Az 1993/3. szám 104–109. oldalán megjelent három probléma közül az első vizsgálata során jutottunk el az ún. Gauss-féle haranggörbéhez. Mivel ennek szerepe, mint a későbbiekben látni fogjuk, nemcsak a binomiális eloszlás közelítésekor jelentős, ezért szeretnénk megadni függvényként.

Próbáljuk meg az előzőekben „szemléletesen látott" tényt analizálni, ami természetesen a matematikai analízis segítségével történik. Tehát írjuk föl, mit szeretnénk becsülni:

$$P(X=k)=\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}=\frac{n!}{k!(n-k)!}\cdot p^k\cdot(1-p)^{n-k}.$$

Ennek becsléséhez először a faktoriálisok nagyságáról kellene tudni valamit. Először A. Moivre-nak sikerült erre egy közelítő formulát találnia, amely 1730-ban jelent meg *Miscellanea Analytica* című művében. Ebben a könyvben szerepel még a következőkben előforduló normális eloszlással kapcsolatos fontos képlet:

$$\int_0^\infty e^{-x^2}\mathrm{d}x=\frac{1}{2}\pi,$$

illetve a komplex számokat ismerők számára nevezetes $(\cos\alpha+\sin\alpha)^n=\cos n\alpha+\sin n\alpha$ Moivre-féle képlet is. A sors játéka, hogy ez a faktoriálisokra vonatkozó képlet, amelyet néhány évvel később James Stirling is megadott, nem Moivre, hanem Stirling-formula néven vált ismertté: $n!\approx(\frac{n}{e})^n\cdot\sqrt{2\pi n}$, ahol két „csúnya" szám is szerepel: az e és a π. Az utóbbit nyilván minden olvasó ismeri, az előbbit csak az analízisben járatosabbak, akik már sorozatok határértékvel foglalkoztak. Az e szám leggyakoribb definíciója:

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n=e.$$

A formula bizonyítását nem fogjuk ezen a helyen leírni, akit érdekel, elolvashatja (pl. [2]-ben 19-23. o.). Annyit mutatunk csak meg, hogy milyen induló ötlet kell hozzá. Nem $n!$-t hanem $\log n!$-t fogjuk megbecsülni, mert akkor a szorzatból összeg lesz, ami sokkal kezelhetőbb. A $\sum\log n$ kifejezés, ahol log a természetes logaritmus, melynek alapja éppen a fent definiált „e" szám, egy közelítő összege $x\to\log x$ függvénygörbe alatti területnek. Így a fenti kifejezés a $\log x$ függvény integráljával becsülhető (lásd a *3. ábrát*).

Ha tudjuk, hogy a $\log x$ egy primitív függvénye $x\log x-x$, akkor már majdnem megvan a keresett formula, hiszen az $\left(\frac{n}{e}\right)^n$ tényező ebből adódik. A hiba becslése adja a másik tényezőt. Megjegyezzük, hogy már $n>10$ esetén is kisebb a hiba, mint 1% (10-nél 0,83%, 67-nél 0,12%), így nagy n-ekre jól lehet használni a becslést.

A továbbiakban ezt a formulát fogjuk használni az $n!$ helyett:

$$\frac{n!}{k!(n-k)!}\cdot p^k\cdot(1-p)^{n-k}=\frac{n^n\cdot\sqrt{n}}{\sqrt{2\pi}\cdot\sqrt{k(n-k)}\cdot k^k\cdot(n-k)^{n-k}}\cdot p^k\cdot(1-p)^{n-k}.$$

Az egyszerűség kedvéért legyen $p=1/2$.

Ekkor

$$f(k)=P(X=k)=\frac{n^n\cdot\sqrt{n}}{\sqrt{2\pi}\cdot\sqrt{k(n-k)}\cdot k^k\cdot(n-k)^{n-k}}\cdot\frac{1}{2^n}.$$

Most ha elvégezzük az I. rész 2. ábráin is használt transzformációt, mivel a csúcs $n/2$-ben van, s a binomiális eloszlás szórása $\sigma=\sqrt{np(1-p)}$, azaz $\sqrt{n}/2$:

$$z=\frac{k-n/2}{\sqrt{n}/2}=\frac{2k-n}{\sqrt{n}},$$

és

$$\varphi_n(z)=\sigma\cdot f(k).$$

$n\to\infty$ esetén $\varphi_n(z)$ tart a következő $\varphi(z)$ függvényhez:

$$\varphi(z)=\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{e^{-z^2/2}}\cdot\left(\frac{e^-z}{e^z}\right)^{\frac{z}{2}}=\frac{1}{\sqrt{2\pi}}\cdot e^{\frac{-z^2}{2}}.$$

