

Sünnipäevaparadoks (materjal harjutustunniks)

Ahto Buldas Aivo Jürgenson

19. aprill, 2006

Kujutlege, et ruumis viibib k juhuslikult sinna sattunud inimest ja nad hakkavad rääkima, kes mis päeval sündinud on (jätame eranliku kuupäeva 29. veebruari mängust välja). Küsime, et kui suur peab olema k , et tõenäosusega vähemalt $\frac{1}{2}$ sattuksid ruumi kaks samal päeval sündinud inimest. Selgub, et vastus on üllatavalt väike: $k \approx 23$.

Järgnevalt tõestame selle fakti eeldusel, et sünnipäevad on jaotunud umbes ühtlaselt, st mis tahes fikseeritud kuupäeval on sündinud umbes sama palju inimesi. Intuitiivselt on selge, et kui sünnipäevad ei oleks ühtlase jaotusega, siis see asjaolu ainult suurendaks tõenäosust, et kahel inimesel on samal päeval sünnipäev. Siiski anname ka sellele intuitiivsele faktile range tõestuse.

1 Tõestus ühtlase jaotuse korral

Sündmus X - "2 inimesel grupist on samal päeval sünnipäev".

$$\Pr[X] \geq \frac{1}{2}$$

Grupp inimesi on siis

$$G = \{I_1, I_2, \dots, I_n\}$$

Inimeste arv $k = |G|$

Tuleb välja, et

$$k \geq 23 \implies \Pr[X] \geq \frac{1}{2}$$

Hakkame järjest valima juhuslikult inimesi ning arvutame tõenäosuse, et nende inimeste sünnipäevad on erinevad.

Tähistame, et inimesel $I_i \in G$ on sünnipäev päeval $P_k \in S$ kui seost

$$f : G \rightarrow S$$

$$f(I_i) = P_k$$

Erinevaid sünnipäevasid on $|S| = 365 = N$.

Valime juhuslikult ja ühtlaselt $G \rightarrow I_1$ ning leiame selle inimese sünnipäeva $f(I_1)$.

Nüüd valime juhuslikult ja ühtlaselt $G \rightarrow I_2$ ning leiame selle inimese sünnipäeva $f(I_2)$.

Tõenäosus, et sünnipäev $f(I_1) = f(I_2)$ on $\frac{1}{N}$, kuna soodsaid võimalusi on 1 ning mittedoodsaid on 365 . Vastandsündmuse, sünnipäevad erinevad, tõenäosus on siis

$$\Pr[f(I_1) \neq f(I_2)] = 1 - \frac{1}{N}$$

Kui me teeme veel ühe katse, ning valime juhuslikult ja ühtlaselt $G \rightarrow I_3$ ning leiame selle inimese sünnipäeva $f(I_3)$, siis tõenäosus, et tema sünnipäev erineb eelmistest on

$$\Pr[f(I_1) \neq f(I_2) \neq f(I_3)] =$$

$$\Pr[f(I_1) \neq f(I_2) \wedge (f(I_2) \neq f(I_3) \wedge f(I_1) \neq f(I_3))] = \frac{N-1}{N} \cdot \frac{N-2}{N}$$

sest teisel sünnipäeval on erinemiseks $N-1$ soodsat võimalust ning kolmandal sünnipäeval on erinemiseks $N-2$ soodsat võimalust. See võrdub aga

$$\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)$$

Seega kui teeme k katse, ning valime juhuslikult ja ühtlaselt $G \rightarrow I_k$ siis on sellel k -ndal sünnipäeval $N-k+1$ soodsat võimalust eelmistest erinemiseks. Tähistame sündmuse, et "kõigi k inimese sünnipäevad erinevad üksteisest" tähega Y ning seega tõenäosus võrdub

$$\begin{aligned} \Pr[Y] &= \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right) \dots \left(1 - \frac{k-1}{N}\right) \\ &= \prod_{i=1}^k \left(1 - \frac{i-1}{N}\right) \end{aligned}$$

