DRUGI MOMENT

Katja Kristan

18. januar 2008

1 Varianca in neenačba Čebiševa

Poleg matematičnega upanja, je pomembna številska karakteristika slučajne spremenljivke tudi varianca. Lahko ji rečemo tudi razpršenost ali pa disperzija. Varianca meri kako odstopa vrednost slučajne spremenljivke od matematičnega upanja (npr. varianca konstantne slučajne spremenljivke je 0).

Definicija 1 ***Varianca*** *realne slučajne spremenljivke X je*

$$\mathrm{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

Prva enakost sledi iz definicije, drugo pa dobimo po lažjem izračunu. Standardna deviacija X je

$$\sigma = \sqrt{\mathrm{Var[X]}}.$$

V primerjavi z matematičnim upanjem, pa varianca ni linearni funkcional. Zato moramo, če želimo izračunati varianco vsote slučajnih spremenljivk, vedeti nekaj o njihovi (paroma) medsebojni odvisnosti.

Definicija 2 ***Kovarianca*** *slučajnih spremenljivk X in Y je*

$$\mathrm{Cov}[X, Y] = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$$

Lema 3 *Varianca vsote slučajnih spremenljivk je enaka*

$$\mathrm{Var}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + \sum_{i \neq j} \mathrm{Cov}[X_i, X_j].$$

Dokaz:

$$\mathrm{Var}[\sum_{i=1}^{n} X_i] = \mathbf{E}[\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j] - \mathbf{E}[\sum_{i=1}^{n} X_i]\mathbf{E}[\sum_{j=1}^{n} X_j] =$$

$$= \sum_{i=1}^{n} \mathbf{E}[X_i^2] + \sum_{i \neq j} \mathbf{E}[X_i X_j] - \sum_{i=1}^{n} (\mathbf{E}[X_i])^2 - \sum_{i \neq j} \mathbf{E}[X_i]\mathbf{E}[X_j] =$$

$$= \sum_{i=1}^{n} \mathrm{Var}[X_i] + \sum_{i \neq j} \mathrm{Cov}[X_i, X_j]$$

■

Če so $X_1, \ldots, X_n$ neodvisne spremenljivke, je kovarianca vsakega para enaka 0. V tem primeru je

$$\mathrm{Var}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} [\mathrm{Var}[X_i]].$$

Po drugi strani pa iz $\mathrm{Cov}[X, Y] = 0$, ne sledi neodvisnost spremenljivk X in Y.

Zdaj, ko vemo kaj je varianca, se lahko posvetimo Čebiševi neenačbi. Uporabljamo jo, kadar želimo oceniti verjetnost, da se bo slučajna spremenljivka odklonila od njenega matematičnega upanja za vsaj dano število t. Ali drugače: Neenačba Čebiševa ocenjuje kakšna je verjetnost, da se slučajna spremenljivka veliko razlikuje od matematičnega upanja.

Lema 4 (Čebiševa neenakost) *Če ima slučajna spremenljivka X matematično upanje $\mathbf{E}[X]$ in končno varianco* Var[X]*, velja za vsak pozitiven t ocena*

$$P[|X - \mathbf{E}[X]| \geq t] \leq \frac{\mathrm{Var[X]}}{t^2}.$$

Dokaz:

$$\mathrm{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] \geq t^2 P[|X - \mathbf{E}[X]| \geq t]$$

■

Ocena ni natančna. Če je $t^2 < \mathrm{Var}[X]$, je celo prazna, saj je tedaj njena desna stran večja od 1, verjetnost pa tako ali tako ne more biti večja od 1. Najboljši rezultat pa nam da ocena, če je X enak μ z verjetnostjo p ali enak $\mu \pm t$ z verjetnostjo $\frac{1-p}{2}$.

2 Ocenjevanje srednjega binomskega koeficienta

Med binomskimi koeficienti $\binom{2m}{k}$, kjer je $k = 0, 1, \ldots, 2m$, je $\binom{2m}{m}$ največji in se pogosto uporablja v različnih formulah (npr. Catalanova števila imajo enostavno enačbo z binomskimi koeficienti. Ta števila tvorijo zaporedje naravnih števil, ki se pojavlja v mnogih preštevalnih in velikokrat rekurzivnih problemih v kombinatoriki). Metoda s pomočjo drugega momenta je preprost način s katerim navzdol omejimo koeficient $\binom{2m}{m}$. Sicer pa obstajajo tudi drugi pristopi in nekateri med njimi nam dajo celo bolj natančno oceno. Vendar je trik s Čebišejevo neenačbo bolj preprost in tudi ocena je dovolj natančna.

