Přednáška 10 a+b

Opakování číselných charakteristik náhodných veličin – střední hodnota, rozptyl.

Čebyševova nerovnost

Předpokládejme, že náhodná veličina X má konečný rozptyl a uvažujme $\varepsilon > 0$. Potom platí

$$P(|X - EX| \geq \varepsilon) \leq \frac{\operatorname{var} X}{\varepsilon^2}$$

Důkaz: Pro spojitou náhodnou veličinu. Položme $\mu = EX$ a počítejme podle definice

$$\operatorname{var} X = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{|x-\mu| \geq \varepsilon} (x-\mu)^2 f(x)\,dx$$

$$+ \int_{|x-\mu| < \varepsilon} (x-\mu)^2 f(x)\,dx \geq \int_{|x-\mu| \geq \varepsilon} \varepsilon^2 f(x)\,dx = \varepsilon^2 P(|X-\mu| \geq \varepsilon)$$

$\blacksquare$

Důsledky: Pro $\varepsilon = k\sigma$, kde $\sigma = \sqrt{\operatorname{var} X} \neq 0$, je

$$P(|X - EX| \geq k\sigma) \leq \frac{1}{k^2}.$$

Kovariance

Pro náhodné veličiny X, Y s existujícími středními hodnotami a rozptyly definujeme jejich kovarianci předpisem

$$\operatorname{cov}(X,Y) = E\big((X-EX)(Y-EY)\big)$$

Speciálně $\operatorname{cov}(X,X) = \operatorname{var} X$.

<u>Věta (Vlastnosti rozptylu.)</u>

Jestliže pro náhodné veličiny X, Y a Z existují jejich rozptyly, pak pro reálná čísla a, b, c, d platí

(1) $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$

(2) $\operatorname{cov}(X,Y) = E(X\cdot Y) - (EX)(EY)$

(3) $\operatorname{cov}(X+Y,Z) = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z)$

(4) $\operatorname{cov}(a+bX, c+dY) = bd\operatorname{cov}(X,Y)$

(5) $\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y + 2\operatorname{cov}(X,Y)$

Jsou-li veličiny X a Y nezávislé, pak $\operatorname{cov}(X,Y)=0$ a

$$\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y$$

<u>Důkaz:</u> (1) je zřejmé.

(2) $\operatorname{cov}(X,Y) = E\{(X-E(X))(Y-E(Y))\} = E\{XY - E(X)Y - XE(Y) + E(X)(E(Y))\} = E(XY) - 2E(X)E(Y) + E(X)E(Y) = E(XY) - (EX)(EY)$

(3) $\operatorname{cov}(X+Y, Z) = E((X+Y)\cdot Z) - E(X+Y)\,E(Z)$

$= E(X\cdot Z) + E(Y\cdot Z) - EX\,EZ - EY\,EZ =$

$= \operatorname{cov}(X, Z) + \operatorname{cov}(Y, Z)$

(4) $\operatorname{cov}(a+bX, c+dY) = E\{(a+bX - E(a+bX))(c+dY - E(c+dY))\} = E\{(bX - bEX)(dY - dEY)\} =$

$= E\{bd\,(X-EX)(Y-EY)\} = bd \operatorname{cov}(X,Y).$

(5) $\operatorname{var}(X+Y) = E\left(\left((X+Y) - E(X+Y)\right)^2\right) =$

$= E\left(\left((X-EX) + (Y-EY)\right)^2\right) =$

$= E(X-EX)^2 + 2E\left((X-EX)(Y-EY)\right)$

$+ E(Y-EY)^2 = \operatorname{var} X + 2\operatorname{cov}(X,Y) + \operatorname{var} Y$

Pokud jsou X a Y nezávislé, jsou nezávislé i $X-EX$, $Y-EY$ a platí

$\operatorname{cov}(X,Y) = E\left((X-EX)(Y-EY)\right) = E(X-EX)\cdot E(Y-EY) = 0.$

Proto

$$\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y$$

Střední hodnota a variace binomického rozdělení $Bi(n, p)$

Binomické rozdělení je rozdělení náhodné veličiny X, která je součtem n náhodných nezávislých veličin s alternativním rozdělením.

Alternativní rozdělení $A(p)$ má

$$E = p$$
$$\sigma = p(1-p)$$

Proto binomické rozdělení $Bi(n, p)$ má

$$E = np$$
$$\sigma = np(1-p)$$

Použití Čebyševovy věty

Jaká je pravděpodobnost, že při 1200 hodech aspoň 150 krát a nejvýše kostkou padne šestka aspoň 150 krát a nejvýše 250 krát.

