Palloja uurniin

- Pallot ja uurnat -mallia ajatellen edellä tarkastelimme tuleeko yhteenkään uurnaan kahta palloa
- *Maksimikuormaksi* nimitetään suurinta määrää palloja, jotka osuvat samaan uurnaan
- Tarkast. tilannetta, jossa $m = n$
- Tällöin keskim. kuorma on 1
- Haetaan maksimikuormalle ylärajaa, joka pätee 1:tä lähestyvällä tn.:llä kun n kasvaa
- Seuraava raja on asymptoottisesti tiukka, mutta vakiokerroin 3 ei ole paras mahdollinen

Lemma 5.1: *Kun n palloa sijoitetaan riippumattomasti tasaisen jakauman mukaan n:ään uurnaan satunnaisesti niin tn., että maksimikuorma on yli arvon $3 \ln n / \ln \ln n$ on korkeintaan $1/n$ riittävän suurille n.*

Todistus. Tn., että mv. uurnaan osuu väh. M palloa on kork.

$$\begin{aligned}\binom{n}{M}\left(\frac{1}{n}\right)^M &= \frac{n!}{M!\,(n-M)!\,n^M} \\ &\le \frac{1}{M!} \\ &\le \left(\frac{e}{M}\right)^M,\end{aligned}$$

missä viimeinen epäyhtälö seuraa, koska

$$\frac{k^k}{k!} < \sum_{i=0}^{\infty} \frac{k^i}{i!} = e^k,$$

jonka perusteella $k! > (k/e)^k$.

Tn.:ttä sille että yhteenkin uurnista osuu väh. $M \ge 3 \ln n / \ln \ln n$ palloa rajoittaa ylhäältä yhdisteen tn.:n perusteella

$$\begin{aligned}&n\left(\frac{e}{M}\right)^M \\ &\le n\left(\frac{e \ln \ln n}{3 \ln n}\right)^{3 \ln n / \ln \ln n} \\ &\le n\left(\frac{\ln \ln n}{\ln n}\right)^{3 \ln n / \ln \ln n} \\ &= e^{\ln n}(e^{\ln \ln \ln n - \ln \ln n})^{3 \ln n / \ln \ln n} \\ &= e^{-2 \ln n + 3(\ln n)(\ln \ln \ln n)/ \ln \ln n},\end{aligned}$$

joka on kork. $1/n$ riittävän suurilla n. □

SANKOJÄRJESTÄMINEN (bucket sort)

- Olk. järjestettävänä $n = 2^m$ alkiota, jotka ovat riippumattomia ja tasaisesti jakautuneita joukosta $\{ 0, \ldots, 2^k - 1 \}$, missä $k \ge m$
- Luvut voidaan sankojärjestää $O(n)$ odotusarvoisessa ajassa
- Odotusarvo otetaan syötteen satunnaisen järjestyksen yli, itse algoritmi on deterministinen
- Joidenkin oletusten vallitessa sankojärjestäminen alittaa vertailuihin perustuvan järjestämisen $\Omega(n \log n)$ alarajan
- Ensivaiheessa alkiot jaetaan n:ään sankoon s.e. sankoon j tulevat kaikki alkiot joiden binääriesityksen m ensimmäistä bittiä vastaavat lukua j

- ⋆ Sangot voidaan toteuttaa esim. linkitettyinä listoina
- ⋆ Jos alkio voidaan sijoittaa oikeaan sankoon vakioajassa, niin ensimmäinen askel vaatii lineaarisen ajan
- ⋆ Mihin tahansa sankoon osuvien alkioiden lkm noudattaa binomijakaumaa $B(n, 1/n)$
- ⋆ Toisessa vaiheessa sankojen sisällöt järjestetään millä tahansa normaalilla neliöllisen ajan vaativalla algoritmilla
- ⋆ Kun sankojen järjestetyt alkiot liitetään peräkkäin saadaan alkuperäinen aineisto järjestyksessä
- ⋆ Tulee vielä osoittaa, että toisen vaiheen odotusarvoinen aikavaatimus on $O(n)$
- ⋆ Olk. X_j sankoon j osuvien lukujen lkm

