

Chernoffin rajojen johtaminen

★ Chernoffin raja s-muuttujalle X saadaan soveltamalla Markovin epäyhtälöä e^{tX} :n hyvin valitulla arvolla t

★ Mille tahansa $t > 0$ pätee

$$\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$

★ Erityisesti

$$\Pr(X \geq a) \leq \min_{t>0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}$$

★ Vastaavasti pätee

$$\Pr(X \leq a) \leq \min_{t>0} (\mathbf{E}[e^{tX}] / e^{ta})$$

★ Idean soveltamiseksi jakaumaan tarvitaan mgf:n arvio ja arvon t valinta

★ Optimaalisen arvon t sijaan usein tavoitellaan "sopivan" rajan antavaa arvoa

94

POISSON-KOKEIDEN SUMMA

★ Bernoulli-toistokokeen (eli binomijakautuneen s-muuttujan) yleistys on Poisson-toistokoe

★ Nyt kaikilla indikaattorimuuttujilla ei tarvitse olla samaa jakaumaa

★ Seuraavat tulokset pätevät tietysti myös binomijakautuneelle s-muuttujalle

★ Olk. X_1, \dots, X_n jono riippumattomia Poisson-kokeita s.e. $\Pr(X_i = 1) = p_i$ ja $X = \sum_{i=1}^n X_i$ sekä

$$\begin{aligned} \mu &= \mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^n X_i\right] \\ &= \sum_{i=1}^n \mathbf{E}[X_i] = \sum_{i=1}^n p_i \end{aligned}$$

95

★ Chernoffin rajojen johtamiseksi tarvitsemme mgf:n

$$\begin{aligned} M_{X_i}(t) &= \mathbf{E}[e^{tX_i}] \\ &= p_i e^t + (1 - p_i) \\ &= 1 + p_i(e^t - 1) \\ &\leq e^{p_i(e^t - 1)} \end{aligned}$$

★ Lauseen 4.3 (yleistyksen) perusteella

$$\begin{aligned} M_X(t) &= \prod_{i=1}^n M_{X_i}(t) \\ &\leq \prod_{i=1}^n e^{p_i(e^t - 1)} \\ &= \exp\left(\sum_{i=1}^n p_i(e^t - 1)\right) \\ &= e^{(e^t - 1)\mu} \end{aligned}$$

96

★ Tämän Poisson-toistokokeen mgf:n arvon perusteella saamme konkreettisia Chernoffin rajoja

Lause 4.4: Olk. Poisson-kokeiden X_1, \dots, X_n , $\Pr(X_i = 1) = p_i$, summa $X = \sum_{i=1}^n X_i$ ja $\mu = \mathbf{E}[X]$. Tällöin pätevät seuraavat Chernoffin rajat:

1. mille tahansa $\delta > 0$

$$\Pr(X \geq (1 + \delta)\mu) < \left(\frac{e^\delta}{(1 + \delta)^{1 + \delta}}\right)^\mu;$$

2. arvoilla $0 < \delta \leq 1$

$$\Pr(X \geq (1 + \delta)\mu) \leq \exp(-\mu\delta^2/3);$$

3. arvoilla $R \geq 6\mu$

$$\Pr(X \geq R) \leq 2^{-R}.$$

★ Kohta 1 antaa tiukimman rajan, mutta kohtien 2 ja 3 rajat ovat helpommin ilmaistavia ja laskettavia

97

Todistus. Kohdan 1 todistamiseksi sovelletaan Markovin epäyhtälöä:

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

Koska $\delta > 0$, niin voimme sijoittaa $t = \ln(1 + \delta) > 0$, jolloin saadaan ensimmäinen väite.