(A levezetést ld. [3]-ban.) Ezt a függvényt hívják a standard normális eloszlás sűrűségfüggvényének. Ez egy folytonos függvény, nem lépcsős, mint a binomiális eloszlásé, amelyet ezzel közelítünk:

4. ábra

Hogyan lehet ezzel most ténylegesen közelítő számításokat végezni? Számunkra általában binomiális összegek kellenek, pl.

$$\sum_{k\le x}P(X=k)=\sum_{k\le x}\frac{n!}{k!(n-k)!}\cdot p^k\cdot(1-p)^{n-k}.$$

Ha most a fenti közelítést akarjuk használni, akkor lévén, hogy folytonos függvényről van szó, a transzformációt elvégezve az $Y = \dfrac{X - np}{\sqrt{np(1-p)}}$ változó már standard normális eloszlású, azaz $P(Y < z) = \displaystyle\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x$

Az analízisből ismeretes azonban, hogy a fenti integrál csak közelítéssel számolható ki, nincs analitikus képlettel felírható primitív függvénye az $x \to e^{-x^2}$ függvénynek. Mivel nagyon gyakran használjuk ezt az integrált, van egy elfogadott jelölés: $P(Y < z) = \Phi(z) = \displaystyle\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x$. A Φ függvény értékeit táblázatban szokták megadni. Mivel φ szimmetrikus az origóra, ezért $\Phi(0) = 0,5$. Ebből a szimmetriából következik az is, hogy $\Phi(-z) = 1 - \Phi(z)$. Ezzel lehet negatív érékekre is meghatározni Φ-t. (Lásd a *4. ábrát, és az 1. táblázatot.*

Mivel a binomiális eloszlás lépcsős függvény, míg a normális eloszlás folytonos, jobb illeszkedést kapunk az ún. *korrekciós formulával* (lásd az *5. ábrát* is):

$$P(X \le x) = P\left(Y \le \frac{x + 0,5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{x + 0,5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{x + 0,5 - \mu}{\sigma}\right),$$

$$P(X < x) = P\left(Y \le \frac{x - 0,5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{x - 0,5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{x - 0,5 - \mu}{\sigma}\right),$$

$$P(X > x) = 1 - \Phi\left(\frac{x + 0,5 - np}{\sqrt{np(1-p)}}\right) = 1 - \Phi\left(\frac{x + 0,5 - \mu}{\sigma}\right),$$

$$P(X \ge x) = 1 - \Phi\left(\frac{x - 0,5 - np}{\sqrt{np(1-p)}}\right) = 1 - \Phi\left(\frac{x - 0,5 - \mu}{\sigma}\right),$$

5. ábra

Ezek szerint lehet számolni a binomiális eloszlás közelítésével. Van még egy fontos kérdés, hogy milyen nagy n-re jó ez a közelítés. A helyzet bonyolult, mert ez p-től is függ. Van egy még Laplace-tól származó gyakorlati kritérium, miszerint, ha

$$np(1-p) > 9,$$

akkor már legalább két tizedes pontossággal jó a közelítés, s természetesen ha $np(1-p)$ sokkal nagyobb, mint 9, akkor a közelítés rendje is lényegesen javul. Ennek elemzése azonban meghaladja ennek a cikknek a kereteit, s nem is volt célunk ezt megtenni, mégis a használat feltételeit tisztázni kellett. A részleteket lásd pl. [3] 195-207. oldalig, vagy [4] 142-147. oldal.

Gyakorlásul nézzük meg, mi van, ha 10%-kal több megrendelést veszünk fel, hogy tele legyen a repülőgép. A 10% esetünkben 25 fő. Vajon mit lehet mondani, ha 275 megrendelést veszek föl, akkor milyen eséllyel fér be minden utazni szándékozó a gépbe? Esetünkben a várható érték $np = 275 \cdot 0,9 = 247,5$, míg a szórás $\sqrt{np(1-p)} = \sqrt{2}75 \cdot 0,09 = 0,3\sqrt{2}75 \approx 4,975$.

Tehát $\Phi\left(\dfrac{250,5 - 247,5}{4,975}\right) = \Phi(0,603) = 0,7257$, azaz kb.72% az esélye, hogy 275 megrendelést felvéve még nem kerülünk kellemetlen helyzetbe.

