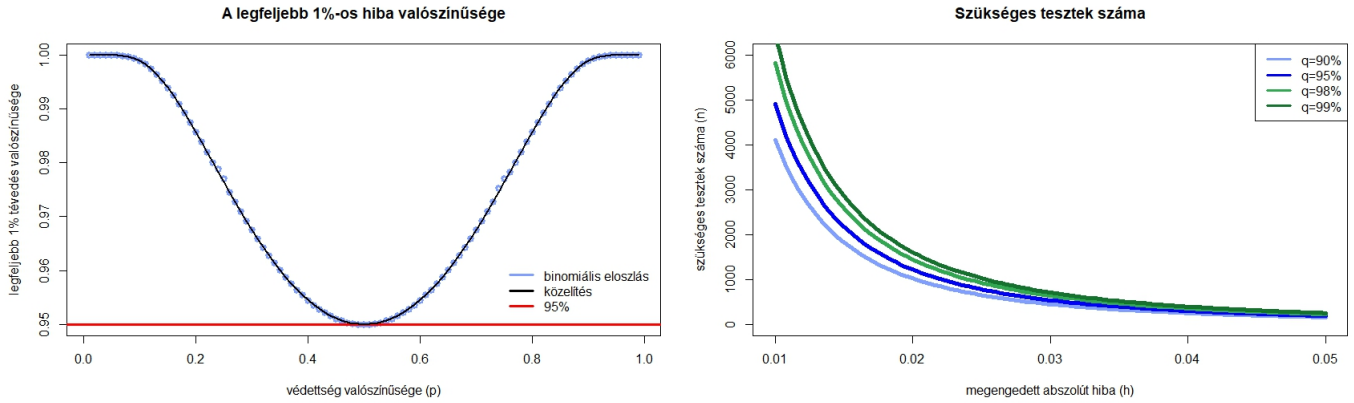
Itt megjelent Y_n standardizáltja, használhatjuk a centrális határeloszlástételt (bár a tételben a és b rögzített volt, itt pedig a határok is végtelenhez tartanak; a Berry–Esséen-tétel akkor lenne jól használható, ha például $0,1 < p < 0,9$-et tudnánk)

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - p\right| < 0,01\right) \approx \Phi\left(\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) = 2\Phi\left(\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1,$$

a $\Phi(-a) = 1 - \Phi(a)$ azonosság alapján. Tehát az kell, hogy

$$2\Phi\left(\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 \geq 0,95$$

$$\Phi\left(\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}}\right) \geq 0,975$$



7. ábra. A szükséges kérdések száma normális közelítéssel néhány megbízhatósági szint mellett a megengedett abszolút hiba függvényében (balra), illetve annak valószínűsége, hogy legfeljebb 1%-ot téved a becslés, a valós p védettségi arány függvényében, $n = 9604$ teszt esetén

$$\frac{0,01\sqrt{n}}{\sqrt{p(1-p)}} \geq \Phi^{-1}(0,975) = 1,96$$

Itt használtuk, hogy a Φ monoton növő. Ebből

$$n \geq 196^2 \cdot p(1-p) = 38416 \cdot p(1-p)$$

Ehhez pedig $p(1-p) \leq 1/4$ alapján elég, hogy

$$n \geq \left(\frac{1,96 \cdot 0,5}{0,01}\right)^2 = 9604.$$

Vagyis $n = 9604$ embert elég megkérdezni. Számítógéppel ellenőrizve ez lényegében a pontos érték (7. ábra).