Přesný výpočet X je náhodná veličina počet šestek při 1200 hodech. Má binomické rozložení $Bi(1200, \frac{1}{6})$. Platí

$$P(150 \leq X \leq 250) = \sum_{k=150}^{250} \binom{1200}{k} \frac{1}{6^k} \cdot \frac{5^{n-k}}{6^{n-k}}$$

To však těžko spočítáme. Proto udělejme odhad této pravděpodobnosti pomocí Čebyševovy věty.

$$EX = 1200 \cdot \frac{1}{6} \quad (= np) \quad = 200$$

$$\operatorname{var} X = 1200 \cdot \frac{1}{6} \cdot \frac{5}{6} \quad (= np(1-p)) = \frac{500}{3}$$

Jde nám o pravděpodobnost jevu $|X-200| \leq 50$:

$$P(|X-200| \leq 50) = 1 - P(|X-200| \geq 51) \geq$$

$$\geq 1 - \frac{\operatorname{var} X}{51^2} = 1 - \frac{500}{3 \cdot 51^2} \approx 0{,}94.$$

Tedy pravděpodobnost, že $|X-200| \leq 50$ je aspoň 0,94. $\blacksquare$

Korelační koeficient

Kovariance vypovídá o závislosti náhodných veličin X a Y. Proto zavádíme korelační koeficient

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X}\sqrt{\operatorname{var} Y}}$$

Poznámka 1 $\operatorname{cov}(X,Y)$ má roli skalárního součinu na prostoru náhodných veličin s konečnou variací. $\sqrt{\operatorname{var} X}$ má ~~[illegible]~~ roli normy a $\rho_{X,Y}$ má roli úhlu. (Pozor $\operatorname{cov}(X,X)$ může být 0 i pro X nenulovou)

Korelační koeficient X a Y se rovná korelačnímu koeficientu normovaných náhodných veličin

$$\frac{1}{\sqrt{\operatorname{var} X}}(X - EX) \quad a \quad \frac{1}{\sqrt{\operatorname{var} Y}}(Y - EY)$$

<u>Věta</u> Je-li korelační koeficient definován, pak

$$|\rho_{X,Y}| \leq 1.$$

Rovnost nastane právě když $P(Y = kX + c) = 1$ pro nějaké konstanty k a c.

<u>Variační matice</u>

Uvažujme náhodný vektor $X = (X_1, X_2, .. X_n)^T$, jehož všechny komponenty mají konečný rozptyl. Variační matice náhodného vektoru X je

$$\operatorname{var} X = \begin{pmatrix} \operatorname{var} X_1 & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{var} X_2 & \cdots & \operatorname{cov}(X_2, X_n) \\ & & \ddots & \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \cdots & \operatorname{var} X_n \end{pmatrix}$$

nebo pomocí střední hodnoty

$$EX = (EX_1, EX_2, .., EX_n)^T$$

jako

$$\operatorname{var} X = E\{(X-EX)(X-EX)^T\}$$

<u>Věta</u>: Uvažujme náhodný vektor $X=(X_1,\dots X_n)^T$ matici B typu $k\times n$ a vektor $c\in\mathbb{R}^k$. Potom

$$\operatorname{var}(BX+c) = B(\operatorname{var} X)B^T.$$

Momenty a momentová funkce

k-tý moment náhodné veličiny X

$$\mu'_k = EX^k$$

k-tý centrální moment veličiny X

$$\mu_k = E(X-EX)^k$$

Pro spojitou náhodnou veličinu je

$$EX^k = \int_{-\infty}^{\infty} x^k f(x)\,dx,$$

kde f je hustota rozložení X. Pro diskrétní náhodnou veličinu je

$$EX^k = \sum_i x_i^k f(x_i),$$

kde f je pravděpodobnostní funkce.

<u>Momentová vytvořující funkce</u> náhodné veličiny X je

$$M_X(t) = E\, e^{tX} = \begin{cases} \sum_i e^{tx_i} f(x_i) \\ \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \end{cases}$$

<u>Věta</u>

Nechť X je náhodná veličina, pro kterou na intervalu $(-a, a)$ existuje její analytická ~~funkce~~ momentová funkce. Pak na tomto intervalu je

$$M_X(t) = \sum \frac{t^k}{k!} E X^k.$$

<u>Důkaz:</u> Pro spojitou náhodnou veličinu počítáme:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_{-\infty}^{\infty} \sum_{k=0}^{\infty} \frac{t^k x^k}{k!} f(x) dx =$$

$$= \sum_{k=0}^{\infty} \frac{t^k}{k!} \int_{-\infty}^{\infty} x^k f(x) dx = \sum_{k=0}^{\infty} \frac{t^k}{k!} E X^k \qquad \blacksquare$$