Kohdan 2 todistamiseksi riittää osoittaa, että

$$\frac{e^\delta}{(1 + \delta)^{1+\delta}} \leq e^{-\delta^2/3}.$$

Ottamalla logaritmi puolittain saadaan ekvivalentti ehto

$$f(\delta) = \delta - (1 + \delta) \ln(1 + \delta) + \delta^2/3 \leq 0.$$

98

Derivoimalla saadaan

$$\begin{aligned} f'(\delta) &= 1 - \frac{1 + \delta}{1 + \delta} - \ln(1 + \delta) + 2\delta/3 \\ &= -\ln(1 + \delta) + 2\delta/3 \\ f''(\delta) &= -1/(1 + \delta) + 2/3. \end{aligned}$$

$f''(\delta) < 0$ välillä $0 \leq \delta \leq 1/2$, eli $f'(\delta)$ vähenee. Kun $1/2 < \delta < 1$ on puolestaan $f''(\delta) > 0$ ja $f'(\delta)$ kasvaa.

Koska $f'(0) = 0$ ja $f'(1) = 2/3 - \ln 2 < 0$, niin $f'(\delta) \leq 0$ välillä $0 \leq \delta \leq 1$.

Koska edelleen $f(0) = 0$, niin edellisen perusteella $f(\delta) \leq 0$ välillä $0 < \delta \leq 1$, joka todistaa kohdan 2.

Kohdan 3 todistamiseksi olk. $R = (1 + \delta)\mu$. Tällöin arvoilla $R \geq 6\mu$ pätee

$$\delta = R/\mu - 1 \geq 5.$$

99

Kohdan 1 perusteella $\Pr(X \geq R) =$

$$\begin{aligned} \Pr(X \geq (1 + \delta)\mu) &\leq \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^\mu \\ &\leq \left(\frac{e}{1 + \delta} \right)^{(1+\delta)\mu} \\ &\leq \left(\frac{e}{6} \right)^R \leq \frac{1}{2^R}. \end{aligned}$$

□

★ Häntätodennäköisyyttä odotusarvon alapuolelta rajoitettaessa pätee

Lause 4.5: Olk. Poisson-kokeiden X_1, \dots, X_n , $\Pr(X_i = 1) = p_i$, summa $X = \sum_{i=1}^n X_i$ ja $\mu = \mathbf{E}[X]$. Tällöin arvoilla $0 < \delta < 1$ pätevät

1. $\Pr(X \leq (1 - \delta)\mu) \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right)^\mu$;
2. $\Pr(X \leq (1 - \delta)\mu) \leq \exp(-\mu\delta^2/2)$.

100

★ Tässä taas kohta 1 antaa tiukemman rajan, mutta kohdan 2 arvio on yleensä helpompi käyttää ja riittävä

★ Yhdistämällä lauseiden 4.4 ja 4.5 kohdat 2 saadaan usein käytetty Chernoffin rajojen muoto

Korollari 4.6: Olk. Poisson-kokeiden X_1, \dots, X_n , $\Pr(X_i = 1) = p_i$, summa $X = \sum_{i=1}^n X_i$ ja $\mu = \mathbf{E}[X]$. Tällöin arvoilla $0 < \delta < 1$ pätee

$$\Pr(|X - \mu| \geq \delta\mu) \leq 2 \exp(-\mu\delta^2/3).$$

★ Odotusarvoa $\mathbf{E}[X]$ ei tarvitse tietää eksaktisti, vaan

- ◇ lauseessa 4.4 voidaan käyttää ylärajaa $\mu \geq \mathbf{E}[X]$ ja
- ◇ lauseessa 4.5 alarajaa $\mu \leq \mathbf{E}[X]$

101

KOLIKONHEITTO (TAAS)

- ★ Olk. X kruunien lkm n :ssä painottamattoman kolikon heitossa
- ★ Korollarin 4.6 perusteella, kun $\delta = \sqrt{(6 \ln n)/n}$,

$$\Pr(|X - n/2| \geq \sqrt{6n \ln n/2}) \leq 2 \exp\left(-\frac{1}{3} \frac{n}{2} \frac{6 \ln n}{n}\right) = \frac{2}{n}$$