Ezután próbáljuk meg az eredeti kérdést megválaszolni, tehát nem a megrendelésszám ismert, hanem annak az esélyét kérdezzük, vajon nem lép-e fel a visszamondás egy bizonyos rögzített esélynél kisebb valószínűséggel. Már egy kicsit tájékozottabbak vagyunk, hiszen pl. 99%-ra rögzítjük az esélyt, akkor tudjuk, hogy 252 megrendelés még bőven jó, s a 275 meg már túl sok. Természetesen végigpróbálhatnánk a kettő között minden számot, s kiderülne, hol lépjük át 0,99-es küszöböt. Azonban éppen azért csináltuk végig a fenti közelítő számolást, hogy annak hasznát vegyük. Tehát a kérdésünk a következőképpen szól:

Legfeljebb milyen nagy n-ekre lehet 0,99 eséllyel 250-nél kisebb vagy egyenlő az utazók száma? Ha most az utazók számát normális eloszlással közelítjük, amelynek várható értéke $\mu = n \cdot 0,9$, szórása $\sigma = \sqrt{n \cdot 0,9 \cdot 0,1} = 0,3\sqrt{n}$, akkor ehhez az kell, hogy az $Y = \dfrac{X - 0,9n}{0,3\sqrt{n}}$ új változó kisebb legyen, mint $\dfrac{250 - 0,9n}{0,3\sqrt{n}}$ legalább 99% eséllyel. Mivel az új változó standard normális eloszlású, azért látható az *1. táblázatból*, hogy 2,33-nál már több, mint 99% eséllyel lesz kisebb a véletlen számunk, amely az utazók számát jelöli az n jelentkező közül.

Tehát $\dfrac{250 - 0,9n}{0,3\sqrt{n}} > 2,33$, vagyis $250 - 0,9n > 0,699\sqrt{n}$. Mivel a bal oldal pozitív, azért $250 - 0,9n$ is az kell legyen, tehát $n < 277$. Négyzetre emlve:

$$62500 - 450n + 0,81n^2 > 0,4886 \Leftrightarrow 81n^2 - 45048,86n + 6250000 > 0.$$

Ennek megoldása $n < 265,13$ vagy $n > 291,03$. Az utóbbit kizárja az $n < 277$ kikötés, tehát $n < 265,13$ vagyis mivel n egész, ezért $n < 265$.

Ezzel választ adtunk a kérdésre, ha a biztonsági szint 99%-os, akkor legfeljebb 265 megrendelést vehetünk föl. Ezzel a várható utasszám $250 \cdot 0,9 = 225$-ről $265 \cdot 0,9 = 238,5$-re emelkedett.

Ez $13,5 \times 30\,000 = 405\,000 Ft$ bevétel növekedést jelent járatonként! S a „túlcsordulási" rizikó csak 1%. Természetesen kiszámolható más rizikó faktorral is a feladat, ezt gyakorlásul ajánjuk az olvasónak.

Ha 5%-os a rizikó faktor, akkor hány megrendelés vehető föl, s mi a helyzet 0,1% esetén? Mekkora átlagos bevétel növekedést jelentenek járatonként a fenti esetek?

Látható, hogy a rizikó faktor és a bevétel növekedés egyirányú, ha csökkentem a rizikót, csökken a bevétel is. Ebben már a managementnek kell döntenie. Nyilván a visszamondás egy esetleges hosszú távú utast taszít át a konkurenciához, s így nem szabad gyakorinak lennie.

Végezetül ne felejtsük el, hogy a függetlenség fontos feltétel volt. Ha az egyes utasok nem függetlenek, akkor kell információ arról, hogyan függnek össze. Ha valaki csoportosan utazik pl. családdal, munkatársakkal, akkor gyakran a kiesése az egész csoport kiesését vonja maga után. Ekkor még többen esnek ki, vagyis az esélyeink a visszamondásnál nőnek, tehát becslésünk még inkább igaz lesz. Ha emiatt még több megrendelést akarunk felvenni, akkor ennek kalkulálásához további információ kell az együttes visszamondások számáról, gyakoriságáról. Ez egy lehetséges útja modellünk finomabbá tételének. Tudni kell azonban, hogy minden ilyen lépés, bár az eredmények egyre közelebb lesznek a valósághoz, növeli a matematikai nehézségeinket, amelyeket manapság leginkább a számítógépek segítségével lehet megoldani. Szimulációs és egyéb programokkal, vagy egyszerűen csak a sokmillió számítás elvégeztetésével. Nekünk azonban most csak az ismerkedés volt a célunk.

