

Matematika III – 12. přednáška

Náhodné veličiny – číselné charakteristiky, normální rozdělení, limitní věty

Michal Bulant

Masarykova univerzita
Fakulta informatiky

11. 12. 2013

Obsah přednášky

- 1 Číselné charakteristiky náhodných veličin
- 2 Limitní věty a odhady
- 3 Normální rozdělení a rozdělení odvezená
- 4 Náhodný výběr

Doporučené zdroje

- Jan Slovák, Martin Panák, Michal Bulant, *Matematika drsně a svižně*, MU Brno, 2013, 774 s. (též jako e-text).
- Karel Zvára, Josef Štěpán, **Pravděpodobnost a matematická statistika**, Matfyzpress, 4. vydání, 2006, 230 stran, ISBN 80-867-3271-1.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Teorie pravděpodobnosti a matematická statistika (sbírka příkladů)**, Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, **Statistika**, Masarykova univerzita, 2004, distanční studijní opora ESF, <http://www.math.muni.cz/~budikova/esf/Statistika.zip>.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, **Základní statistické metody**, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.

Střední hodnota transformované náhodné veličiny

Střední hodnotu můžeme přímo vyjádřit také pro funkce $Y = \psi(X)$ náhodné veličiny X . V diskrétním případě můžeme přímo spočít

$$\begin{aligned}
 E(Y) &= \sum_j y_j P(Y = y_j) \\
 &= \sum_j y_j \sum_{\psi(x_i)=y_j} P(X = x_i) \\
 &= \sum_i \psi(x_i) P(X = x_i) = \sum_i \psi(x_i) f_X(x_i).
 \end{aligned}$$

Je tedy $E(\psi(X))$ přímo spočítatelná pomocí pravděpodobnostní funkce f_X .

Podobně vyjadřujeme střední hodnotu funkce ze spojitě náhodné veličiny:

$$E(\psi(X)) = \int_{-\infty}^{\infty} \psi(x) f_X(x) dx,$$

pokud tento integrál absolutně konverguje.

Příklad

Spočtěme střední hodnotu binomického rozdělení.

Řešení

Pro $X \sim \text{Bi}(n, p)$ je

$$\begin{aligned}
 E(X) &= \sum_{k=0}^n k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \\
 &= np \sum_{k=1}^n \frac{(n-1)!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{n-k} = \\
 &= np \sum_{j=0}^{n-1} \frac{(n-1)!}{(n-1-j)!j!} p^j (1-p)^{n-1-j} = \\
 &= np(p + (1-p))^{n-1} = np.
 \end{aligned}$$

Základní vlastnosti střední hodnoty

Věta

Nechť $a, b \in \mathbb{R}$ a X, Y jsou náhodné veličiny s existující střední hodnotou. Pak

- $E(a) = a,$
- $E(a + bX) = a + bE(X),$
- $E(X + Y) = E(X) + E(Y),$
- jsou-li X a Y **nezávislé**, pak $E(XY) = E(X) \cdot E(Y).$

Důkazy těchto tvrzení jsou přímočaré, zkuste si je udělat!
Analogická tvrzení platí i pro náhodné vektory.

Příklad

Spočtíme ještě jednu střední hodnotu binomického rozdělení, tentokrát s využitím vlastností střední hodnoty.

Řešení

Vyjádříme počet zdarů v n pokusech jako počet zdarů v jednotlivých pokusech

$$X = \sum_{k=1}^n Y_k,$$

přičemž náhodné veličiny Y_k mají všechny alternativní rozdělení $A(p)$. Snadno spočítáme $E(Y_k) = 1 \cdot p + 0 \cdot (1 - p) = p$. Dále víme, že střední hodnota součtu je součtem středních hodnot, proto

$$E(X) = \sum_{k=1}^n E(Y_k) = np.$$

Kvantily

Dalšími užitečnými charakteristikami jsou tzv. **kvantily**. Pro ryze monotónní distribuční funkci F_X (tj. spojitou náhodnou veličinu X s všude nenulovou hustotou, jako je tomu např. u normálního rozdělení) jde prostě o inverzní funkci $F_X^{-1} : (0, 1) \rightarrow \mathbb{R}$. To znamená, že hodnota $y = F^{-1}(\alpha)$ je taková, že $P(X < y) = \alpha$. Obecněji, je-li $F_X(x)$ distribuční funkce náhodné veličiny X , pak definujeme **kvantilovou funkci**²

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R}; F(x) \geq \alpha\}, \quad \alpha \in (0, 1).$$

Zřejmě jde o zobecnění předchozí definice.