- ★ Lkm siis keskittyy (**concentrates**) vahvasti odotusarvonsa ympäristöön
- ★ Tšebyševin epäyhtälöllä saimme aiemmin $\Pr(|X - n/2| \geq n/4) \leq 4/n$
- ★ Chernoffin rajalla saamme eksponentiaalisesti pienemmän poikkeamatn:n

$$\Pr\left(|X - \frac{n}{2}| \geq \frac{n}{4}\right) \leq 2 \exp\left(-\frac{1}{3} \frac{n}{2} \frac{1}{4}\right) = 2e^{-n/24}$$

102

PARAMETRIN ESTIMOINTI

- ★ Haluamme esim. arvioida tn.:ttä, että tietty geenimutaatio tapahtuu
- ★ Laboratoriotesteissä selviää DNA näytteestä onko siinä k.o. mutaatio
- ★ Testit ovat kuitenkin kalliita ja haluaisimme luotettavan arvion pienellä näytteiden määrällä
- ★ Olk. p arvo, jota estimoimme
- ★ Olk. näytteiden lkm n ja $X = \tilde{p}n$ mutaatioiden havaintofrekvenssi
- ★ Jos n on riittävän suuri, niin odotamme \tilde{p} :n olevan lähellä p :tä
- ★ Väli $[\tilde{p} - \delta, \tilde{p} + \delta]$ on $(1 - \gamma)$ -luottamusväli parametrille p , jos

$$\Pr(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) \geq 1 - \gamma$$

103

- ★ $X = \tilde{p}n$ on binomijakautunut parametrein n ja p , joten $\mathbf{E}[X] = np$
- ★ Jos $p \notin [\tilde{p} - \delta, \tilde{p} + \delta]$, niin toinen seuraavista on tapahtunut:
 - ◇ $p < \tilde{p} - \delta$:
 $X = n\tilde{p} > n(p + \delta) = \mathbf{E}[X](1 + \delta/p)$
 - ◇ $p > \tilde{p} + \delta$: $X < \mathbf{E}[X](1 - \delta/p)$
- ★ Laskemalla tn.:t yhteen ja soveltamalla lauseita 4.4 ja 4.5 saamme ylärajan

$$\begin{aligned} \Pr(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) &< e^{-np(\delta/p)^2/2} + e^{-np(\delta/p)^2/3} \\ &= e^{-n\delta^2/(2p)} + e^{-n\delta^2/(3p)} \end{aligned}$$

- ★ Koska p :n arvoa ei tunneta, käytetään ylärajaa $p \leq 1$, jolloin voimme valita

$$\gamma = e^{-n\delta/2} + e^{-n\delta/3}$$

- ★ Kääntäen voimme ratkaista tästä δ :n, kun γ on kiinnitetty ja n tunnetaan

104

Tarkempia rajoja joillekin erikoistapauksille

- ★ Edellisiä tiukempia rajoja saadaan symmetrisesti jakautuneille s -muuttujille

Lause 4.7: Olk. $X = \sum_{i=1}^n X_i$, missä X_1, \dots, X_n ovat riippumattomia s -muuttujia s.e. $\Pr(X_i = 1) = \Pr(X_i = -1) = 1/2$. Tällöin kaikilla $a > 0$ pätee

$$\Pr(X \geq a) \leq \exp\left(-\frac{a^2}{2n}\right).$$

Todistus. Odotusarvon määritelmän perusteella mille tahansa $t > 0$ pätee

$$\mathbf{E}[e^{tX_i}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t}.$$

Huomioimalla, että e^t :n Taylor-kehityksen

105

perusteella $e^t = \sum_{i=0}^{\infty} \frac{t^i}{i!}$, saadaan

$$\begin{aligned} \mathbf{E}[e^{tX_i}] &= \frac{1}{2} \left(1 + t + \frac{t^2}{2} + \frac{t^3}{3!} + \dots \right) \\ &\quad + \frac{1}{2} \left(1 - t + \frac{t^2}{2} - \frac{t^3}{3!} + \dots \right) \\ &= \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!} \\ &\leq \sum_{i=0}^{\infty} \frac{(t^2/2)^i}{i!} \\ &= \exp\left(\frac{t^2}{2}\right). \end{aligned}$$