- ⋆ Sangon j alkioiden järjestäminen vaatii kork. ajan cX_j^2, missä c on vakio
- ⋆ Symmetrian perusteella kaikkien sankojen jakauma on sama, joten $\mathbf{E}[X_j^2] = \mathbf{E}[X_1^2]$ kaikilla j
- ⋆ Toisen vaiheen järjestämisten odotusarvoisesti vaatima aika on siis

$$\mathbf{E}\left[\sum_{j=1}^{n} cX_j^2\right] = c\sum_{j=1}^{n} \mathbf{E}[X_j^2] = cn\mathbf{E}[X_1^2]$$

- ⋆ Jos $X \sim B(n,p)$, niin $\mathbf{E}[X^2] = n(n-1)p^2 + np$ (kirjan sivu 48)

$$\mathbf{E}[X_1^2] = \frac{n(n-1)}{n^2} + 1 = 2 - \frac{1}{n} < 2$$

- ⋆ Toisen vaiheen odotusarvoinen aika on siis kork. $2cn$
- ⋆ Kaikkiaan sankojärjestäminen toimii odotusarvoisesti lineaarisessa ajassa

Poisson-jakauma

- ⋆ Tarkastellan tn.:ttä että jokin uurna jää kokonaan palloitta ja tyhjien uurnien odotusarvoa
- ⋆ Tn. että tiettyyn uurnaan ei osu yhtään palloa on

$$\left(1 - \frac{1}{n}\right)^m \approx e^{-m/n}$$

- ⋆ Koska kaikilla uurnilla on sama tn. jäädä tyhjäksi, niin lineaarisuuden perusteella tyhjien uurnien lkm:n X odotusarvo on

$$\mathbf{E}[X] = n\left(1 - \frac{1}{n}\right)^m \approx ne^{-m/n}$$

- ⋆ Tämä arvio on melko hyvä jopa pienillä arvoilla m ja n

- ⋆ Yleistäen: tn., että uurnaan j osuu $X_j = r$ palloa on

$$\binom{m}{r}\left(\frac{1}{n}\right)^r\left(1 - \frac{1}{n}\right)^{m-r} = \frac{1}{r!}\frac{m(m-1)\cdots(m-r+1)}{n^r}\left(1 - \frac{1}{n}\right)^{m-r}$$

- ⋆ Kun $m, n \gg r$, niin tämä tn. on likimäärin

$$\Pr(X_j = r) \approx \frac{e^{-m/n}(m/n)^r}{r!}$$

- ⋆ Palloja per uurna on keskimäärin m/n
- ⋆ *Diskreetti Poisson-jakautunut* s-muuttuja X parametrilla μ, $X \sim \mathrm{Poisson}(\mu)$, noudattaa seuraavaa jakaumaa kaikilla $j = 0, 1, 2, \ldots$

$$\Pr(X = j) = \frac{e^{-\mu}\mu^j}{j!}$$

⋆ Yo. määritelmä antaa todella jakauman, s.o., todennäköisyydet summautuvat arvoon 1

$$\begin{aligned}\sum_{j=0}^{\infty} \Pr(X=j) &= \sum_{j=0}^{\infty} \frac{e^{-\mu}\mu^j}{j!} \\ &= e^{-\mu}\sum_{j=0}^{\infty}\frac{\mu^j}{j!} = 1,\end{aligned}$$

koska Taylor-kehitelmän perusteella $e^x = \sum_{j=0}^{\infty}(x^j/(j!))$

⋆ S-muuttujan X odotusarvo $\mathbf{E}[X]$ on

$$\begin{aligned}&= \sum_{j=0}^{\infty} j\Pr(X=j) = \sum_{j=1}^{\infty} j\frac{e^{-\mu}\mu^j}{j!} \\ &= \mu\sum_{j=1}^{\infty}\frac{e^{-\mu}\mu^{j-1}}{(j-1)!} = \mu\sum_{j=0}^{\infty}\frac{e^{-\mu}\mu^j}{j!} = \mu\end{aligned}$$

