

Matricák és tankok

Salvi Péter

2026. szeptember 19.

A közeli ábécében adott értékű vásárlás után gyűjtögetős matricákat adnak, amelyeket aztán a sorsszámuk alapján egy albumba be lehet ragasztani. Zsófi már 12 matricát gyűjtött (nem mind különböző), a legnagyobb sorsszámú közülük a 100-as. Albuma viszont még nincsen, ezért kíváncsi: vajon összesen hány matricának lehet benne hely?

Formalizáljuk a feladatot. Az $1 \dots n$ számokból választunk visszatevéssel k darabot, és ezekből a maximum m . Hogyan lehet megtippelni n értékét a k és m ismeretében?

1. A legnagyobb valószínűség

Egy elég természetes átfogalmazása a feladatnak, hogy milyen n értékre lesz $P(M = m \mid N = n)$ maximális.¹ Ez a valószínűség

$$P(M = m \mid N = n) = \left(\frac{m}{n}\right)^k - \left(\frac{m-1}{n}\right)^k = \frac{m^k - (m-1)^k}{n^k}. \quad (1)$$

A középső képletben az első tag annak a valószínűsége, hogy k db. legfeljebb m sorsszámú matricát kapunk, de ebben benne van az az eset is, amikor minden matrica sorszáma kisebb m -nél, ennek az esetnek a valószínűségét tehát kivonjuk a második tagban.

Viszont ennek a kifejezésnek csak a nevezőjében szerepel az n , ezért n minél nagyobb, az eredmény annál kisebb, tehát a lehető legkisebb n -re kapjuk a maximális értéket, vagyis

$$N = m, \quad (2)$$

ami a mi esetünkben 100. Ez az eredmény nem igazán intuitív, hiszen független k -től, és nyilvánvalóan alulbecsüli n tényleges értékét.

Érdemes ezért inkább azt megnézni, hogy az (1) képletben látott ún. *likelihood* függvény az N milyen intervallumán veszi fel a maximális értékének legalább (mondjuk) 10%-át. Tehát ha $L(N) := P(M = m \mid N = n)$, akkor a legnagyobb n értéket keressük, amelyre

$$\frac{L(n)}{L(m)} = \left(\frac{m}{n}\right)^k \geq 0.1 \quad (3)$$

¹A jelölés jelentése: „Mi a valószínűsége annak, hogy M az m értéket veszi fel, ha tudjuk, hogy N értéke n ?” A nagybetűk valószínűségi változókat, a kisbetűk konkrét értékeket jelölnek.

még teljesül. Ez a feladatban szereplő értékekre az $n = 121$, mivel $(100/121)^{12} \approx 0.102$, de $(100/122)^{12} \approx 0.092$. Ez alapján azt mondhatjuk, hogy a megfigyelt $m = 100$ érték hihető, ha n értéke a $100 \dots 121$ intervallumba esik.

2. Várható érték

Egy másik megközelítés, hogy visszafelé gondolkozunk. Mi a várható értéke M -nek, ha ismerjük n -et és k -t? A várható érték nem más, mint a lehetséges értékek összege súlyozva a valószínűségükkel, tehát

$$E(M) = \sum_{m=1}^n mP(M=m). \quad (4)$$

Egy kis trükkel ezt átírhatjuk egy kényelmesebb alakra:

$$E(M) = \sum_{m=1}^n P(M \geq m). \quad (5)$$

Ez ugyanaz, mert ha kibontjuk,

$$\begin{aligned} P(M \geq 1) &= P(M=1) + P(M=2) + P(M=3) + \dots + P(M=n), \\ P(M \geq 2) &= P(M=2) + P(M=3) + \dots + P(M=n), \\ P(M \geq 3) &= P(M=3) + \dots + P(M=n), \\ &\vdots \\ P(M \geq n) &= P(M=n), \end{aligned}$$

tehát az összegben minden $P(M=m)$ pontosan m -szer szerepel.

Erről viszont könnyen látható, hogy

$$P(M \geq m) = 1 - P(M < m) = 1 - \left(\frac{m-1}{n}\right)^k, \quad (6)$$

tehát akkor

$$E(M) = n - \sum_{m=1}^n \left(\frac{m-1}{n}\right)^k, \quad (7)$$

és arra vagyunk kíváncsiak, hogy ez milyen esetben lesz kb. akkora, mint az általunk megfigyelt m érték (100). Ez számítógéppel könnyen kiszámolható, de lehet közvetlen képletet is adni rá, ha ki tudjuk fejezni n -et. Ehhez megint egy kis trükkre lesz szükség. A

$$\sum_{m=1}^n \left(\frac{m-1}{n}\right)^k \quad (8)$$

összeg nem más, mint egy közelítése az $\left(\frac{x}{n}\right)^k$ függvény alatti területnek a $[0, n]$ intervallumon, ha azt 1-széles téglalapokra felosztjuk úgy, hogy minden téglalap bal felső sarka érintse a függvényt. Ez tehát kb.