Tämän estimaatin perusteella

$$\mathbf{E}[e^{tX}] = \prod_{i=1}^n \mathbf{E}[e^{tX_i}] \leq \exp\left(\frac{t^2 n}{2}\right)$$

106

ja Markovin epäyhtälön perusteella

$$\begin{aligned} \Pr(X \geq a) &= \Pr(e^{tX} \geq e^{ta}) \\ &\leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}} \\ &= \exp\left(\frac{t^2 n}{2} - ta\right). \end{aligned}$$

Lopulta sijoillamalla $t = a/n$ saadaan väite $\Pr(X \geq a) \leq \exp(-a^2/2n)$. \square

★ Symmetrian perusteella myös

$$\Pr(X \leq -a) \leq \exp(-a^2/2n)$$

Korollaari 4.8: Olk. $X = \sum_{i=1}^n X_i$, missä X_1, \dots, X_n ovat riippumattomia s -muuttujia s.e. $\Pr(X_i = 1) = \Pr(X_i = -1) = 1/2$.

Tällöin kaikilla $a > 0$ pätee

$$\Pr(|X| \geq a) \leq 2 \exp\left(-\frac{a^2}{2n}\right).$$

107

Korollaari 4.9: Olk. $Y = \sum_{i=1}^n Y_i$, missä Y_1, \dots, Y_n ovat riippumattomia s -muuttujia s.e. $\Pr(Y_i = 1) = \Pr(Y_i = 0) = 1/2$. Merk. $\mu = \mathbf{E}[Y] = n/2$. Tällöin

1. kaikilla $a > 0$ pätee

$$\Pr(Y \geq \mu + a) \leq \exp\left(-\frac{2a^2}{n}\right);$$

2. kaikilla $\delta > 0$ pätee

$$\Pr(Y \geq (1 + \delta)\mu) \leq \exp(-\delta^2\mu).$$

Todistus. Olk. X_i kuten edellä ja $Y_i = (X_i + 1)/2$, jolloin

$$\begin{aligned} Y &= \sum_{i=1}^n Y_i = \frac{1}{2} \left(\sum_{i=1}^n X_i \right) + \frac{n}{2} \\ &= \frac{1}{2} X + \mu. \end{aligned}$$

108

Lauseen 4.7 perusteella

$$\begin{aligned} \Pr(Y \geq \mu + a) &= \Pr(X \geq 2a) \\ &\leq \exp\left(-\frac{4a^2}{2n}\right), \end{aligned}$$

joka todistaa väittämän kohdan 1. Kohdan 2 todistamiseksi valitaan $a = \delta\mu = \delta n/2$. Taas lauseen 4.7 perusteella

$$\begin{aligned} \Pr(Y \geq (1 + \delta)\mu) &= \Pr(X \geq 2\delta\mu) \\ &\leq \exp\left(-\frac{4\delta^2\mu^2}{2n}\right) \\ &= \exp(-\delta^2\mu). \end{aligned}$$

\square

★ Vastaava tulos pätee kun rajoitetaan tn:iä odotusarvon alapuolelta

109

- ★ Olk. annettuna m henkilöä ja n ominaisuutta
- ★ Tilastollista koetta varten tavoittelemme henkilöiden jakoa kahteen joukkoon A ja \bar{A} s.e.
 - ◇ kaikkien ominaisuuksien $i = 1, \dots, n$ suhteen joukot ovat niin tasapainoisia kuin mahdollista
 - ◇ ts., ominaisuuden i omaavien henkilöiden lkm kummassakin joukossa on jokseenkin sama
- ★ Toisen joukon henkilöt alistetaan testiin ja toinen toimii kontrolliryhmänä
- ★ Määritellään $n \times m$ matriisi $\mathbf{A} = (a_{ij}) \in \{0, 1\}^{n \times m}$, jossa sarakkeet vastaavat henkilöitä ja rivit ominaisuuksia
 - ◇ $a_{ij} = 1$ jos h:löllä j on ominaisuus i