Trditev 5 *Za vsak $m \geq 1$ velja*

$$\binom{2m}{m} \geq \frac{2^{2m}}{(4\sqrt{m} + 2)}.$$

Dokaz: Naj bo X naključna spremenljivka za katero velja $X = X_1 + X_2 + \cdots + X_{2m}$, kjer so spremenljivke X_i neodvisne med sabo in vsaka od njih doseže vrednost 0 in 1 z verjetnostjo $\frac{1}{2}$. Torej je $\mathbf{E}[X] = m$ in $\mathrm{Var}[X] = \frac{m}{2}$. Za $t = \sqrt{m}$ nam da neenačba Čebiševa

$$P[|X - m| < \sqrt{m}] \geq \frac{1}{2}$$

Verjetnost, da X doseže vrednost $m + k$, kjer je $|k| < \sqrt{m}$, je $\binom{2m}{m+k}2^{-2m} \leq \binom{2m}{m}2^{-2m}$, saj je $\binom{2m}{m}$ največji binomski koeficient. Torej imamo

$$\frac{1}{2} \leq \sum_{|k|<\sqrt{m}} P[X = m + k] \leq (2\sqrt{m} + 1)\binom{2m}{m}2^{-2m}.$$

■

3 Pragovna funkcija

Vrnimo se sedaj k slučajnim grafom in razmislimo naslednje vprašanje. Kakšna je verjetnost, da graf $G(n, p)$ vsebuje trikotnik? Omenimo še, da je lastnost, da graf vsebuje trikotnik, monotona, kar pomeni, da če ima graf G neko določeno lastnost in je $G \subset H$, potem ta lastnost velja tudi za graf H. Za majhne p graf $G(n, p)$ verjetno trikotnika ne bo vseboval, medtem ko je za velike p pojav trikotnika v grafu bolj verjeten.

Naj bo T število trikotnikov v grafu $G(n, p)$. Za dano trojico točk je verjetnost, da tvorijo trikotnik p^3. Zaradi linearnosti matematičnega upanja, je pričakovano število trikotnikov

$$\mathbf{E}[T] = \binom{n}{3}p^3.$$

Če je $p(n) \ll \frac{1}{n}$, potem se pričakovana vrednost števila trikotnikov bliža 0, ko večamo n. Zato se tudi verjetnost, da nek graf $G(n, p)$ vsebuje trikotnike nagiba k 0, če je $p(n) \ll \frac{1}{n}$ pri velikih n. Po drugi strani pa, če predpostavimo, da je $p(n) \gg \frac{1}{n}$, matematično upanje števila trikotnikov narašča v neskončnost z naraščanjem n, kar pa ne pomeni, da graf $G(n, p)$ sigurno vsebuje trikotnike. Lahko se zgodi, da nekaj grafov vsebuje veliko trikotnikov in tako dvigne pričakovano vrednost. To lahko ponazorimo tudi z naslednjim življenskim primerom.

Požarno zavarovanje: Letna cena zavarovanja proti požaru na gospodinjstvo narašča. To odraža rast škode, ki jo povzroči ogenj v gospodinjstvu vsako leto. Ampak ali to pomeni, da verjetnost, da bo ogenj povzročil nesrečo, narašča? Ali to celo pomeni, če gledamo v limiti, da bo skoraj vsako gospodinjstvo gorelo vsako leto? Težko. Dvig pričakovane cene škode je posledica parih požarnih nesreč vsako leto, katerih cena se viša.

Na srečo pa se naši trikotniki ne obnašajo tako nepredvidljivo kot požarne nesreče. Za večino slučajnih grafov velja, da je število trikotnikov, ki jih vsebujejo, relativno blizu pričakovane vrednosti. Pravzaprav nam ravno ta metoda s pomočjo drugega momenta dokazuje, da če je pričakovana vrednost števila trikotnikov dovolj velika, potem slučajni graf skoraj sigurno vsebuje nekaj trikotnikov.

Lema 6 *Naj bodo $X_1, X_2, \ldots$ nenegativne slučajne spremenljivke za katere velja*

$$\lim_{n\to\infty} \frac{\mathrm{Var}[X_n]}{(\mathbf{E}[X_n])^2} = 0.$$

Potem

$$\lim_{n\to\infty} P[X_n > 0] = 1.$$

Dokaz: Naj bo $t = \mathbf{E}[X_n]$ in uporabimo Čebiševo neenačbo:

$$P[|X_n - \mathbf{E}[X_n]| \geq \mathbf{E}[X_n]] \leq \frac{\mathrm{Var}[X_n]}{(\mathbf{E}[X_n])^2}.$$

Tako dobimo

$$\lim_{n\to\infty} P[X_n \leq 0] \leq \lim_{n\to\infty} \frac{\mathrm{Var}[X_n]}{(\mathbf{E}[X_n])^2} = 0.$$