A másik módszerrel:

Csebisev-egyenlőtlenség véges szórású Y-ra és $t > 0$-ra:

$$\mathbb{P}(|Y - \mathbb{E}(Y)| \geq t) \leq \frac{D^2(Y)}{t^2}.$$

Itt $Y = \sum X_i$ binomiális eloszlású, várható értéke pn. Tehát a Csebisev-egyenlőtlenség alapján

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - p\right| \geq 0,01\right) = \mathbb{P}\left(\left|\sum_{i=1}^n X_i - pn\right| \geq 0,01n\right) \leq \frac{D^2(\sum X_i)}{0,01^2 n^2} = \frac{np(1-p)}{0,01^2 n^2} = \frac{p(1-p)}{0,01^2 n}.$$

Ezért elég, hogy minden p-re

$$\frac{p(1-p)}{n \cdot 0,01^2} \leq 0,05.$$

Mivel $p(1-p) \leq 1/4$ minden p-re, ehhez elég, hogy

$$n \geq \frac{1}{4 \cdot 0,01^2 \cdot 0,05} = 50000.$$

Ennek a példának egy részletesebb kifejtése:

`https://ematlap.hu/tudomany-tortenet-2020-12/992-mennyit-teszteljunk-2-v3`

Az eredeti kérdés 2%-os tévedési valószínűséggel, Csebisev-egyenlőtlenséggel:

Tegyük fel, hogy minden megkérdezett egymástól függetlenül p valószínűséggel dohányzik ($p \in [0,1]$ ismeretlen). A megkérdezettek száma legyen n (ez ismert, sőt szabadon megválasztható).

Legyen X_i ($i = 1, \ldots, n$) annak indikátora, hogy az i. ember dohányzik, azaz X_i értéke 1, ha az i. megkérdezett dohányzik, 0 különben. A p-re adott becslés a dohányzók száma osztva az összes megkérdezett számával, vagyis $\sum_{i=1}^n X_i/n$.

Vagyis a feltétel, amit teljesíteni kell: minden $p \in [0,1]$-re

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - p\right| < 0,02\right) \geq 0,9 \qquad \Leftrightarrow \qquad \mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - p\right| \geq 0,02\right) \leq 0,1. \tag{2}$$

Kell:

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - pn\right| \geq 0,02n\right) \leq 0,1$$

Csebisev-egyenlőtlenség véges szórású Y-ra és $t > 0$-ra:

$$\mathbb{P}(|Y - \mathbb{E}(Y)| \geq t) \leq \frac{D^2(Y)}{t^2}.$$

Itt $Y = \sum X_i$ binomiális eloszlású, várható értéke pn. Tehát a Csebisev-egyenlőtlenség alapján

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - p\right| \geq 0,02\right) = \mathbb{P}\left(\left|\sum_{i=1}^n X_i - pn\right| \geq 0,02n\right) \leq \frac{D^2(\sum X_i)}{0,02^2 n^2} = \frac{np(1-p)}{0,02^2 n^2} = \frac{p(1-p)}{0,02^2 n}.$$

Ezért elég, hogy minden p-re

$$\frac{p(1-p)}{n \cdot 0,02^2} \leq 0,1.$$

Mivel $p(1-p) \leq 1/4$ minden p-re, ehhez elég, hogy

$$n \geq \frac{1}{4 \cdot 0,02^2 \cdot 0,1} = 6250.$$

Ez a $0,02$-ben, vagyis az eltérésben négyzetes (ez jobb módszereknél is így van). A valószínűségre kevésbé érzékeny. Viszont ha p kicsi, vagy 1-hez közeli, akkor a $0,02$ nagyon rossz becslés, jóval pontosabb közelítés is adható ennyi mintaelemből.

(8) Egy egyetemre 1000 diák jár. Mindegyikük $0,002$ valószínűséggel lesz beteg egy adott napon. Mekkora az esélye, hogy holnap legfeljebb 4-en lesznek betegek?

Ha feltesszük a függetlenséget, akkor a pontos eloszlás Binom(1000;0,002). Ebből kézzel egy meglehetősen kellemetlen számolás adja meg a keresett valószínűséget ($P(X \leq 4)$). Ezért lehet célszerű a Poisson eloszlás alkalmazása (tudjuk, hogy $np \to \lambda$ esetén a binomiális eloszlás tart a λ paraméterű Poissonhoz), ott barátságosabbak a képletek. De persze számítógéppel (pl. az R program segítségével) mindkét érték könnyen megkapható: $P(X \leq 4) \sim 0,95$, bármelyik modellt is használjuk.