A második probléma tisztázásához idézzük fel, miről van szó:

2. Probléma: (ismerve a sokaság eloszlását, mit lehet mondani egy minta eloszlásáról, amelyet ebből a sokaságból vesznek?)

Egy, az Országos Rendőrfőkapitányság és a Főügyészség által közölt statisztikából tudjuk, hogy Magyarországon 1990-ben 341 061 bűncselekmény vált ismertté (feljelentés, eljárás indítás), s ebből 187 655 esetben az elkövető ismeretlen maradt. Ha most kiválasztunk 1000 bűncselekményt a 341 061-ből, akkor vajon mekkora eséllyel lesz legalább 400 esetben ismeretlen az elkövető? Mit kell érteni kiválasztáson? Hogyan lehet megvalósítani a kiválasztás „véletlenszerűségét", amelyre a modellünk mond valamit?

Nagyon gyakori, hogy egy nagy sokaságról tudunk valamit, s arra vagyunk kíváncsiak, vajon hogyan tükröződik ez egy valahogyan kiválasztott jóval kisebb elemszámú mintában. Az alapvető kérdés persze éppen az lesz, hogy hogyan válasszuk ki ezt a mintahalmazt a teljes sokaságból, mint alaphalmazból. Azok a számítások, amelyeket majd elvégzünk, alapvetően arra támaszkodnak, hogy a *mintahalmazt valamilyen módon „egyenletesen véletlenszerűen" tudjuk kiválasztani,* amin természetesen azt értjük, hogy semelyik elemnek sincs „prioritása" a kiválasztási folyamat során. Ezt persze könnyebb kijelenti, mint megvalósítani. A legegyszerűbb eset, s természetesen egy bevezető példában ezt választottuk, amikor egyetlen egy tulajdonság fennállása szerint csoportosítjuk a sokaság elemeit. Vagyis az alaphalmaz elemei két osztályba sorolhatók, s az lesz a kérdés, hogy egy véletlenszerűen választott mintában milyen gyakorisággal várható a kétféle elem felbukkanása. Konkrétan a mi feladatunkban a felfedett, illetve a ki nem derített bűntények szerint osztályozzuk a bejelentett bűneseteket. Tudjuk, hogy a 341 061 esetből több mint a fele, 187 655 kiderítetlen maradt. Ez 0,55, vagyis az összes esetnek 55%-a.

Vajon az 1000 „véletlenszerűen" választott eset között hány felderítetlent találunk? Nyilván, ha biztosat akarunk mondani, akkor ez a szám 0 és 1000 között lesz. Vajon, ha egy kis rizikót vállalunk, akkor mennyire csökkenthető ez az intervallum?

Mennyire igaz az, amit sokszor gondolunk, hogy „körülbelül" 55% lesz a mintában is a kiderítetlen esetek száma? Mit jelent itt a „körülbelül" szó, lehet-e valamilyen kvantitatív értelmet adni neki? Mennyire valószínű, hogy például legalább 400 felderítetlen eset lesz?

Látható, hogy a kérdés nagyon hasonló az első problémához. Legalábbis ami a gondolkodásmódot illeti. A számolás sem lesz sokkal nehezebb, de itt is kell majd közelítést használni.

Kíséreljük meg modellezni a problémát. Egyszerű kombinatorikai kérdésről van szó: 341 061 esetből kell kiválasztani 1000 esetet, ami a teljes véletlenszerűség feltételezése mellett $\binom{341\,061}{1000}$-féleképpen tehető meg. Annak az esélye pedig, hogy éppen k felderítetlen eset lesz ezek között, a klasszikus Laplace-féle formulával számolva (kedvező esetek száma osztva az összes esetek számával): $\dfrac{\binom{187\,655}{k} \cdot \binom{153\,406}{100-k}}{\binom{341\,061}{1000}}$. Feladatunkra a válasz:

$$\sum_{k=400}^{1000} \frac{\binom{187\,655}{k} \cdot \binom{153\,406}{100-k}}{\binom{341\,061}{1000}}.$$

Ezt kellene különböző k értékek mellett $(0 < k < 1000)$ kiszámolni. Nyilván elég fáradságos munka lenne. Ezért érdemes megpróbálni valamilyen közelítést keresni. Azt az eloszlást, amely 0-tól n-ig terjedő természetes számokon

értelmezett a következőképpen:

$$P(X=k)=\frac{\binom{K}{k}\cdot\binom{N-K}{n-K}}{\binom{N}{n}}, k=0,1,\ldots,n \tag{1}$$

szokás *hipergeometrikus eloszlásnak* nevezni, lásd pl. [5]-ben.