$$\int_0^n \left(\frac{x}{n}\right)^k dx = \frac{n}{k+1} \left[\left(\frac{x}{n}\right)^{k+1} \right]_0^n = \frac{n}{k+1}, \quad (9)$$

amiből

$$E(M) \approx n - \frac{n}{k+1} = \frac{nk}{k+1}. \quad (10)$$

Ha most feltételezzük, hogy a megfigyelt m értékünk a várható érték, akkor a fenti képletből n -et kifejezve azt kapjuk, hogy

$$n = m \frac{k+1}{k}, \quad (11)$$

azaz a mi esetünkben $100 \cdot \frac{13}{12} \approx 108$.²

3. Medián

A fenti gondolatmenet szépséghibája, hogy a várható érték általában nem egész szám; logikusabbnak tűnik tehát a mediánt alkalmazni. Az M mediánja adott $N = n$ érték mellett az a legkisebb m érték, melyre

$$P(M \leq m | N = n) = \left(\frac{m}{n}\right)^k \geq \frac{1}{2}. \quad (12)$$

Ezután már csak meg kell keresni azt az n értéket, ahol a medián értéke m (vagy ehhez legközelebb esik). Egy ilyen programot mutat a Függelék; a mi példánkban ez n -re 105-öt ad.

4. Bayes-tétel

Manapság divatos az ilyen kérdésekre úgy gondolni, hogy milyen mértékben változik meg a véleményünk. Ehhez természetesen szükség van arra, hogy legyen valamilyen véleményünk az N eloszlásáról, ezt hívják *prior*nak; a mi esetünkben a legegyszerűbb azt mondani, hogy nincsen semmilyen előfeltevésünk N -ről, tehát $P(N = n) \propto 1$ (egyenletes, „improper” prior). A Bayes-tétel szerint

$$P(N = n | M = m) = \frac{P(M = m | N = n) \cdot P(N = n)}{P(M = m)}, \quad (13)$$

ahol $P(A | B)$ annak a valószínűsége, hogy A , ha feltételezzük, hogy B . Mivel m -et rögzítettnek vesszük, a számláló arányos n^{-k} -val, amiből normalizálással

$$P(N = n | M = m) = \frac{n^{-k}}{\sum_{j=m}^{\infty} j^{-k}} = \frac{n^{-k}}{\zeta(k, m)}, \quad (14)$$

ahol ζ a Hurwitz-féle zéta-függvény. Ennek az eloszlásnak a maximuma nyilvánvalóan $N = m$ esetén lesz, de ez nem azt jelenti, hogy ez az esély nagy. A konkrét esetünkben $P(N = 100 | M = 100) \approx 10\%$. Érdekesebb ezért itt is az átlagot (≈ 109) vagy a mediánt (106) nézni, ezek több információt adnak a *posterior* eloszlás középpontjáról. (Ezek kiszámításához ld. a programot a Függelékben.)

Persze valószínűbb a konkrét esetünkben olyan priort használni, amely jobban tükrözi az elvárásainkat: tudjuk, hogy biztosan nincs 500-nál több matrica, és talán érdemes a nagyobb albumokat valószínűtlenebbé tenni, pl. $P(N = n) = 1/n$ priorral. Ezek végigszámolása már az olvasó feladata.

²Az Euler–Maclaurin formula segítségével egy kicsit tudunk javítani a közelítésen, ez alapján $E(M) \approx \frac{nk}{k+1} + \frac{1}{2}$, amiből n -re az $(m - \frac{1}{2}) \frac{k+1}{k}$ képletet kapjuk.

Epilógus

Honnan jönnek a címben levő tankok? A második világháború során fontos kérdés volt, hogy a németek milyen ütemben gyártják a tankokat. A kémek gyanúsan nagy mennyiségeket jelentettek. Sok elfogott tankon vagy alkatrészen azonban szerepelt dátum és gyártási szám, és az azonos időszakban adott gyártási számok alapján – a fenti 2. ponthoz nagyon hasonló módon – igen pontos becsléseket tudtak adni a gyártás valódi ütemére vonatkozólag. Ezeket utólag össze is tudták hasonlítani a tényleges adatokkal: 1941 júniusában pl. 244-et adott a statisztika a valódi 271 helyett, míg a kémek ugyanezt több, mint 1500-ra becsülték (ld. Wikipedia: *German tank problem* szócikk).

A. Medián alapú becslés

```
#include <math.h>
#include <stdio.h>

int main() {
    int m = 100, k = 12, n = m;
    while (1) {
        int median = 1;
        while (pow((double)median / n, k) < 0.5)
            median++;
        if (median == m)
            break;
        n++;
    }
    printf("%d\n", n);
}
```

B. Egyenletes priorú eset számítása

```
#include <math.h>
#include <stdio.h>

#include <gsl/gsl_sf.h>

int main() {
    int m = 100, k = 12, n = m;
    double zkm = gsl_sf_hzeta(k, m);
    while (gsl_sf_hzeta(k, n + 1) > zkm / 2)
        n++;
    printf("Átlag: %lf, medián: %d\n",
           gsl_sf_hzeta(k - 1, m) / zkm,
           n);
}
```