Nejčastěji jsou používané kvantily s $\alpha = 0.5$, tzv. **medián**, s $\alpha = 0.25$, tzv. **první kvartil**, $\alpha = 0.75$, tzv. **třetí kvartil**, a podobně pro **decily** a **percentily** (kdy je α rovno násobkům desetin a setin). K těmto hodnotám se vrátíme v popisné statistice později.

²Uvědomte si, že jsme se již s kvantily setkali, jen jsme jím tak zatím neříkali.

Rozptyl a směrodatná odchylka

Tyto číselné charakteristiky rozdělení náhodné veličiny nepopisují nějakou střední či typickou hodnotu (jako střední hodnota či medián), ale míru „kolísání“ náhodné veličiny kolem střední hodnoty.

Rozptylem (variancí) náhodné veličiny X , která má konečnou střední hodnotu, nazýváme číslo

$$D(X) = \text{var } X = E([X - E(X)]^2),$$

odmocnina z rozptylu $\sqrt{D(x)}$ se pak nazývá **směrodatná odchylka**.

Základní vlastnosti rozptylu

Věta

Pro náhodnou veličinu X a reálná čísla a, b platí:

- 1 $D(X) = E(X^2) - E(X)^2,$
- 2 $D(a + bX) = b^2 D(X),$
- 3 $\sqrt{D(a + bX)} = |b| \sqrt{D(X)}.$

Důkaz.

Důkaz je přímočarý. Poznamenejme, že tvrzení 1 se často používá k výpočtům $D(X)$. □

Kovariance

O závislosti dvou náhodných veličin do jisté míry vypovídá tzv. **kovariance**, definovaná předpisem

$$C(X, Y) = \text{cov}(X, Y) = E([X - E(X)][Y - E(Y)]).$$

Veličinám X, Y , pro něž je $C(X, Y) = 0$, říkáme **nekorelované**.

Věta

Pro náhodné veličiny s existujícími rozptyly platí:

- ① $C(X, Y) = C(Y, X)$,
- ② $C(X, X) = D(X)$,
- ③ $C(X, Y) = E(XY) - E(X)E(Y)$,
- ④ $C(a + bX, c + dY) = bd \cdot C(X, Y)$,
- ⑤ $D(X + Y) = D(X) + D(Y) + 2C(X, Y)$, *speciálně, jsou-li X, Y nezávislé, je $D(X + Y) = D(X) + D(Y)$, tj. $C(X, Y) = 0$ a X, Y jsou nekorelované.*

Koeficient korelace

Koeficient korelace je jen speciální název pro kovarianci dvou normovaných náhodných veličin:

$$R(X, Y) = \rho_{X,Y} = C \left(\frac{X - E(X)}{\sqrt{D(X)}}, \frac{Y - E(Y)}{\sqrt{D(Y)}} \right).$$

Věta

- ① $R(X, X) = 1,$
- ② $R(a + bX, c + dY) = \text{sgn}(bd)R(X, Y),$
- ③ *jsou-li* X, Y *nezávislé, je* $R(X, Y) = 0,$
- ④ *(Cauchyova nerovnost)* $|R(X, Y)| \leq 1.$

Příklad

Spočtěme rozptyl binomického rozdělení.

Řešení

Stejně jako dříve lze psát $X = \sum_{k=1}^n Y_k$, kde Y_1, \dots, Y_n jsou nezávislé náhodné veličiny vyjadřující úspěch v k -tém pokusu.

Snadno vypočteme $E(Y_k^2) = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$, proto

$D(Y_k) = E(Y_k^2) - E(Y_k)^2 = p - p^2 = p(1 - p)$. Protože pro **nezávislé** Y_k platí $D(\sum Y_k) = \sum D(Y_k)$, je $D(X) = np(1 - p)$.

Příklady k procvičení

Příklad

Náhodná veličina X je dána pravděpodobnostní funkcí

$$p(x) = \begin{cases} \frac{1}{3} & \text{pro } x = -2 \\ \frac{1}{2} & \text{pro } x = 3 \\ \frac{1}{6} & \text{pro } x = 1 \\ 0 & \text{jinak.} \end{cases}$$

Určete $E(X)$, $E(2X + 5)$, $E(X^2)$, $D(X)$ a $D(2X + 1)$.