Korollaari 4.11: Satunnaisvektorille \vec{b} , jonka komponentit on vedetty riippumattomasti ja tasaisen jakauman mukaan joukosta $\{-1, 1\}$ pätee $\Pr(\|\mathbf{A}\vec{b}\|_\infty \geq \sqrt{4m \ln n}) \leq 2/n$.

Todistus. Tarkastellaan \mathbf{A} :n i :ttä riviä a_{i1}, \dots, a_{im} . Olk. $k = \sum_{j=1}^m a_{ij}$ ykkösten lkm tuolla rivillä. Jos $k \leq \sqrt{4m \ln n}$, niin selvästi $|\vec{a}_i \cdot \vec{b}| = |c_i| \leq \sqrt{4m \ln n}$.

Jos taas $k > \sqrt{4m \ln n}$, niin c_i :n k termiä $a_{ij}b_j \neq 0$ ovat riippumattomia s-muuttujia, joilla on tn. $1/2$ saada arvo -1 ja 1 .

Korollaarin 4.8 perusteella, koska $m \geq k$,

$$\Pr(|c_i| \geq \sqrt{4m \ln n}) \leq \exp\left(-\frac{4m \ln n}{2k}\right) \leq \frac{2}{n^2}.$$

Tn., että yksikin n :stä rivistä ylittää rajan on yhdisteen tn.:n perusteella $2/n$. □

- ★ Henkilöiden ositus (A, \bar{A}) esitetään vektorina $\vec{b} \in \{-1, 1\}^m$, missä $b_j = 1$ jos h:lö j on joukossa A
- ★ Olk. $\mathbf{A}\vec{b} = \vec{c} \in \mathbb{Z}^m$, eli $c_i = \sum_j a_{ij}b_j$
- ★ Tavoitteemme on minimoida

$$\|\mathbf{A}\vec{b}\|_\infty = \max_i |c_i|$$

- ★ Äärimmäisen yksinkert. satunnaisalgoritmi valitsee kunkin b_j riippumattomasti s.e. $\Pr(b_j = 1) = 1/2$
- ★ Vaikka koko \mathbf{A} jätetään täysin huomiotta, saavutetaan tällä menetelmällä niinkin tiukka raja kuin $O(\sqrt{m \ln n})$ $\|\mathbf{A}\vec{b}\|_\infty$:lle suurella tn.:llä

5. Pallot, urnat ja satunnaisverkot

- ★ Tarkastellaan m :n pallon sijoittamista toisistaan riippumatta n :n urnaan, kun kullakin urnalla on sama tn. tulla valituksi
- ★ Klassinen tn-laskennan malli, jolla on paljon sovelluksia modernissa algoritmianalyysissä
- ★ Kysymykset:
 - ◇ Kuinka monta palloa johonkin urnaan sattuu?
 - ◇ Mikä on suurin määrä palloja yhdessä urnassa?
- ★ Sovellamme mallia mm. tietoraken-teisiin (hajautukseen)

Syntymäpäiväparadoksi

- ★ Millä tn.:llä esim. 30:n ihmisen joukossa ainakin kahdella on sama syntymäpäivä?
- ★ Nyt siis $m = 30$ ja $n = 365$
- ★ Luonnollisestikin tämä on yksinkertaistus reaali maailman tilanteesta
- ★ Valitaan 30 eri syntymäpäivää 365:n päivän joukosta kiinnittämättä huomiota siihen mikä päivä osuu kellekin
- ★ Mahdollisuuksia, että kaikilla 30:lla on eri syntymäpäivä on