Ez a következőképpen interpretálható: van N elem (az alapsokaság elemszáma), amelyből K valamilyen megadott tulajdonságú, míg természetesen $N-K$ azon elemek száma, amelyeknek nincs meg az adott tulajdonsága. Eztuán kiválasztunk véletlenszerűen n elemet (ez a minta), s azt kérdezzük, mekkora annak az esélye, hogy éppen k darab lesz $(0 \geq k \geq n)$ a K közül a mintában, feltéve, hogy $n < K$. Ha X jelöli a tulajdonsággal rendelkezők számát a mintában, akkor éppen az (1) alatti eloszlást kapjuk. A levezetésben valójában egy urnamodellt használunk, csak most eltérően az első problémától, visszatevés nélkül húzunk.

Ha figyelembe vesszük, hogy az urna elemszáma jóval nagyobb, mint a húzások száma, akkor lényegében nem játszik jelentős szerepet a „vissza nem tétel". Ezt felismerve várhatóan jó közelítés lesz a binomiális eloszlás, ahol tehát visszatesszük a húzás után a kihúzott elemet.

Ennek pontos matematikai elemzésére még visszatérünk.

Azaz:

$$P(X=k)=\frac{\binom{K}{k}\cdot\binom{N-K}{n-K}}{\binom{N}{n}}\approx\binom{n}{k}\cdot\binom{K}{N}^{k}\cdot\binom{N-K}{N}^{n-k}.$$

Esetünkben ez konkrétan:

$$\frac{\binom{187\,655}{k}\cdot\binom{153\,406}{100-k}}{\binom{341\,061}{1000}}\approx\binom{1000}{k}\cdot\binom{187\,655}{341\,061}^{k}\cdot\binom{153\,406}{341\,061}^{1000-k}=\binom{1000}{k}\cdot 0,55^{k}\cdot 0,45^{1000-k}.$$

Ezzel már kissé könnyebb számolni. Természetesen számítógéppel kiszámolható az eredmény, de ha az előző részben leírt standard normális eloszlással közelítünk, akkor egyszerű feladat lesz intervallumot megadni előírt rizikó mellett. Mivel esetünben $np(1-p)=247,5>9$, tehát teljesül a Laplace-kritérium, azért a közelítés „megengedett".

Átlagosan 550 körül lesz a fel nem fedett bűnesetek száma az ezerből, s hogy mekkora intervallumba kell esnie, az a rizikófaktortól függ. Esetünkben nem az intervallum, hanem a rizikófaktor a kérdés. A standard normális eloszlással becsülve adódik:

$$\sum_{k=400}^{1000}\binom{1000}{k}\cdot 0,55\cdot 0,45^{1000-k}\approx 1-\Phi\left(\frac{399,5-550}{\sqrt{2}247,5}\right)\approx 1-\Phi(-9,57)=\Phi(9,57).$$

Ez a szám viszont már nincs is benne a szokásos Φ táblázatokban, nagyobb mint 0,9999999, vagyis nagyon valószínű.

A binomiális eloszlás normálissal történő közelítéséről már volt szó. A hipergeometrikus eloszlás kapcsolata a binomiális eloszlással jóval könnyebben áttekinthető, ezért itt bemutatjuk. Akik kevésbé érdeklődnek a matematikai levezetések iránt – bár sokszor éppen abból lehet megérteni a dolog lényegét – azok most tovább lapozhatnak a 2. táblázatig.

Az alábbi közelítést szeretnénk megvizsgálni:

$$P(X=k)=\frac{\binom{K}{k}\cdot\binom{N-K}{n-K}}{\binom{N}{n}}\approx\binom{n}{k}\cdot\binom{K}{N}^{k}\cdot\binom{N-K}{N}^{n-k}.$$

A kérdés természetesen az, hogy milyen feltételek mellett lehet ezt használni, illetve mekkora numerikus hibát követünk el. Fejtsük ki a közelítő egyenlőség bal oldalát:

$$\frac{\binom{K}{k}\cdot\binom{N-K}{n-K}}{\binom{N}{n}}=\binom{n}{k}\cdot\frac{K\cdot(K-1)\cdot\cdots\cdot(K-k+1)}{N\cdot(N-1)\cdot\cdots\cdot(N-k+1)}\cdot$$

$$\cdot\frac{(N-K)\cdot(N-K-1)\cdot\cdots\cdot(N-K-n+k+1)}{(N-k)\cdot(N-k-1)\cdot\cdots\cdot(N-n+1)}$$