Nüüd kasutame omadust, et

$$1 + a \leq e^a$$

Seda saab vaadata nii graafiliselt, kui ka selle järgi, et

$$e^x = 1 + x + \frac{x^2}{2!} + \dots$$

Seega siis saame asendada korrutise teguri $\left(1 - \frac{i-1}{N}\right) = (1+a) \leq e^a = \exp\left(-\frac{i-1}{N}\right)$. Kui me korrutise teguri asendame suurema teguriga, siis läheb korrutis ise ka suuremaks. Seega, meie sündmuse tõenäosus on väiksem-võrdne sellest suuremast väärtusest:

$$\Pr[Y] = \prod_{i=1}^k \left(1 - \frac{i-1}{N}\right) \leq \prod_{i=1}^k \exp\left(-\frac{i-1}{N}\right)$$

Korrutamisel astendajad liidetakse, seega

$$\Pr[Y] \leq \exp\left(\sum_{i=1}^k -\frac{i-1}{N}\right) = \exp\left(\frac{1}{N} \sum_{i=1}^k -(i-1)\right)$$

Summamärgi all on jada $-0 - 1 - 2 - 3 - 4 - 5 \dots - (k-1)$ mille summa võrdub $\frac{0-(k-1)}{2} \cdot k$ ning seega tõenäosus

$$\Pr[Y] \leq \exp\left(-\frac{k(k-1)}{2N}\right)$$

Nüüd, vastandsündmus X on see, et vähemalt kaks sünnipäeva võrduvad üksteisega. Seega

$$\Pr[X] = 1 - \Pr[Y]$$

Kuna $\Pr[Y]$ on meil hinnatud, siis saame ka $\Pr[X]$ võrratusega ning

$$\Pr[X] = 1 - \Pr[Y] \geq 1 - \exp\left(-\frac{k(k-1)}{2N}\right).$$

Meie eesmärk on leida selline k , et $\frac{1}{2} \leq \Pr[X]$. Võtame üldkujul, et $\epsilon \leq \Pr[X]$

$$\epsilon \leq \Pr[X] \geq 1 - \exp\left(-\frac{k(k-1)}{2N}\right)$$

$$1 - \exp\left(-\frac{k(k-1)}{2N}\right) \geq \epsilon$$

$$\exp\left(-\frac{k(k-1)}{2N}\right) \leq 1 - \epsilon$$

Võtame mõlemalt poolt naturaallogaritmi, saame

$$-\frac{k(k-1)}{2N} \leq \ln(1 - \epsilon)$$

Korrutame $-2N$ mõlemalt poolt, saame

$$k^2 - k + 2N \cdot \ln(1 - \epsilon) \geq 0$$

Lahendades vastava ruutvõrrandi k suhtes, saame

$$k_{1,2} = \frac{1}{2} \pm \sqrt{\frac{1}{4} - 2N \cdot \ln(1 - \epsilon)}$$

Kuna ruutvõrrandi ruutliikme kordaja on positiivne, siis on parabool avatud ülesse ning võrratusele sobivad lahendid on vasakult- ja paremalt poolt nullkohti.

Meid huvitab parempoolne ala. Kuna $1 - \epsilon < e$, siis vastav naturaallõgarm on negatiivne, ning ruutjuure all on lõppkokkuvõtteks positiivne arv. Jätame ruutjuure alt ära $\frac{1}{4}$, kuna see on teise liidetavaga võrreldes väga väike suurus ning saame, et

$$k \geq \sqrt{-2N \cdot \ln(1 - \epsilon)} - \frac{1}{2}$$

Juhul, kui $\epsilon = \frac{1}{2}$, siis

$$k \geq \sqrt{2N} - \frac{1}{2}$$

Me oleme huvitatud täisarvudest ning juhul, kui $N = 365$, siis

$$k \geq \left\lfloor \sqrt{2 \cdot 365} - \frac{1}{2} \right\rfloor \simeq [22, 522] = 23$$