Lemma 5.2: *Äärellisen monen riippumattoman Poisson-satunnaismuuttujan summa on Poisson-satunnaismuuttuja.*

Todistus. Olkoot $X \sim \text{Poisson}(\mu_1)$ ja $Y \sim \text{Poisson}(\mu_2)$ toisistaan riippumattomia s-muuttujia. Nyt

$\Pr(X+Y=j)$

$$\begin{aligned}&= \sum_{k=0}^{j}\Pr((X=k)\cap(Y=j-k)) \\ &= \sum_{k=0}^{j}\frac{e^{-\mu_1}\mu_1^k}{k!}\frac{e^{-\mu_2}\mu_2^{j-k}}{(j-k)!} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{j!}\sum_{k=0}^{j}\frac{j!}{k!(k-j)!}\mu_1^k\mu_2^{j-k} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{j!}\sum_{k=0}^{j}\binom{j}{k}\mu_1^k\mu_2^{j-k} \\ &= \frac{e^{-(\mu_1+\mu_2)}(\mu_1+\mu_2)^j}{j!},\end{aligned}$$

missä viimeinen yhtälö seuraa binomilauseesta

$$(x+y)^r = \sum_k \binom{r}{k} x^k y^{r-k}.$$

Todistus saatetaan loppuun induktiolla. □

Lemma 5.3: *Olk. $X \sim \text{Poisson}(\mu)$. X:n mgf on $M_X(t) = \exp(\mu(e^t-1))$.*

Todistus. Kaikilla t pätee

$$\begin{aligned}\mathbf{E}\left[e^{tX}\right] &= \sum_{k=0}^{\infty}\frac{e^{-\mu}\mu^k}{k!}e^{tk} \\ &= e^{\mu(e^t-1)}\sum_{k=0}^{\infty}\frac{e^{-\mu e^t}(\mu e^t)^k}{k!} \\ &= e^{\mu(e^t-1)}.\end{aligned}$$

□

⋆ Derivoimalla edellinen saadaan

$$\begin{aligned}M_X'(t) &= e^{\mu(e^t-1)}\cdot\mu e^t \\ M_X''(t) &= \mu e^{\mu(e^t-1)+t}(\mu e^t+1)\end{aligned}$$

⋆ Näin ollen, sijoittamalla $t=0$, saadaan $\mathbf{E}[X] = \mu$ sekä $\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = (\mu^2+\mu) - \mu^2 = \mu$

⋆ Lemma 5.2 olisi voitu todistaa myös lemman 5.3 avulla

⋆ Lauseen 4.3 perusteella

$$\begin{aligned}M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= \exp\left((\mu_1+\mu_2)(e^t-1)\right),\end{aligned}$$

⋆ Tämä on $\text{Poisson}(\mu_1+\mu_2)$-jakautuneen s-muuttujan mgf

⋆ Koska mgf lauseen 4.2 perusteella määrää 1-käsitteisesti jakauman, niin $X+Y$ on Poisson-satunnaismuuttuja, jonka parametri on $\mu_1+\mu_2$

Lause 5.4: *Olk.* $X \sim \mathrm{Poisson}(\mu)$. *Tällöin*

1. *Jos* $x > \mu$, *niin* $\Pr(X \geq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$;
2. *Jos* $x < \mu$, *niin* $\Pr(X \leq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$.

