PRAVDEPODOBNOSŤ a ŠTATISTIKA

Zákony veľkých čísel a centrálna limitná veta

doc. RNDr. Štefan Peško, CSc.

Katedra matematických metód a operačnej analýzy, FRI ŽU

11. apríla 2019

Motivácia

Ak opakujeme nezávisle určitý pokus, môžeme z pozorovaných hodnôt zostaviť rozdelenie relatívnych početností niektorej náhodnej udalosti a pokúsiť sa zistiť jej rozdelenie alebo charakteristiky.

Očakávame, že pri dodržaní istých podmienok sa s rastúcim počtom opakovaní bude empirické rozdelenie početností približovať k teoretickému rozdeleniu.

Túto myšlienku presnejšie upravujú zákony veľkých čísel.

Konvergencia podľa pravdepodobnosti

Hovoríme, že postupnosť náhodných premenných $\{\mathbf{X}_n\}_{n=1}^{\infty}$ konverguje podľa pravdepodobnosti ku konštante c, ak pre každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} \mathcal{P}(|\mathbf{X}_n - c| < \varepsilon) = 1. \tag{1}$$

resp.

$$\lim_{n\to\infty} \mathcal{P}(|\mathbf{X}_n - c| \geq \varepsilon) = 0. \tag{2}$$

Pre konvergenciu podľa pravdepodobnosti budeme používať symbol $\mathbf{X}_n \xrightarrow{\mathcal{P}} c$.

Tvrdenie 16 (Bernoulli)

Majme postupnosť $\{\mathbf{X}_n\}_{n=1}^{\infty}$ navzájom nezávislých náhodných premenných, ktoré majú to isté binomické rozdelenie $\mathbf{X}_n \sim Bi(n, p)$. Potom pre každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} \mathcal{P}\left(\left|\frac{\mathbf{X}_n}{n} - p\right| \geq \varepsilon\right) = 0. \tag{3}$$

Dôkaz: Pre každé n platí

$$E\left(\frac{\mathbf{X}_n}{n}\right) = \frac{np}{n} = p,\ D\left(\frac{\mathbf{X}_n}{n}\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Z Čebyšovej nerovnosti [1] dostaneme

$$\mathcal{P}\left(\left|\frac{\mathbf{X}_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2},$$

odkiaľ je tvrdenie zrejmé.

[1] $\mathcal{P}(|\mathbf{Y} - E(\mathbf{Y})| \geq \beta) \geq \frac{D(\mathbf{Y})}{\beta^2}$ pre náhodnú premennú $\mathbf{Y}$ existujú $E(\mathbf{Y})$ a $D(\mathbf{Y})$ a ľubovoľné kladné reálne číslo β.

Tvrdenie hovorí, že s rastúcim počtom nezávislých pokusov, postupnosť relatívnych početností výskytu náhodnej udalosti A „nejako“ konverguje k $\mathcal{P}(A)$, ak je rovnaká vo všetkých pokusoch. Keď zapíšme (3) pomocou doplnkovej udalosti, dostaneme

$$\lim_{n\to\infty} \mathcal{P}\left(\left|\frac{\mathbf{X}_n}{n} - p\right| < \varepsilon\right) = 1.$$

Ak si zvolíme ľubovoľne malé $\varepsilon > 0$, vždy máme jednotkovú pp., že v limite sa relatívna početnosť líši od odhadovanej pp. o menej než je toto malé ε.

Príklad 7.1

Aká je pravdepodobnosť, že pri 100 hodoch mincou je odhad pp. hodu znaku 1/2 pri chybe odhadu 1% ?

Priaznivou udalosťou je hod znaku. Počet priaznivych hodov v n pokusoch udáva náhodná premenná $\mathbf{X}_n \sim Bi(n, p)$ kde $n = 100, p = 1/2$. Chyba odhadu $\varepsilon = 0.01$.