■

Zdaj moramo oceniti varianco števila trikotnikov v grafu $G(n,p)$. Število trikotnikov T zapišemo kot $T = \sum_i T_i$, kjer so $T_1, T_2, \ldots$ indikatorske spremenljivke za vse $\binom{n}{3}$ možne trikotnike v grafu $G(n,p)$. Varianca vsote slučajnih spremenljivk je

$$\mathrm{Var}[T] = \sum_i \mathrm{Var}[T_i] + \sum_{i\neq j} \mathrm{Cov}[T_i, T_j].$$

Za vsak trikotnik velja

$$\mathrm{Var}[T_i] \leq \mathbf{E}[T_i^2] = p^3$$

in za vsak par trikotnikov, ki ima skupno eno povezavo velja

$$\mathrm{Cov}[T_i, T_j] \leq \mathbf{E}[T_i T_j] = p^5,$$

(torej sta T_i in T_j indikatorski spremenljivki dveh trikotnikov na petih določenih povezavah).

Indikatorske spremenljivke trikotnikov, ki nimajo skupne povezave, so neodvisne in je torej kovarianca takih parov enaka 0. Zato seštevamo le kovarianco tistih parov trikotnikov, ki imajo skupno povezavo. Število le teh je $12\binom{n}{4}$ in tako dobimo

$$\mathrm{Var}[T] \leq \binom{n}{3}p^3 + 12\binom{n}{4}p^5 \leq n^3p^3 + n^4p^5$$

$$\frac{\mathrm{Var}[T]}{(\mathbf{E}[T])^2} \leq \frac{n^3p^3 + n^4p^5}{(\binom{n}{3}p^3)^2} = O(\frac{1}{n^3p^3} + \frac{1}{n^2p}),$$

kar limitira proti 0, če je $p(n) \gg \frac{1}{n}$. Po lemi 6 pa iz tega sledi, da se verjetnost, da graf $G(n,p)$ vsebuje trikotnike, približuje 1 z naraščanjem n proti neskončnosti.

Lastnost, ali graf vsebuje trikotnik, lahko preverimo tudi s pomočjo pragovne funkcije. Ta pojem sta vpeljala Erdös in Rényi.

Preden pa definiramo pragovno funkcijo, ponovimo še kaj pomeni, da je lastnost A monotona. Naj bosta G in H grafa, za katera velja $V(H) = V(G)$ in $E(H) \subseteq E(G)$. Lastnost A je monotona, če velja: H ima lastnost A, potem sledi, da ima lastnost A tudi G.

Definicija 7 *Funkcija $r : \mathbb{N} \to R$ je **pragovna funkcija** za monotono lastnost A grafa $G(n,p)$, če za vsak $p : \mathbb{N} \to [0,1]$ velja:*

1. $p(n) \ll r(n) \Rightarrow \lim_{n\to\infty} P[A \text{ velja za } G(n,p(n))] = 0$,
2. $r(n) \ll p(n) \Rightarrow \lim_{n\to\infty} P[A \text{ velja za } G(n,p(n))] = 1$.

Pragovna funkcija lahko obstaja lahko pa tudi ne. Tudi če že obstaja, ni nujno, da je enolično določena. Naprimer za našo lastnost, da $G(n,p)$ vsebuje trikotnik, je pragovna funkcija $r(n) = 1/n$. Vendar pa bi lahko namesto tega zapisa uporabili tudi $r(n) = c/n$(za vsak $c > 0$).

Lahko pa bi se ukvarjali tudi s pragovno funkcijo bolj splošnih lastnosti, kot je na primer pojav podgrafa v grafu $G(n,p)$ (vendar ne nujno induciranega, saj je ta problem veliko bolj zapleten). Izkaže se, da je naš ukrep primeren tudi za bolj splošne lastnosti podgrafa, pod pogojem, da je ta uravnotežen. Preden pa povemo kaj je uravnotežen graf, pa definirajmo še gostoto grafa.

Definicija 8 *Naj bo H graf z v vozlišči in e povezavami. Definiramo **gostoto** grafa H kot*

$$\rho(H) = \frac{e}{v}.$$

Grafu H rečemo, da je uravnotežen, če noben njegov podgraf nima večje gostote kot graf H sam.