Příklad

Nekorelované náhodné veličiny X a Y mají rozptyly $D(X) = a$ a $D(Y) = 2$. Určete konstantu a , jestliže rozptyl náhodné veličiny $Z = 3Y - X$ je $D(Z) = 25$.

Normovaná náhodná veličina a limitní věty

Všimněme si, že výraz $\frac{X_n - np}{\sqrt{np(1-p)}}$ vystupující v Moivre-Laplaceově

větě je totéž, co $\frac{X_n - E(X_n)}{\sqrt{D(X_n)}}$ a jde tedy o tzv. **normovanou**

náhodnou veličinu (tj. veličinu lineárně transformovanou tak, aby měla střední hodnotu 0 a rozptyl 1). Moivre-Laplaceova věta pak říká, že pro $n \rightarrow \infty$ se rozložení této náhodné veličiny blíží normovanému normálnímu rozdělení $N(0, 1)$.

Jde o speciální případ limitních vět, ukazujících, že za určitých podmínek platí „zákony velkých čísel“, kdy se obdobným způsobem transformované náhodné veličiny chovají jako normální rozdělení.

Další momenty

Někdy je užitečné studovat řadu dalších charakteristik rozdělení náhodných veličin. Za rozumných předpokladů jsou definovány **k -té obecné momenty**

$$\mu'_k = E(X^k)$$

a **k -té centrální momenty**

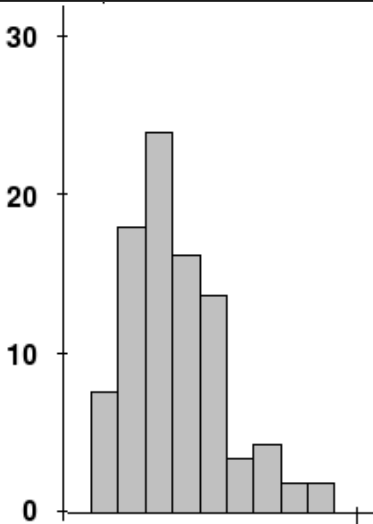
$$\mu_k = E([X - E(X)]^k).$$

Pomocí momentů pak definujeme např. **šikmost** (asymetrii) náhodné veličiny X jako

$$\frac{\mu_3}{\sqrt{D(x)}^3}$$

nebo **špičatost** (exces) jako

$$\frac{\mu_4}{D(x)^2} - 3.$$



Kladná šikmost distribuce (více vysokých kladných hodnot než odpovídá normálnímu rozdělení s nulovou šikmostí).

Momentová vytvořující funkce

Definice

Reálnou funkci proměnné $t \in \mathbb{R}$ $M_X(t) = E(e^{tX})$ nazveme **momentovou vytvořující funkcí** náhodné veličiny X .

Poznámka

Je-li X např. spojitá, platí

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f(x) dx = \\ &= \int_{-\infty}^{\infty} \left(1 + tx + \frac{t^2 x^2}{2!} + \dots\right) f(x) dx = \\ &= 1 + t\mu'_1 + \frac{t^2 \mu'_2}{2!} + \dots \end{aligned}$$

a jde vlastně o *exponenciální vytvořující funkci* posloupnosti k -tých obecných momentů μ'_k .

Věta

Pro momentovou vytvořující funkci platí:

- $\mu'_k = \frac{d^k}{dt^k} M_X(t) |_{t=0}$.
- *Platí-li $M_X(t) = M_Y(t)$ pro všechna $t \in \langle -b, b \rangle$, mají náhodné veličiny stejné rozdělení, tj. $F_X(x) = F_Y(x)$.*
- $M_{a+bX}(t) = e^{at} M_X(bt)$.
- *Jsou-li X, Y nezávislé, je $M_{X+Y}(t) = M_X(t)M_Y(t)$.*

Příklad

Určete momentovou vytvořující funkci binomického rozdělení.

Řešení

$$\begin{aligned}M(t) &= E(e^{tX}) = \sum_{k=0}^n e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = \\&= \sum_{k=0}^n \binom{n}{k} (pe^t)^k (1-p)^{n-k} = \\&= (pe^t + (1-p))^n = (p(e^t - 1) + 1)^n.\end{aligned}$$

Snáze jsme mohli funkci určit s využitím předchozích vět a momentové vytvořující funkce alternativního rozdělení, neboť pro $Y \sim A(p)$ je $E(e^{tY}) = e^{t \cdot 1} \cdot p + e^{t \cdot 0}(1-p) = p(e^t - 1) + 1$.