Todistus. Kaikilla $t > 0$ ja $x > \mu$

$$\begin{aligned}\Pr(X \geq x) &= \Pr(e^{tX} \geq e^{tx}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{tx}} \\ &= \exp\left(\mu(e^t - 1) - tx\right).\end{aligned}$$

Sijoittamalla $t = \ln(x/\mu) > 0$ saadaan

$$\begin{aligned}\Pr(X \geq x) &\leq \exp\left(x - \mu - x\ln(x/\mu)\right) \\ &= \frac{e^{-\mu}(e\mu)^x}{x^x}.\end{aligned}$$

Toinen kohta todistetaan vastaavasti. □

Poisson-approksimaatio

- Pallot ja uurnat -mallissa tulee väkisinkin riippuvuuksia
  - Jos uurna 1 on tyhjä, niin muiden uurnien tyhjyyden tn. on pienempi sillä m palloa on jakautunut $n-1$ uurnaan
  - Jos tiedämme $n-1$ uurnan pallojen lkm:n, niin viimeisen uurnan pallojen lkm on määrätty
- Uurnien kuormat eivät ole toisistaan riippumattomia
- Tiedämme, että yhden uurnan kuormaa approksimoi jakauma $\mathrm{Poisson}(m/n)$
- Tavoitteemme on osoittaa, että kaikkien uurnien yhteisjakauma approksimoituu hyvin kun oletamme kunkin uurnan kuorman olevan riippumattomasti $\mathrm{Poisson}(m/n)$-jakautunut

Lause 5.5: *Olk.* $X_n \sim B(n, p)$, *missä* p *riippuu* n*:stä ja* $\lim_{n\to\infty} np = \lambda$ *on* n*:stä riippumaton vakio. Tällöin mille tahansa kiinteälle* k *pätee*

$$\lim_{n\to\infty} \Pr(X_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

- Olk. m pallojen ja b uurnien lkm s.e. m on b:stä riippuva ja $\lim_{n\to\infty} m/b = \lambda$
- Olk. X_n tiettyyn uurnaan osuvien pallojen lkm
- $X_n \sim B(m, 1/b)$
- Lauseen 5.5 perusteella saamme

$$\lim_{n\to\infty} \Pr(X_n = r) = \frac{e^{-m/n}(m/n)^r}{r!}$$

- Tämä on sama kuin aiemmin (kalvo 126) laskemamme arvio

Eksakti tapaus: Olk. $X_i^{(m)}$, $1 \leq i \leq n$, uurnan i kuorma, kun m palloa sijoitetaan riippumattomasti ja tasaisen jakauman mukaan satunnaisesti n:ään uurnaan

Poisson-tapaus: Olk. $Y_1^{(m)}, \ldots, Y_n^{(m)}$ n riippumatonta s-muuttujaa s.e. $Y_i^{(m)} \sim \mathrm{Poisson}(m/n)$

- Eksaktissa tapauksessa palloja on m kpl, kun taas Poisson-tapauksessa niiden kokonaismäärän odotusarvo on m
- Jos palloja kuitenkin Poisson-tapauksessa on kaikkiaan m, niin
  - muuttujien jakauma on sama kuin jos m palloa olisi sijoitettu satunnaisesti n:ään uurnaan

Lause 5.6: *$(Y_1^{(m)}, \ldots, Y_n^{(m)})$:n jakauma ehdolla $\sum_i Y_i^{(m)} = k$ on $(X_1^{(k)}, \ldots, X_n^{(k)})$ riippumatta arvosta m.*

Todistus. Olk. $(k_1, \ldots, k_n)$ mielivaltainen s.e. $\sum_i k_i = k$. Tn., että $(X_1^{(k)}, \ldots, X_n^{(k)}) = (k_1, \ldots, k_n)$ on

$$\binom{k}{k_1; k_2; \ldots; k_n} \Big/ n^k = \frac{k!}{(k_1!) \cdots (k_n!) n^k}.$$

Tarkastellaan nyt tn.:ttä

$$\Pr\left((Y_1^{(m)}, \ldots, Y_n^{(m)}) = (k_1, \ldots, k_n) \,\middle|\, \sum_{i=1}^n Y_i^{(m)} = k \right) = \frac{\Pr\left((Y_1^{(m)} = k_1) \cap \cdots \cap (Y_n^{(m)} = k_n) \right)}{\Pr\left(\sum_i^n Y_i^{(m)} = k \right)}.$$