Erről az alakról már látszik, hogy mi lesz a közelítés lényege, az első tört esetén mindenütt elhagyjuk a kivonandót, mondván, hogy ha K és N nagy, míg n (és ezzel k is) kicsi, akkor k darab K/N szorzata az első tényező. Hasonlóképpen a második tényező esetén a számlálóban minden tényezőt $(N-K)$-nak véve, míg a nevezőben mindent N-nek véve adódik a közelítés. Aki tanult határérték számítást, az látja, hogy ha n és k fix számok, míg N és vele K tart a végtelenbe (úgy, hogy az arányuk közben állandó maradjon), akkor tényleg határértékben egyenlő lesz a hipergeometrikus eloszlás bármelyik tagja a binomiális eloszlás megfelelő tagjával.

A valóságban azonban K és N nem végtelen nagy, ezért jó lenne megvizsgálni, mekkora hibát követünk el ezzel a becsléssel. Ez nyilván függ N, K valamint n és k viszonyától. Nyilván az utolsó tényezők, azaz $K-k+1$ és $N-k+1$ térnek el legjobban K és N-től. A hibának tehát jó felső becslése a $\left(\frac{K}{N}-\frac{K-k+1}{N}\right)^k\cdot\left(\frac{N-K}{N}-\frac{N-K-n+k+1}{N}\right)^{n-k}$ szorzat. Ennél biztos kevesebbet hibázunk. Ennek értéke: $\left(\frac{k-1}{N}\right)^k\cdot\left(\frac{n-k-1}{N}\right)^{n-k}$, amit ismét felülről becsülhetünk. Ezúttal az első tényezőben k helyett a legnagyobb lehetséges értéket, n-et írva, és a -1-et elhagyva, valamint a második tényezőben a legkisebb k-t, 0-t írva és ismét elhagyva a -1-et, adódik, hogy a hiba univerzálisan bármelyik tagot nézve, azaz *tetszőleges k esetén biztos kisebb, mint* $\left(\frac{n}{N}\right)^n$. Ez egy nagyon durva felső becslés, de még ez is mutatja, hogy nem követünk el nagy hibát, még viszonylag nagy húzás százalék mellett sem.

Esetünkben $n=1000$, míg $N=341\,061$, a hiba kisebb, mint $0,0029^{1000}$, ami $\approx 10^{-2537}$, azaz elenyésző, ennél a normális eloszlással történő közelítés jóval nagyobb hibát produkál.

Összefoglalva tehát levezetésünket, megállapítható, hogy ha az elemszám legalább egy nagyságrenddel nagyobb a húzásszámnál, akkor semelyik tag közelítésében nem követünk el $0,1^n$-nél nagyobb hibát, ami legalább öt húzás esetében az eredményt legfeljebb a hatodik tizedesjegyben változtatja meg. Valójában még ennél is kisebb a hiba, nem beszélve arról, ha a húzásszámnál több nagyságrenddel nagyobb az elemek száma. Természetesen, ha ez nem áll fenn, akkor nagy különbségek is felléphetnek, s a közelítés abszolút rossz, lásd például az alábbi esetet:

Legyen egy urnában 20 golyó, 12 piros 8 fehér. 10-szer húzunk visszatevés nélkül. Vajon mekkora eséllyel lesz k piros a húzott golyók között?

Már tudjuk, hogy ezt hogyan számoljuk: az esély$= \dfrac{\binom{12}{k}\cdot\binom{8}{10-k}}{\binom{20}{10}}$.

A megfelelő közelítés ebben az esetben az a binomális eloszlás lenne, amelynek paraméterei: $n=20, p=0,6$, azaz annak az esélye, hogy éppen k piros lesz a 10 húzás során: $B_{20;0.6}(k)=\binom{20}{k}\cdot 0,6^k\cdot 0,4^{10-k}$.