Příklad

Naposled spočtěme střední hodnotu a rozptyl binomického rozdělení, tentokrát s využitím vytvořující funkce.

Řešení

$M(t) = (p(e^t - 1) + 1)^n$, proto je

$$\frac{d}{dt}M(t) = n(p(e^t - 1) + 1)^{n-1}e^t p,$$

což pro $t = 0$ dá $E(X) = \mu'_1 = np$.

Podobně spočítáme i $D(x) = \mu'_2 - (\mu'_1)^2$.

Příklady k procvičení

Příklad

Náhodná veličina X má na intervalu $(0, a)$ konstantní hustotu pravděpodobnosti (a jinde nulovou). S využitím vlastností střední hodnoty a rozptylu určete:

- 1 $E(2X + 3)$,
- 2 $E(3X^2 - 2X + 1)$,
- 3 $D(2X + 3)$,
- 4 $D(X^2 + 1)$,

Příklad

Náhodná veličina X má Poissonovo rozdělení (tj. pravděpodobnostní funkci $p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$). Určete její (momentovou vytvořující funkci,) střední hodnotu a rozptyl.

Číselné charakteristiky náhodných vektorů

$E(X) = (E(X_1), \dots, E(X_n))$ se nazývá vektor středních hodnot,

$$\text{var}(X) = \begin{pmatrix} D(X_1) & C(X_1, X_2) & \cdots & C(X_1, X_n) \\ \dots & \dots & \dots & \dots \\ C(X_n, X_1) & C(X_n, X_2) & \cdots & D(X_n) \end{pmatrix}$$

varianční (rozptylová) matice a

$$\text{cor } X = \begin{pmatrix} 1 & R(X_1, X_2) & \cdots & R(X_1, X_n) \\ \dots & \dots & \dots & \dots \\ R(X_n, X_1) & R(X_n, X_2) & \cdots & 1 \end{pmatrix}$$

je korelační matice.

Snadno je po rozepsání po jednotlivých složkách vidět, že

$$\text{var}(X) = E((X - E(X)) \cdot (X - E(X))^T)$$

Ukážeme na příkladech, že pravděpodobnostní struktura náhodného vektoru (X, Y) není určena pouze marginálními rozděleními veličin X a Y . Podstatný je rovněž pravděpodobnostní vztah mezi X a Y , který je částečně popsán např. prostřednictvím korelačního koeficientu.

Příklad

Jsou-li X a Y náhodné veličiny, nabývající hodnot 0 a 1, pak

$$P(X = 1, Y = 1) - P(X = 1)P(Y = 1) = E(XY) - E(X)E(Y) = C(X, Y).$$

Odtud je snadno vidět, že pokud jsou X a Y nekorelované, jsou i nezávislé (což obecně neplatí).

Uveďme ještě příklad, ilustrující, že nekorelovanost nemusí implikovat nezávislost:

Příklad

Buďte A a X nezávislé náhodné veličiny, splňující $X \sim N(0, 1)$ a $P(A = 1) = P(A = -1) = 1/2$. Položíme-li $Y = AX$, pak

$$P(Y < y) = \frac{1}{2}P(X < y) + \frac{1}{2}P(-X < y) = \Phi(y),$$

proto má rovněž Y rozdělení $N(0, 1)$.

Dále $C(X, Y) = E(XY) - E(X)E(Y) = E(AX^2) = E(A)E(X^2) = 0 \cdot 1 = 0$, přitom $P(X = Y) = P(X = -Y) = 1/2$ a X, Y zřejmě nejsou nezávislé.

Příklad

Nechť (X, Y) je náhodný vektor, který má rovnoměrné rozdělení na jednotkovém kruhu $K = \{(x, y) : x^2 + y^2 \leq 1\}$. Zřejmě je hustota tohoto rozdělení rovna $1/\pi$ pro $(x, y) \in K$ a 0 jinde a je rovněž vidět, že X, Y **nejsou nezávislé**. Označme $R = R(X, Y)$ a $\Phi = \Phi(X, Y)$ polární souřadnice náhodného vektoru (X, Y) a určíme rozdělení vektoru (R, Φ) .