Věta Pro nezávislé náhodné veličiny X a Y platí

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Dále

$$M_{a+bX}(t) = e^{at} M_X(bt).$$

Důkaz:

$$M_{X+Y}(t) = E e^{t(X+Y)} = E\left(e^{tX} \cdot e^{tY}\right) = E(e^{tX}) \cdot E(e^{tY})$$

neboť střední hodnota součinu nezávislých náhodných veličin je součinem jejich středních hodnot.

$$M_{a+bX}(t) = E e^{t(a+bX)} = E\left(e^{ta} \cdot e^{tbX}\right) =$$

$$= e^{ta} \cdot E\left(e^{tbX}\right) = e^{ta} M_X(tb).$$

■

Příklad 1 momentová vytvořující funkce pro $X \sim Bi(n,p)$

Často je jednodušší počítat momentovou funkci než přímo počítat momenty.

Pro alternativní rozdělení náh. veličiny $Y \sim A(p)$

$$M_Y(t) = E e^{tY} = e^0 (1-p) + e^t p = p(e^t - 1) + 1$$

Současně $M_Y(t) = 1 + tEY + \frac{t^2}{2} EY^2 + \frac{t^3}{3} EY^3 + \dots$

Tedy

$$EY = \frac{d}{dt}\left(M_Y(t)\right)_{t=0} = \left(pe^t\right)_{t=0} = p$$

$$EY^2 = \frac{d^2}{dt^2}\left(M_Y(t)\right)_{t=0} = \left(pe^t\right)_{t=0} = p$$

$$\operatorname{var} Y = EY^2 - (EY)^2 = p - p^2 = p(1-p)$$

Binomické rozdělení $X \sim Bi(n,p)$ je součtem n nezávislých veličin $Y_i \sim A(p)$. Proto

$$M_X(t) = \left(p(e^t-1)+1\right)^n = 1 + \binom{n}{1} p(e^t-1) + \binom{n}{2} p^2 (e^t-1)^2 + \dots + \binom{n}{n} p^n (e^t-1)^n$$

$$EX = \left(\frac{d}{dt} M_X(t)\right)_{t=0} = \left(npe^t\right)_{t=0} = np$$

$$EX^2 = \left(\frac{d^2}{dt^2} M_X(t)\right)_{t=0} = \left(\frac{d^2}{dt^2} np(e^t-1) + \frac{n(n-1)}{2} p^2 \frac{d^2}{dt^2}(e^t-1)^2\right)_{t=0}$$

$$= \left(npe^t\right)_{t=0} + \frac{n(n-1)}{2} p^2\, 2\left(2e^{2t} - e^t\right)_{t=0} =$$

$$= np + n(n-1)p^2$$

Proto

$$\operatorname{var} X = EX^2 - (EX)^2 = np + n(n-1)p^2 - n^2p^2$$

$$= np(1-p),$$

což jsme odvodili již dříve.

<u>Příklad 2</u> momentová vytvořující funkce

pro $Z \sim N(0,1)$

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx =$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2 - 2tx + t^2 - t^2}{2}\right) dx$$

$$= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2}\right) dx$$

$$= \exp\left(\frac{t^2}{2}\right)$$

Odtud

$$EZ = \left(\frac{d}{dt} M_Z(t)\right)_{t=0} = \left(t \cdot \exp\left(\frac{t^2}{2}\right)\right)_{t=0} = 0$$

$$EZ^2 = \left(\frac{d^2}{dt^2} M_Z(t)\right)_{t=0} = \left(\exp\left(\frac{t^2}{2}\right) + t^2 \exp\left(\frac{t^2}{2}\right)\right)_{t=0} = \underline{1}$$

Tedy $\operatorname{var} Z = EZ^2 - (EZ)^2 = 1$.

$Y_1, Y_2, \ldots, Y_n$ nezávislé náhodné veličiny se (stejným rozdělením a) střední hodnotou 0 a rozptylem 1.

Předpokládejme, že $E|Y_i|^3$ je konečný.

Pro náhodnou veličinu

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

spočteme momentovou funkci

$$M_{S_n}(t) = \prod_{i=1}^{n} E\, e^{(t/\sqrt{n})Y_i} = \left(M_Y(t/\sqrt{n})\right)^n,$$

kde M_Y je společná momentová funkce všech Y_i.