$$\binom{365}{30} 30!$$

- ★ Mahdollisuuksia valita kaikille syntymäpäivä on kaikkiaan 365^{30}

114

- ★ Henkilö kerrallaan laskien
 - ★ Kun ensimmäinen syntymäpäivä on valittu on tn., että toisen henkilön syntymäpäivä on muu, on $1 - 1/365$
 - ★ Yleistäen kun $k - 1$ syntymäpäivää on valittu, niin tn., ettei k :s syntymäpäivä mene näiden kanssa päällekkäin on $1 - (k - 1)/365$
 - ★ 30:n hengen tapauksessa tn. saada 30 eri syntymäpäivää on siis
- $$\left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{29}{365}\right) \approx 0,2937$$
- ★ Siis yli 70% tn., että ainakin kahdella 30:sta hengestä on sama syntymäpäivä
 - ★ Yleisemmin jos palloja on m ja urnia n , niin tn., ettei yhteenkään urnaan osu kahta palloa on $\prod_{j=1}^{m-1} \left(1 - \frac{j}{n}\right)$

115

★ Kun $k \ll n$, niin $1 - k/n \approx e^{-k/n}$

★ Jos siis $m \ll n$, niin

$$\begin{aligned} \prod_{j=1}^{m-1} \left(1 - \frac{j}{n}\right) &\approx \prod_{j=1}^{m-1} e^{-j/n} \\ &= \exp\left(-\sum_{j=1}^{m-1} \frac{j}{n}\right) \\ &= \exp\left(-\frac{1}{n} \frac{m(m-1)}{2}\right) \\ &\approx e^{-m^2/2n} \end{aligned}$$

- ★ Jotta kaikilla m :llä hengellä olisi eri syntymäpäivä tn.:llä $1/2$ on oltava $m^2/(2n) = \ln 2$, eli $m = \sqrt{2n \ln 2}$
- ★ Kun $n = 365$, niin tämä approksimointi antaa $m = 22,49$, kun eksakti laskenta antaa tuloksen $m = 23$

116

- ★ Lasketaan henkilö kerrallaan
- ★ Olk. E_k tapahtuma, että k :nnen henkilön syntymäpäivä ei ole sama kuin ensimmäisten $k - 1$ hengen
- ★ Tällöin tn., ettei k :lla ensimmäisellä hengellä ole eri syntymäpäiviä on

$$\begin{aligned} \Pr(\bar{E}_1 \cup \cdots \cup \bar{E}_k) &\leq \sum_{i=1}^k \Pr(\bar{E}_i) \\ &\leq \sum_{i=1}^k \frac{i-1}{n} \\ &= \frac{k(k-1)}{2n} \end{aligned}$$

- ★ Jos $k \leq \sqrt{n}$ tämä tn. on alle $1/2$, joten $\lfloor \sqrt{n} \rfloor$:llä henkilöllä on eri syntymäpäivät ainakin tn.:llä $1/2$

117

- ★ Oletetaan nyt, että $\lceil \sqrt{n} \rceil$:llä ensimmäisellä henkilöllä on eri syntymäpäivät
- ★ Jokaisella seuraavalla henkilöllä on väh. tn.:llä $\sqrt{n}/n = 1/\sqrt{n}$ sama syntymäpäivä kuin yhdellä $\lceil \sqrt{n} \rceil$ ensimmäisestä henkilöstä
- ★ Siis tn., että kaikilla $\lceil \sqrt{n} \rceil$ seuraavalla henkilöllä on eri syntymäpäivä kuin $\lceil \sqrt{n} \rceil$ ensimmäisellä on kork.

$$\left(1 - \frac{1}{\sqrt{n}}\right)^{\lceil \sqrt{n} \rceil} < \frac{1}{e} < \frac{1}{2}$$

- ★ Eli kun henkilöitä on $2 \lceil \sqrt{n} \rceil$, niin tn., että kaikilla on eri syntymäpäivät on kork. $1/e \approx 0,368$