2. táblázat. Visszatevés nélküli húzás

Lehetséges értékek	esélyek
0	$\frac{\binom{12}{0}\cdot\binom{8}{10}}{\binom{20}{10}}=0$
1	$\frac{\binom{12}{1}\cdot\binom{8}{9}}{\binom{20}{10}}=0$
2	$\frac{\binom{12}{2}\cdot\binom{8}{8}}{\binom{20}{10}}=0,0003572$
3	$\frac{\binom{12}{3}\cdot\binom{8}{7}}{\binom{20}{10}}=0,0095261$
4	$\frac{\binom{12}{4}\cdot\binom{8}{6}}{\binom{20}{10}}=0,0750179$
5	$\frac{\binom{12}{5}\cdot\binom{8}{5}}{\binom{20}{10}}=0,2400572$
6	$\frac{\binom{12}{6}\cdot\binom{8}{4}}{\binom{20}{10}}=0,3500834$
7	$\frac{\binom{12}{7}\cdot\binom{8}{3}}{\binom{20}{10}}=0,2400572$
8	$\frac{\binom{12}{8}\cdot\binom{8}{2}}{\binom{20}{10}}=0,0750179$
9	$\frac{\binom{12}{9}\cdot\binom{8}{1}}{\binom{20}{10}}=0,0095261$
10	$\frac{\binom{12}{10}\cdot\binom{8}{0}}{\binom{20}{10}}=0,0003572$

A hipergeometrikus eloszlás számolásakor figyelembe kell venni, hogy ha $10-k>8$, akkor a számláló 0 lesz, hiszen nem lehet 8 elemből 9 elemet kiválasztani.
Abban érdemes megállapodni, hogy $\binom{n}{k}=0$, ha $k>n$.

Összehasonlításképpen a binomiális eloszlás megfelelő értékeivel:

Lehetséges érték	Húzás visszatevéssel	Húzás visszatevés nélkül
0	0,0001	0
1	0,0016	0
2	0,0106	0,0003572
3	0,0425	0,0095261
4	0,1115	0,0750179
5	0,2007	0,2400572
6	0,2508	0,3500834
7	0,2150	0,2400572
8	0,1209	0,0750179
9	0,0403	0,0095261
10	0,0060	0,0003572

A második problémára adott válaszunk tehát az, hogy nagyon nagy (>0,9999999) eséllyel lesz több, mint 400 a fel nem fedett bűesetek száma, amely várhatóan 550 körüli lesz az 1000 esetből, s hogy mekkora intervallumba esik nagy eséllyel (tehát kicsi a rizikója annak, hogy ezen kívül essen) azt ki lehet számítani, ha előre adott, mit akarunk nagy esélyűnek nevezni. Itt is igaz, hogy jelentősen megrövidül az eredeti (0-1000) intervallum, ha egy igen pici rizikót merünk vállalni.

Egyvalamivel még adósak vagyunk, nem mondtuk meg, hogy hogyan lehet véletlenszerűen választani, ami az egész gondolatmenet alapja volt. Gondoljuk meg, ha ez nem teljesül, akkor akár úgy is választhatok, hogy a fel nem fedett bűnesetek közül veszem mind az 1000 esetet, s így biztos, hogy 100% lesz a fel nem derítettek száma.

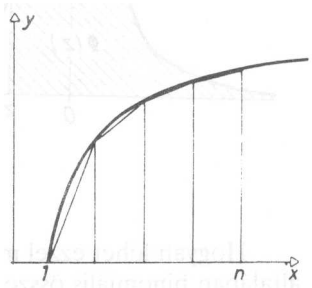
Egy, a lottónál szokásos lehetőség, az urnából húzás lenne. Ez azonban aligha valósítható meg 341 061 golyóval, közben az egyenletességet folyamatosan keveréssel biztosítva. Jobb lehetőség kínálkozik a számítógépek felhasználásával, amit TV-s játékokból ismerhetünk. Az eljárás a következő: Sorbarendezzük valahogy az összes esetet, az embereket pl. a személyi számuk szerint, majd elkezdjük a gépen a listát „futtatni", s valamikor tetszés szerinti pillanatban a listát megállítani. Ahol éppen tart a futás, az lesz az első elem. Ezután ezt töröljük, s az egész eljárás kezdődik elölről. (Ha nem töröljük az elemet, akkor a visszatevéses eljárás is szimulálható így számítógép segítségével.)