$$\Pr\left(Y_i^{(m)} = k_i \right) = e^{-m/n} (m/n)^{k_i} / k_i!,$$

koska $Y_i^{(m)} \sim \text{Poisson}(m/n)$ ja riippumattomia. Lemman 5.2 perusteella myös niiden summa on Poisson-jakautunut parametrilla m. Täten yo. ehdollinen tn. on

$$\frac{\prod_{i=1}^n e^{-m/n} (m/n)^{k_i} / k_i!}{e^{-m} m^k / k!} = \frac{k!}{(k_1!) \cdots (k_n!) n^k},$$

joka todistaa väitteen. □

- ⋆ Seuraavaa (Stirlingin approksimaatiota löysempää) kertoman arviota käytetään lauseen 5.7 todistamisessa

Lemma 5.8: $n! \leq e\sqrt{n} \left(\frac{n}{e} \right)^n$.

Lause 5.7: *Olk. $f(x_1, \ldots, x_n)$ ei-negatiivinen funktio. Tällöin*

$$\mathbf{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)}) \right] \leq e\sqrt{m} \mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \right].$$

Todistus. Lemman 2.5 perusteella

$$\begin{aligned}
&\mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \right] \\
&= \sum_{k=0}^{\infty} \mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \,\middle|\, \sum_{i=1}^n Y_i^{(m)} = k \right] \Pr\left(\sum_{i=1}^n Y_i^{(m)} = k \right) \\
&\geq \mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \,\middle|\, \sum_{i=1}^n Y_i^{(m)} = m \right] \Pr\left(\sum_{i=1}^n Y_i^{(m)} = m \right) \\
&= \mathbf{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)}) \right] \Pr\left(\sum_{i=1}^n Y_i^{(m)} = m \right),
\end{aligned}$$

missä viimeinen yhtälö seuraa lauseen 5.6 perusteella. Poisson-jakautuneiden $Y_i^{(m)}$ summa on Poisson-jakautunut parametrilla m, joten

$$\begin{aligned}
&\mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \right] \\
&\geq \mathbf{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)}) \right] \frac{e^{-m} m^m}{m!} \\
&\geq \mathbf{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)}) \right] \frac{1}{e\sqrt{m}},
\end{aligned}$$

lemman 5.8 perusteella. □

- ⋆ Ed. lause pätee mille tahansa ei-negatiiviselle funktiolle yli pallojen määrän uurnissa
- ⋆ Erityisesti se pätee indikaattorifunktiolle, jonka arvo on 1 jonkin tapahtuman yhteydessä, ja antaa rajoja tapahtumien tn.:lle

Korollaari 5.9: *Jos Poisson-tapauksessa tapahtuman tn. on p, niin sen tn. eksaktissa tapauksessa on kork. $pe\sqrt{m}$.*

Todistus. Olk. f tapahtuman indikaattorifunktio. Tällöin $\mathbf{E}[f]$ on täsmälleen sama kuin tn., että tapahtuma havaitaan ja väite seuraa suoraan lauseesta 5.7. □

- ⋆ Jos siis jotakin tapahtuu pienellä tn.:llä Poisson-tapauksessa, niin sen tapahtumistn. on pieni myös eksaktissa mallissa
- ⋆ Voimme siis käyttää melko yksinkertaista riippumattomien Poisson-jakautuneiden muuttujien mallia johtamaan rajoja pallot ja uurnat -mallille

Lause 5.10: *Olk. $f(x_1, \ldots, x_n)$ ei-negatiivinen funktio s.e. $\mathbf{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)})\right]$ on joko monotonisesti kasvava tai monotonisesti laskeva. Tällöin*

$$\begin{aligned} &\mathbf{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)})\right] \\ &\leq \quad 2\mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)})\right]. \end{aligned}$$

Korollaari 5.11: *Olk. $\mathcal{E}$ tapahtuma, jonka tn. on pallojen lkm:n myötä joko monotonisesta kasvava tai laskeva. Jos Poisson-tapauksessa $\Pr(\mathcal{E}) = p$, niin eksaktissa tapauksessa $\Pr(\mathcal{E}) \leq 2p$.*