2 Rakendus: Räsifunktsioonid

Kui nüüd sünnipäevade pealt minna üldisemaks, siis me lahendasime olukorra, kus meil on sisendhulk X , väljundhulk Y ning meil on mingi seos $f : X \rightarrow Y$ ning

Kui $|X| > |Y|$, siis on põhimõtteliselt tegemist räsifunktsiooniga. Näiteks MD5 : $\{0, 1\}^n \rightarrow \{0, 1\}^{128}$. Kui me püüame leida räsifunktsioonile kollisiooni, siis me püüame leida sellist sisendit X' , mille puhul $\text{MD5}(X) = \text{MD5}(X')$, ehk siis me püüame leida sellist kahte inimest, kellel on sama sünnipäev?

Selleks peab otsinguruumi suurus olema suurem kui $\ln(1, 2 \cdot \sqrt{2^{128}}) \simeq 64$ ning tõenäosus, et nende sees on kaks sellist sisendit, mis annavad samasuguse räsi, on suurem kui $1/2$.

3 Mitteühtlane jaotus

Eraldi küsimus on selles, et kas selline ühtlaselt valimine on kõige mõistlikum. Äkki on meil selline räsifunktsioon turvalisem, mis annab mingid väljundid tõenäolisemalt, kui muud jaotused?

Olgu siis selline funktsioon $f : X \rightarrow Y$, mille korral $y_i \in Y$ saamise tõenäosus ühtlase X jaotuse korral on p_i . Sealjuures defineerime $N = |Y|$ ja $k = |X|$.

$$\sum_{i=1}^N p_i = 1$$

Leiame sellisel juhul tõenäosuse, et k -liikmelise grupi korral ei leidu 2 sellist x_i ja x_j , et $f(x_i) = f(x_j)$.

Saame ehitada "sündmuste puu", mille harud on järjest elementide valimise sündmused. Kõik harud on üksteisest sõltumatud ning kokkuvõttes tähendab

see puu kõiki võimalusi, kus k liikmelise grupi korral on kõik funktsiooni väärtused sellel grupil erinevad.

Sellise puu toimumise tõenäosus on

$$\sigma = \sum_{\{(N)_k\}} k! \cdot p_{n_1} \cdot p_{n_2} \cdots p_{n_k}$$

Kus $\{(N)_k\}$ tähistab k -kaupa kombinatsioone üle indeksite $1, 2, \dots, N$. Selliseid kombinatsioone on kokku $\binom{N}{k}$ tükki. Summamärgi all on $k!$ kuna tõenäosus p_i võib esineda ükskõik millises positsioonis, kuid korrutise väärtus sellest ei muutu ning seega võime sellised korrutised (ehk sündmuste puu erinevad harud) kokku grupeerida.

Seda tõenäosust võib vaadelda kui funktsiooni

$$\sigma = \sigma(p_1, p_2, \dots, p_N, N, k),$$

mis annab meile tõenäosuse, et k -liikmelise grupi korral meie funktsiooni $f(x_i)$ kõik väärtused on erinevad. Kui meil õnnestub näidata, et ühtlase jaotuse korral on see funktsioon maksimaalne, siis me olemegi küsimusele vastanud.

Oletame, et meie esialgne jaotus p_1, p_2, \dots, p_N on mitteühtlane ning $p_1 > p_2$.

Sellisel juhul saame me selle jaotuse muuta "ühtlasemaks", kui me asendame esimese tõenäosuse

$$p'_1 = p_1 - e$$

ja teise tõenäosuse

$$p'_2 = p_2 + e,$$

kus

$$e = \frac{p_1 - p_2}{2}$$

Eraldi küsimus on, et kas me saame selliseid operatsioone juhuslike paaride peal korrates lõpuks ühtlase jaotuse või mitte.