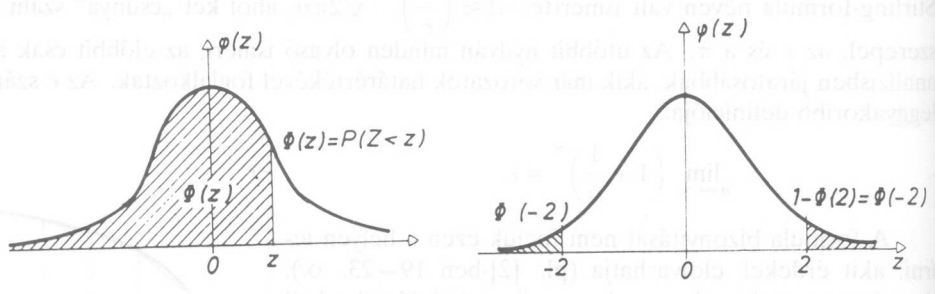
Természetesen itt a megállítás véletlenszerűségére tevődött át a probléma. Ennek biztosítása sem egyszerű. Kis esetszám esetén persze lehet valami „igazi véletlen szerint" megállítani, de mivel a futás hihetetlenül gyors az emberi reakcióidőhöz képest, kis esetszám esetén elég valamikor lenyomni a „megállító" billentyűt, az teljesen véletlenszerű lesz. Nagy esetszámnál lehet, hogy valamiiyen ritmusra állunk be, s ez már esetleg determinisztikussá teszi az eljárást. Jó lenne persze teljesen automatizálni a dolgot, ez is megtehető (legalábbis valamilyen pszeudo-véletlen szinten), de ebben már a számítógép-programozók illetékesek.

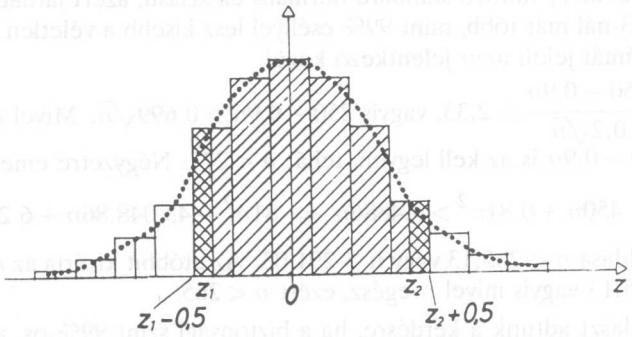
Vancsó Ödön ELTE TTK Matematika Szakmódszertani Csoport

Irodalom

[1] Hajnal – Nemetz – Pintér – Urbán: Matematika IV. (B fakt), Tankönyvkiadó, 1982
[2] Császár Ákos: Végtelen sorok, Egyetemi jegyzet, Tankönyvkiadó, 1977.
[3] H. Ch. Reichel: Wahrscheinlichkeitsrechnung und Statistik, Verlag Hölder – Pichler – Tempsky, Wien, 1989
[4] Rényi Alfréd: Valószínűségszámítás, Tankönyvkiadó, 1968
[5] Nemetz Tibor: Valószínűségszámítás, Tankönyvkiadó.







z → 0	1	2	3	4	5	6	7	8	9
0,00 0,50000	50399	50798	51197	51595	51994	52392	52790	53188	53586
0,10 53983	54380	54776	55172	55567	55962	56356	56749	57142	57535
0,20 57926	58317	58706	59095	59483	59871	60257	60642	61026	61409
0,30 61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0,40 65542	65910	66276	66640	67003	67364	67724	68082	68439	68793
0,50 0,69146	69497	69847	70194	70540	70884	71226	71566	71904	72240
0,60 72575	72907	73237	73565	73891	74215	74537	74857	75175	75490
0,70 75804	76115	76424	76730	77035	77337	77637	77935	78230	78524
0,80 78814	79103	79389	79673	79955	80234	80511	80785	81057	81327
0,90 81594	81859	82121	82381	82639	82894	83147	83398	83646	83891
1,00 0,84134	84375	84614	84849	85083	85314	85543	85769	85993	86214
1,10 86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1,20 88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1,30 90320	90490	90658	90824	90988	91149	91309	91466	91621	91774
1,40 91924	92073	92220	92364	92507	92647	92785	92922	93056	93189
1,50 0,93319	93448	93574	93699	93822	93943	94062	94179	94295	94408
1,60 94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1,70 95543	95637	95728	95818	95907	95994	96080	96164	96246	96327
1,80 96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1,90 97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2,00 0,97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2,10 98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2,20 98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2,30 98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2,40 99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2,50 0,99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2,60 99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2,70 99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2,80 99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2,90 99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3,00 0,99865	99869	99874	99878	99882	99886	99889	99893	99897	99900
3,10 99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3,20 99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3,30 99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3,40 99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3,50 0,99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3,60 99984	99985	99985	99986	99986	99987	99987	99988	99988	99989
3,70 99989	99990	99990	99990	99991	99991	99992	99992	99992	99992
3,80 99993	99993	99993	99994	99994	99994	99994	99995	99995	99995
3,90 99995	99995	99996	99996	99996	99996	99996	99996	99997	99997
4,00 0,99997	99997	99997	99997	99997	99997	99998	99998	99998	99998