Potom po dosadní do upravenej Čebyševovej nerovnosti

$$\mathcal{P}\left(\left|\frac{\mathbf{X}_n}{n} - p\right| < \varepsilon\right) \geq 1 - \frac{np(1-p)}{\varepsilon^2}.$$

dostaneme

$$\mathcal{P}\left(\left|\frac{\mathbf{X}_{100}}{100} - \frac{1}{2}\right| < 0.01\right) = \mathcal{P}(|\mathbf{X}_{100} - 50| < 1) \geq 1 - \frac{\frac{1}{2}\frac{1}{2}}{100 \cdot 1^2} = 0.9975$$

Tvrdenie 17 (Čebyšev)

Majme takú postupnosť $\{\mathbf{X}_i\}_{i=1}^n$, že $E(\mathbf{X}_i) < \infty$ a $D(\mathbf{X}_i) = c < \infty$. Potom pre každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} \mathcal{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i - \frac{1}{n}\sum_{i=1}^{n}E(\mathbf{X}_i)\right| < \varepsilon\right) = 1. \qquad (4)$$

Pri splnení uvedených podmienok

$$\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \xrightarrow{\mathcal{P}} \frac{1}{n}\sum_{i=1}^{n}E(\mathbf{X}_i).$$

Ak majú naviac náhodné premenné $\mathbf{X}_i$ to isté rozdelenie so strednou hodnotou $E(\mathbf{X}_i) = \mu$ a konečným rozptylom $D(\mathbf{X}_i) = \sigma^2$, potom $\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \xrightarrow{\mathcal{P}} \mu$.

Príklad 7.2

Vypočítajte pravdepodobnosť, že aritmetický priemer z 10 000 nezávislých meraní udáva skutočnú hodnotu meranej veličiny μ s presnosťou 0.01 ak rozptyl jednotlivých meraní nepresiahne 0.02 .

Výsledky $n = 10000$ meraní reprezentujme premennými $X_i; i = 1, \ldots, n$ s $E(\mathbf{X}_i) = \mu, c = D(\mathbf{X}_i) = 0.02$.

Po dosadení do upravenej Čebyševovej nerovnosti dostaneme

$$\mathcal{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i - \mu\right| < \varepsilon\right) \geq 1 - \frac{c}{n\varepsilon^2} = 1 - \frac{0.02}{10000(0.01)^2} = 0.98.$$

Centrálna limitná veta

Podstatou všetkých formulácií centrálnej limitnej vety je tvrdenie, že súčet veľkého počtu nezávislých náhodných premenných s konečnou strednou hodnotou a konečným rozptylom má asymptoticky normálne rozdelenie.

Tento výsledok sa využíva v praxi. Ak aj skúmané náhodné veličiny v jednotlivých pokusoch nemajú normálne rozdelenie, tak s ich súčtom sa pri dostatočne veľkom počte pracuje ako s normálne rozdelenou náhodnou premennou.

Tvrdenie 18 (Linderberg-Lévy)

Majme postupnosť $\{\mathbf{X}_i\}_{i=1}^n$ navzájom nezávislých náhodných premenných, ktoré majú rovnaký pravdepodobnostné rozdelenie, konečnú strednú hodnotu $E(\mathbf{X}_i) = \mu$ a konečný rozptyl $D(\mathbf{X}_i) = \sigma^2$. Nech $\mathbf{Y}_n = \sum_{i=1}^{n} \mathbf{X}_i$. Potom normovaná náhodná premenná

$$\mathbf{U}_n = \frac{\mathbf{Y}_n - E(\mathbf{Y}_n)}{\sqrt{D(\mathbf{Y}_n)}}$$

má asymptoticky štandardizované normálne rozdelenie pp. $N(0,1)$ tj. platí

$$\lim_{n\to\infty} \mathcal{P}(\mathbf{U}_n < u) = \Phi(u),\ u \in \Re.$$

Príklad 7.3

Zaujíma nás neznámy podiel p osôb s krvnou skupinou A v danej populácii. U koľkých osôb musíme zistiť, či má alebo nemá skupinu A, aby sme s pp. aspoň 0.9 odhadli neznámu pp. s chybou nanajvýš 0.05?