Nõnda saame me uue jaotuse, mis koosneb tõenäosustest $p'_1, p'_2, p_3, \dots, p_N$ ning mis samamoodi määrab väljundsuuruse käitumise. Niiet me saame arvutada

$$\sigma'(p'_1, p'_2, p_3, \dots, p_N, N, k)$$

Näitame, et

$$\sigma < \sigma'$$

Selleks teisendame σ arvutamise summa mitmeks summaks, kuid jätame märkimata, millised liikmed on meil summamärgi all

$$\begin{aligned} \sigma = & \sum_{p_1 \in \binom{N}{k} \wedge p_2 \notin \binom{N}{k}} + \sum_{p_2 \in \binom{N}{k} \wedge p_1 \notin \binom{N}{k}} + \\ & + \sum_{p_1 \wedge p_2 \in \binom{N}{k}} + \sum_{p_1 \wedge p_2 \notin \binom{N}{k}} \end{aligned}$$

Kuna esimes osasummas on kõigis liidetavates sees element p_1 , siis me saame ta summamärgi ette tuua. Samamoodi teises osasummas p_2 ning kolmandas osasummas $p_1 p_2$. Vastavalt muutuvad väiksemaks ka kombinatsioonide hulgad, üle mille me summeerime

$$\begin{aligned} \sigma = & p_1 \cdot \sum_{p_1 \notin \binom{N}{k-1} \wedge p_2 \notin \binom{N}{k-1}} + p_2 \cdot \sum_{p_2 \notin \binom{N}{k-1} \wedge p_1 \notin \binom{N}{k-1}} + \\ & + p_1 p_2 \cdot \sum_{p_1 \wedge p_2 \notin \binom{N}{k-2}} + \sum_{p_1 \wedge p_2 \notin \binom{N}{k}} \end{aligned}$$

Tähistame need osasummad suure sigma tähe ja indeksiga ning ei kirjuta enam edaspidi täpselt välja, üle millise hulga need summad on arvatud. Seega siis

$$\begin{aligned} \sum_{p_1 \notin \binom{N}{k-1} \wedge p_2 \notin \binom{N}{k-1}} &= \Sigma_1 \\ \sum_{p_2 \notin \binom{N}{k-1} \wedge p_1 \notin \binom{N}{k-1}} &= \Sigma_2 \\ \sum_{p_1 \wedge p_2 \notin \binom{N}{k-2}} &= \Sigma_3 \\ \sum_{p_1 \wedge p_2 \notin \binom{N}{k}} &= \Sigma_4 \end{aligned}$$

Sealjuures $\Sigma_1 = \Sigma_2$ ning seega saame avaldist veelgi lihtsustada.

$$\sigma = (p_1 + p_2)\Sigma_1 + p_1 p_2 \Sigma_3 + \Sigma_4$$

Nüüd on näha, et σ avaldises on p_1 ja p_2 vaid väljaspool summasid ning seega saame lihtsalt arvutada

$$\sigma' = (p_1 - e + p_2 + e)\Sigma_1 + (p_1 - e)(p_2 + e)\Sigma_3 + \Sigma_4$$

Nüüd arvutame $\sigma' - \sigma$ ning koondame ja rühmitame sarnased liikmed:

$$\begin{aligned}
\sigma' - \sigma &= ((p_1 - e)(p_2 + e) - p_1 p_2) \Sigma_3 = \\
&= (+p_1 p_2 + p_1 e - p_2 e - e^2 - p_1 p_2) \Sigma_3 = \\
&= (e(p_1 - p_2) - e^2) \Sigma_3 =
\end{aligned}$$

Kuna $e = \frac{p_1 - p_2}{2}$, siis $2e = p_1 - p_2$ ja asendades saame et

$$= (e \cdot (2e) - e^2) \Sigma_3 = e^2 \Sigma_3.$$

See on aga kindlasti positiivne. Seega on

$$\sigma' > \sigma$$

ning seega on “ühtlasema” jaotusega funktsiooni korral tõenäosus suurem, et kõik k funktsiooni väärtust on erinevad.

Seega siis on “ühtlasema” jaotusega funktsioon “turvalisem”.