Pro $0 < r_1 \leq 1$ a $0 < \varphi_1 \leq 2\pi$ je

$$\begin{aligned} P(R < r_1, \Phi \leq \varphi_1) &= \frac{1}{\pi} \pi r_1^2 \frac{\varphi_1}{2\pi} = \\ &= \int_0^{r_1} \int_0^{\varphi_1} \frac{1}{2\pi} 2r d\varphi dr. \end{aligned}$$

Hustota je tedy rovna $f(r, \varphi) = \frac{r}{\pi}$ pro $0 < r \leq 1$, $0 < \varphi \leq 2\pi$ a rovna 0 všude jinde.

Příklad (pokr.)

Marginální hustoty $g(r)$ a $h(\varphi)$ veličin R a Φ se nyní snadno dopočtou:

$$g(r) = \int_{-\infty}^{\infty} f(r, \varphi) d\varphi = \int_0^{2\pi} \frac{r}{\pi} d\varphi = 2r$$

$$h(\varphi) = \int_{-\infty}^{\infty} f(r, \varphi) dr = \int_0^1 \frac{r}{\pi} dr = \frac{1}{2\pi}.$$

Veličina Φ má rovnoměrné rozdělení na $(0, 2\pi)$, odkud $E(\Phi) = \pi$ a $D(\Phi) = \pi^2/3$, snadno rovněž odvodíme $E(R) = 2/3$, $D(R) = 1/18$.

Všimněme si ale zejména, že $f(r, \varphi) = g(r)h(\varphi)$, což znamená **nezávislost veličin R a Φ** .

Vlastnosti charakteristik náhodného vektoru

Věta

Pro náhodné vektory X, Y stejné dimenze, konstantní matici B a konstantní vektor a (odpovídajících dimenzí) platí

- $E(X + Y) = E(X) + E(Y)$,
- $E(a + BX) = a + B \cdot E(X)$,
- $\text{var}(a + B \cdot X) = B \text{var}(X) B^T$.

Důkaz.

Důkaz vyplývá z vlastností náhodných veličin a ze vztahu $\text{var}(X) = E((X - E(X))(X - E(X))^T)$. □

Motivace

S jedním případem limitní věty jsme se již setkali – de Moivre-Laplaceova věta říká, že binomické rozdělení $Bi(n, p)$ lze za určitých podmínek aproximovat normovaným normálním rozdělením. Obvykle se k aproximaci přistupuje při splnění podmínek $np(1 - p) > 9$ a $\frac{1}{n+1} < p < \frac{n}{n+1}$. V této kapitole zformulujeme zobecnění této věty a rovněž další tvrzení umožňující odhadovat chování náhodných veličin při velkém počtu nezávislých opakování náhodného pokusu.

Čebyševova nerovnost

Věta

Pro **každou** náhodnou veličinu X a libovolné $\epsilon > 0$ platí

$$P(|X - EX| \geq \epsilon) \leq \frac{DX}{\epsilon^2}.$$

Důkaz.

Budeme odhadovat rozptyl DX ve spojitém případě (diskrétní analogicky):

$$\begin{aligned} DX &= \int_{-\infty}^{\infty} (X - EX)^2 f(x) dx \geq \int_{|x-EX| \geq \epsilon} (X - EX)^2 f(x) dx \geq \\ &\geq \int_{|x-EX| \geq \epsilon} \epsilon^2 f(x) dx = \epsilon^2 P(|X - EX| \geq \epsilon). \end{aligned}$$



Pomocí Čebyševovy nerovnosti můžeme odhadovat pravděpodobnost, s jakou se náhodná veličina s neznámým rozdělením odchýlí od své střední hodnoty o více než k -násobek směrodatné odchylky (zřejmě je totiž $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$).

Příklad

Nechť je $E(X) = \mu, D(X) = \sigma^2$.

- 1 Odhadněte $P(|X - \mu| \geq 3\sigma)$.
- 2 Vypočtete $P(|X - \mu| \geq 3\sigma)$, jestliže navíc víte, že $X \sim N(0, 1)$.

Řešení

- 1 $1/9$,
- 2 $0,0027$.

Speciálním případem je Bernoulliho věta, která říká, že je-li $Y_n \sim \text{Bi}(n, p)$, pak posloupnost relativních četností Y_n/n konverguje podle pravděpodobnosti k p .