Izrek 9 *Naj bo H uravnotežen graf z gostoto ρ. Potem je $r(n) = n^{\frac{-1}{\rho}}$ pragovna funkcija za lastnost, da je H podgraf grafa $G(n,p)$.*

Dokaz: Naj ima graf H v vozlišč in e povezav. Potem je gostota $\rho(H) = \frac{e}{v}$. Označimo vozlišča grafa H z $a_1, a_2, \ldots, a_v$. Za vsako urejeno v-terico $\beta = (b_1, b_2, \ldots, b_v)$ različnih vozlišč $b_1, b_2, \ldots, b_v \in V(G(n,p))$, naj A_β označuje dogodek, ko $G(n,p)$ vsebuje pravilno urejeno kopijo H na $(b_1, b_2, \ldots, b_v)$. To je, dogodek A_β se zgodi, če velja $b_i b_j \in E(G(n,p))$, vedno kadar velja $a_i a_j \in E(H)$. Z drugimi besedami,dogodek A_β se zgodi vedno ko je predpis $a_i \longrightarrow b_i$ homomorfizem na grafu.

Naj X_β označuje indikatorske spremenljivke nanašujoče se na A_β in naj bo $X = \sum_\beta X_\beta$ po vseh urejenih v- tericah β. Upoštevati moramo, da so zaradi možne simetrije H, nekatere kopije H lahko štete večkrat in tako X ni točno število kopij H v grafu $G(n,p)$. Kakorkoli pogoj $X = 0$ je ekvivalenten odsotnosti grafa H v $G(n,p)$ in $X > 0$ je ekvivalenten pojavu grafa H v grafu $G(n,p)$.

Verjetnost A_β je p^e. Zaradi linearnosti matematičnega upanja pa velja

$$\mathbf{E}[X] = \sum_\beta P[A_\beta] = \Theta(n^v p^e)$$

(upoštevati je treba, da sta v in e konstanti, medtem ko je p funkcija od n).

Če je $p(n) \ll n^{\frac{-v}{e}}$, potem je

$$\lim_{n\to\infty} \mathbf{E}[X] = 0,$$

s čimer je prvi korak dokazan.

Sedaj pa predpostavimo, da je $p(n) \gg n^{\frac{-v}{e}}$ in uporabimo lemo 3

$$\mathrm{Var}[X] = \sum_{\beta} \mathrm{Var}[X_\beta] + \sum_{\beta\neq\gamma} \mathrm{Cov}[X_\beta, X_\gamma].$$

Ker je $\mathrm{Var}[X_\beta] = \mathrm{Cov}[X_\beta X_\beta]$, lahko pišemo tudi

$$\mathrm{Var}[X] = \sum_{\beta,\gamma} \mathrm{Cov}[X_\beta X_\gamma]$$

Kovarianca je neničelna le za pare kopij, ki si delijo nekaj povezav. Naj si β in γ delita $t \geq 2$ vozlišč. Potem imata dve kopiji H največ $t\rho$ skupnih povezav (saj je H uravnotežen), njuna unija pa vsebuje vsaj $2e - t\rho$ povezav. Tako je

$$\mathrm{Cov}[X_\beta X_\gamma] \leq \mathbf{E}[X_\beta X_\gamma] \leq p^{2e-t\rho}$$

Število parov β in γ, ki si delijo t vozlišč je reda $O(n^{2v-t})$, saj lahko izberemo množico vozlišč moči $2v - t\rho$ na $\binom{n}{2v-t}$ načinov. Za fiksen t dobimo

$$\sum_{|\beta\cap\gamma|=t} \mathrm{Cov}[X_\beta, X_\gamma] = O(n^{2v-t}p^{2e-t\rho}) = O((n^v p^e)^{2-t/v})$$

$$\mathrm{Var}[X] = O(\sum_{t=2}^{v} (n^v p^e)^{2-t/v})$$

in

$$\lim_{n\to\infty} \frac{\mathrm{Var}[X]}{(\mathbf{E}[X])^2} = \lim_{n\to\infty} O(\sum_{t=2}^{v} (n^v p^e)^{-t/v}) = 0,$$

če je $\lim_{n\to\infty} n^v p^e = \infty$. S tem pa je izrek dokazan, saj po lemi 6 velja

$$\lim_{n\to\infty} P[X > 0] = 1$$

in tako se skoraj vedno pojavi kopija H v grafu $G(n,p)$.

■

Vprašanje pojava splošnih podgrafov H v grafu sta rešila Eödos in Rényi. Pragovna funkcija za graf H je določena s podgrafom $\bar{H} \subset H$, ki ima maksimalno gostoto. Zapišimo izrek:

Izrek 10 *Naj bo H graf in $\bar{H} \subset H$ z maksimalno gostoto ρ. Potem je*

$$r(n) = n^{\frac{-1}{\rho(H)}}$$

pragovna funkcija za lastnost, da je H podgraf grafa $G(n,p)$.

Literatura

[1] J. Matoušek, J. Vondrák, *The probalistic method (Lecture notes)*, Praga: ITI, 2002.