Vyberieme z danej populácie náhodne n osôb. Náhodný počet osôb s krvnou skupinou A je náhodnou premennou $\mathbf{Y} \sim Bi(n, p)$. Neznámy podiel p budeme odhadovať pomocou Y/n. Počet oslovených osôb zvolíme tak, aby platilo

$$\mathcal{P}\left(\left|\frac{1}{n}\mathbf{Y} - p\right| < 0.05\right) \approx 0.9.$$

Použijeme (5) tj. $\mathbf{Y} \sim N(np, np(1-p))$ a postupnými úpravami dostaneme

$$0.9 \approx \mathcal{P}\left(\left|\frac{1}{n}\mathbf{Y} - p\right| < 0.05\right) = \mathcal{P}(-0.05n + np < Y < 0.05n + np)$$

$$= \Phi\left(\frac{0.05n}{\sqrt{np(1-p)}}\right) - \left[1 - \Phi\left(\frac{0.05n}{\sqrt{np(1-p)}}\right)\right] = 2\Phi\left(\frac{0.05n}{\sqrt{np(1-p)}}\right) - 1.$$

Teda by malo byť

$$\Phi\left(\frac{0.05n}{\sqrt{np(1-p)}}\right) \approx \frac{1+0.9}{2} = 0.95,$$

čo zaručíme voľbou

$$\frac{0.05n}{\sqrt{np(1-p)}} = \Phi^{-1}(0.95) = 1.644854.$$

Pri voľbe $p = \frac{1}{2}$ máme $p(1-p) \to max$ a tak po úprave dostaneme

$$n > 100 \cdot 1.644854^2 \approx 270.$$

Príklad 7.4

Pomocou centrálnej limitnej vety vyjadrite $\mathcal{P}(\sum_{i=1}^{n} \mathbf{X}_i < a)$ ak sú $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ nezávislé náhodné premenné s rozdelením $N(1,4)$, resp. $A(\frac{1}{5}), R(0,2), R(-2,2)$.

Podľa tvrdenia 18 možeme odhadnúť

$$\mathbf{U} = \sum_{i=1}^{n} \mathbf{X}_i \sim N(nE(\mathbf{X}_1), nD(\mathbf{X}_1)) \implies \mathcal{P}(\mathbf{U} < a) = \Phi\Big(\frac{a - nE(\mathbf{X}_1)}{\sqrt{nD(\mathbf{X}_1)}}\Big)$$

- $\mathbf{X}_i \sim N(1,4)$: $\mathcal{P}(\mathbf{U} < a) = \Phi\big(\frac{a-n}{2\sqrt{n}}\big)$,
- $\mathbf{X}_i \sim A(\frac{1}{5})$: $\mathcal{P}(\mathbf{U} < a) \approx \Phi\big(\frac{5a-n}{2\sqrt{n}}\big)$,
- $\mathbf{X}_i \sim R(0,2)$: $\mathcal{P}(\mathbf{U} < a) \approx \Phi\big(\frac{(a-n)\sqrt{3}}{\sqrt{n}}\big)$,
- $\mathbf{X}_i \sim R(-2,2)$: $\mathcal{P}(\mathbf{U} < a) \approx \Phi\big(\frac{a\sqrt{3}}{2\sqrt{n}}\big)$.

Príklad 7.5

V určitej oblasti je 3% chorých na maláriu. Aká je pravdepodobnosť, že pri kontrole 5 000 ľudí nájdeme $3 \pm 0.5\%$ chorých na maláriu.