- ⋆ Lemmassa 5.1 laskimme yhdisteen tn.:n perusteella, että uurnien maksimikuorma on kork. $3 \ln n / \ln \ln n$ suurella tn.:llä kun $m = n$
- ⋆ Seuraava lemma antaa lähes saman alarajan maksimikuormalle

Lemma 5.12: *Kun n palloa sijoitetaan riippumattomasti tasaisen jakauman mukaan n:ään uurnaan satunnaisesti, on maksimikuorma* vähintään $\ln n / \ln \ln n$ *vähintään tn.:llä $1 - 1/n$ riittävän suurilla n.*

Todistus. Tn., että Poisson-tapauksessa tietyn uurnan kuorma on $M = \ln n / \ln \ln n$ on $1/eM!$. Koska uurnat ovat riippumattomia, niin tn., ettei yhdenkään uurnan kuorma ole väh. M on kork.

$$\left(1 - \frac{1}{eM!}\right)^n \leq \exp\left(-n/(eM!)\right).$$

Valitsemalla M nyt s.e. $e^{-n/(eM!)} \leq n^{-2}$, lauseen 5.7 perusteella eksaktin tapauksen tn., ettei yhdenkään uurnan kuorma ole väh. M on kork. $e\sqrt{n}/n^2 < 1/n$, joka todistaa väitteen.

Toisaalta maksimikuorma on selvästi mo-

notonisesti kasvava pallojen lkm:n myötä, joten voisimme myös käyttää lemmaa 5.10.

Riittää osoittaa, että $M! \leq n/2e \ln n$ eli $\ln M! \leq \ln n - \ln \ln n - \ln(2e)$. Lemman 5.8 perusteella

$$M! \leq e\sqrt{M}\left(\frac{M}{e}\right)^M \leq M\left(\frac{M}{e}\right)^M,$$

kun n (ja M) on riittävän suuri. Tällöin

$$\begin{aligned} \ln M! &\leq M \ln M - M + \ln M \\ &= \frac{\ln n}{\ln \ln n}(\ln \ln n - \ln \ln \ln n) \\ &\quad - \frac{\ln n}{\ln \ln n} + (\ln \ln n - \ln \ln \ln n) \\ &\leq \ln n - \frac{\ln n}{\ln \ln n} \\ &\leq \ln n - \ln \ln n - \ln(2e), \end{aligned}$$

missä kaksi viimeistä epäyhtälöä perustuvat faktaan $\ln \ln n = o(\ln n / \ln \ln n)$. □

Hajautus

- Hajautus sijoittaa universumin U alkiot (esim. salasanat) n:ään lokeroon hajautusftiolla $f : U \to [0, n-1]$
- Tekemällä oletus f:n satunnaisuudesta, voidaan hajautusta mallintaa pallot ja uurnat -mallilla
  - Kaikilla $x \in U$: $\Pr(f(x) = j) = 1/n$
  - Kaikilla x arvot $f(x)$ ovat toisistaan riippumattomia
- Olk. avaimia on m ja lokeroita n kpl
- Avainten odotuarvo lokerossa on m/n
- Muiden kuin etsitty avain odotusarvo lokerossa on $(m-1)/n$, joten kaikkiaan siinä on odotusarvoisesti $1 + (m-1)/n$ avainta
- Valitsemalla $n = m$ lokerossa läpikäytyjen avainten odotusarvo on vakio

- Hitainta haku on lokerossa, joka sisältää suurimman määrän avaimia
- Pallot ja uurnat -mallin analyysin perusteella tiedämme, että maksimikuorma on $\Theta(\ln n / \ln \ln n)$ tn.:llä joka on lähellä yhtä
- Suuri ero haun pahimman ja odotusarvoisen aikavaativuuden välillä voi olla joissain sovelluksissa merkittävä ongelma
- Toinen ketjuttavan hajautuksen merkittävä ongelma on tilan tuhlaus
- Monet tyhjät lokerot syövät tilaa
- Tilaa ja haun tehokkuutta voidaan tasapainotella säätämällä parametria m/n