Nyní

$$M_{S_n}(t/\sqrt{n}) = \left(1 + 0 \cdot \frac{t}{\sqrt{n}} + 1 \cdot \frac{t^2}{2n} + \ldots\right)^n$$

~~$= \left(1 + \frac{t^2}{2n} + \text{členy vyšších řádů}\ \frac{1}{n}\right)$~~

V limitě dostáváme (aspoň na intervalu $(-a, a)$)

$$\lim_{n\to\infty} M_{S_n}(t) = \lim_{n\to\infty} \left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n = e^{\frac{t^2}{2}},$$

což je momentová funkce pro rozdělení $N(0,1)$.

CENTRÁLNÍ LIMITNÍ VĚTA

Nechť $Y_1, Y_2, \ldots, Y_n$ jsou po dvou nezávislé náhodné veličiny se společnou střední hodnotou $EY_i = \mu$ a rozptylem $\operatorname{var} Y_i = \sigma^2 > 0$ a konečným třetím absolutním momentem $E|Y_i|^3 < C$.

Pro distribuční funkce náhodných veličin

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\sigma} (Y_i - \mu)$$

platí

$$\lim_{n \to \infty} P(S_n \leq x) = \Phi(x),$$

kde $\Phi(x)$ je distribuční funkce normálního rozdělení $N(0,1)$.

<u>Aplikace</u> Vztah binomického a normálního rozdělení. Nechť $Y_i \sim A(p)$ jsou nezávislé. Pak $X_n = Y_1 + Y_2 + \ldots + Y_n \sim Bi(n,p)$. Podle centrální limitní věty má náhodná veličina

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - p}{\sqrt{p(p-1)}} = \frac{X_n - np}{\sqrt{np(1-p)}}$$

pro n velké rozložení velmi podobné $N(0,1)$

Tedy X_n má distribuční funkci blízkou distribuční funkci normálního rozdělení

$$N\left(np, \sqrt{np(1-p)}\right).$$

Toto je obsahem <u>Moivrovy - Laplaceovy věty.</u>

<u>Příklad</u>: Chceme zjistit s jakou pravděpodobností p je v populaci zastoupena krevní skupina A.
X náhodná veličina počet osob s krevní skupinou A v souboru n osob

$$X \sim Bi(n, p)$$

Chceme zjistit, pro které n bude už platit

$$P\left(\left|\frac{X}{n} - p\right| \leq 0{,}05\right) \simeq 0{,}9$$

(Chceme s 90% spolehlivostí ~~zjistit~~ najít p s chybou nejvýše 5%.)

$$0{,}9 \simeq P\left(\left|\frac{X}{n} - p\right| \leq 0{,}05\right) = P\left(\frac{X}{n} - p \leq 0{,}05\right)$$

$$- P\left(\frac{X}{n} - p < -0{,}05\right) = P\left(\frac{X - np}{\sqrt{np(p-1)}} \leq \frac{0{,}05n}{\sqrt{np(p-1)}}\right)$$

$$- P\left(\frac{X - np}{\sqrt{np(p-1)}} < \frac{-0{,}05n}{\sqrt{np(p-1)}}\right) \approx$$

$$\Phi\left(\frac{0{,}05\,n}{\sqrt{n p (p-1)}}\right) - \Phi\left(-\frac{0{,}05\,n}{\sqrt{n p (p-1)}}\right) =$$

$$= \Phi\left(\frac{0{,}05\,n}{\sqrt{n p (p-1)}}\right) - \left(1 - \Phi\left(\frac{0{,}05\,n}{\sqrt{n p (p-1)}}\right)\right) =$$

$$= 2\,\Phi\left(\frac{0{,}05\,n}{\sqrt{n p (p-1)}}\right) - 1$$

Chceme tedy

$$\Phi\left(\frac{0{,}05\,n}{\sqrt{n p (p-1)}}\right) \simeq \frac{1}{2}\,(1+0{,}9) = 0{,}95$$

Odtud

$$\frac{0{,}05\sqrt{n}}{\sqrt{\cancel{n}\, p\,(p-1)}} \geq 1{,}64485$$

Protože $p(p-1) \leq \frac{1}{4}$ na intervalu $(0,1)$,

budeme požadovat

$$\frac{0{,}05\sqrt{n}}{\sqrt{\frac{1}{4}}} = 0{,}1\sqrt{n} \geq 1{,}64485$$

$$\sqrt{n} \geq 16{,}4485$$

$$n \geq 270.$$

Vybereme-li tedy vzorek 270 lidí, určíme na něm pravděpodobnost výskytu krevní skupiny A s chybou 5% se spolehlivostí 90%.