Podiel chorých

$$\frac{\mathbf{S}_{5000}}{5000} \sim Bi(5000, 0.03) \approx N(0.03 \cdot 5000, 5000 \cdot 0.03 \cdot 0.97).$$

Potom

$$\mathcal{P}\left(2.5 \leq \frac{\mathbf{S}_{5000}}{5000} \leq 3.5\right) \approx \Phi\left(\frac{150}{\sqrt{145.5}}\right) - \Phi\left(\frac{125}{\sqrt{145.5}}\right) = 0.962$$

Tvrdenie 19 (Ljapunova veta)

Majme postupnosť $\{\mathbf{X}_i\}_{i=1}^n$ navzájom nezávislých náhodných premenných, pričom existujú ich strednú hodnoty $E(\mathbf{X}_i)$, rozptyly $D(\mathbf{X}_i)$ a tretie centrálne momenty $\mu_3(\mathbf{X}_i)$. Nech je splnená Ljapunova podmienka:

$$\lim_{n\to\infty} \frac{\sqrt[3]{\sum_{i=1}^n \mu_3(\mathbf{X}_i)}}{\sqrt{\sum_{i=1}^n D(\mathbf{X}_i)}} = 0. \tag{5}$$

Potom pre normovanú náhodnú premennú

$$\mathbf{U}_n = \frac{\mathbf{Y}_n - E(\mathbf{Y}_n)}{\sqrt{D(\mathbf{Y}_n)}},\ \mathbf{Y}_n = \sum_{i=1}^n \mathbf{X}_i$$

platí

$$\lim_{n\to\infty} \mathcal{P}(\mathbf{U}_n < u) = \Phi(u),\ u \in \Re.$$

Príklad 7.6

Prístroj sa skladá z 50 častí, ktoré nezávisle na sebe môžu mať poruchu. Bolo zistené, že stredné hodnoty a disperzie počtu porúch jednotlivých častí prístroja počas určitého časového intervalu, t.j. náhodných premenných $\{\mathbf{X}_i\}_{i=1}^{n}$ sú $E(\mathbf{X}_i) = 0.05 \cdot i$ a $D(\mathbf{X}_i) = 0.02 \cdot i$. Aká je pravdepodobnosť, že celkový počet porúch častí prístroja počas daného časového intervalu je menšia než 74?

Nech náhodná premenná $\mathbf{Y} = \sum_{i=1}^{50} \mathbf{X}_i$ predstavuje celkový počet porúch častí prístroja počas daného časového intervalu. Za predpokladu, že je splnená Ljapunova podmienka (5) vypočítame

$$E(\mathbf{Y}) = \sum_{i=1}^{50} E(\mathbf{X}_i) = 0.05 \frac{50 \cdot 51}{2} = 63.75,$$

$$D(\mathbf{Y}) = \sum_{i=1}^{50} D(\mathbf{X}_i) = 0.02 \frac{50 \cdot 51}{2} = 25.5.$$

Podľa Ljapunovej vety dostávame nasledujúci odhad $\mathcal{P}(\mathbf{Y} < 74) = \mathcal{P}(\mathbf{U} < \frac{74-63.75}{\sqrt{25.5}}) \approx \Phi(2.03) \doteq 0.979$.

Cvičenie

7.1 Počet chýb na jednej strane textu má strednú hodnotu 8 a rozptyl 4. Aká je pravdepodobnosť, že na 100 stranách bude menej než 750 chýb?

7.2 Odhadnite pp. s akou bude počet šestiek, ktoré padnú v 1000 nezávislých hodoch spravodlivou kockou, v intervale $\langle 147, 186 \rangle$.

7.3* V 5 000 nezávislých hodoch mincou padlo spolu 2 507× hlava. Môžeme považovať mincu za spravodlivú?

7.4* Pravdepodobnosť, že bude výrobok reklamovaný je 0.05. Aká je pp., že z 300 predaných výrobkov ich bude nanajvýš 20 reklamovaných?

7.5* Pri zostavovaní štatistického výkazu bolo spočítaných 10^3 čísel, pričom každé bolo zaokrúhlené na m desatinných miest. Chyba vznikajúca zaokrúhlením má rovnomerné rozdelenie v rozsahu $(-0.5 \cdot 10^{-m}, 0.5 \cdot 10^{-m})$. Vypočítajte hranicu, ktorú s pp. 0.99 neprekročí absolútna hodnota celkovej chyby súčtu.