Věta (Bernoulliho)

Pro náhodnou veličinu s binomickým rozdělením $Y_n \sim \text{Bi}(n, p)$ a pro libovolné $\epsilon > 0$ platí

$$P\left(\left|\frac{Y_n}{n} - p\right| > \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}.$$

Důkaz.

Plyne snadno z Čebyševovy nerovnosti, neboť

$$E(Y_n/n) = np/n = p \text{ a}$$

$$D(Y_n/n) = np(1-p)/n^2 = p(1-p)/n.$$



Příklad

Při zkoušce bylo zjištěno, že mezi 600 kontrolovanými studenty je 5 studentů, kteří neumí ani malou násobilku. Odhadněte pravděpodobnost, že relativní četnost takových studentů se od jejich pravděpodobnosti výskytu liší o více než 0,01? (Můžete předpokládat, že pravděpodobnost výskytu studenta bez znalosti násobilky je menší než 0,02).

Centrální limitní věta

Centrální limitní věta dává odpověď na otázku, proč je normální rozdělení nejdůležitějším rozdělením. Ukazuje totiž, že rozdělení součtu dostatečně velkého počtu nezávislých a stejně rozdělených náhodných veličin lze aproximovat normálním rozdělením.

Věta

*Nechť je Y_1, Y_2, \dots posloupnost **nezávislých stejně rozdělených náhodných veličin se střední hodnotou μ a rozptylem σ^2 . Pak pro normované náhodné veličiny***

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

platí

$$\lim_{n \rightarrow \infty} P(S_n < t) = \Phi(t),$$

kde Φ je distribuční funkce rozdělení $N(0, 1)$.

Příklad

Mezi učiteli matematiky v ČR je jich 10% s příjmem přesahujícím celostátní průměr. Kolik matematiků je třeba pozvat na konferenci, aby s pravděpodobností aspoň 0,95 mezi nimi bylo 8 až 12 procent s nadprůměrným příjmem?

Řešení

$Y_n \sim \text{Bi}(n; 0,1)$, $E(Y_n) = 0,1 \cdot n$, $D(Y_n) = 0,1 \cdot 0,9 \cdot n$. Pak

$$0,95 \leq P(0,08n \leq Y_n \leq 0,12n) =$$

$$= P\left(\frac{0,08 - 0,1}{\sqrt{0,09n}} n \leq \frac{Y_n - 0,1n}{\sqrt{0,09n}} \leq \frac{0,12 - 0,1}{\sqrt{0,09n}} n\right) =$$

$$= P\left(\frac{-\sqrt{n}}{15} \leq \frac{Y_n - 0,1n}{\sqrt{0,09n}} \leq \frac{\sqrt{n}}{15}\right) \approx \Phi\left(\frac{\sqrt{n}}{15}\right) - \Phi\left(-\frac{\sqrt{n}}{15}\right).$$

Je tedy $\Phi\left(\frac{\sqrt{n}}{15}\right) \geq 0,975$, což je ekvivalentní $\sqrt{n}/15 \geq 1,96$, tj.
 $n \geq 865$.

Řešení (Pomocí Čebyševovy/Bernoulliovy nerovnosti)

Nyní využijme Bernoulliovu nerovnost – ta dává

$$P\left(\left|\frac{Y_n}{n} - 0,1\right| \leq 0,02\right) \geq 1 - \frac{0,1 \cdot 0,9}{n \cdot 0,02^2},$$

což má být alespoň 0,95. Odtud

$$n \geq \frac{0,09}{0,05 \cdot 0,02^2} = 4500.$$

Opět vidíme, že odhad prostřednictvím Bernoulliovy nerovnosti je podstatně slabší než odhad s využitím centrální limitní věty (resp. de Moivre-Laplaceovy věty).

Normální rozdělení

Příklad

Dokažte, že funkce

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

je hustotou (tzv. normované/standardizované normálně rozdělené náhodné veličiny), tj. že platí

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

Řešení

Tento příklad je možné řešit několika způsoby, my (neformálně) naznačíme postup pomocí integrace funkce dvou proměnných přechodem k polárním souřadnicím. Dvěma způsoby tak spočítáme

$$\iint_A e^{-(x^2+y^2)} dA \quad \text{pro} \quad A = \mathbb{R}^2,$$

odkud porovnáním dostaneme

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}.$$

Momenty normálního rozdělení

Přímý výpočet střední hodnoty a rozptylu normovaného normálního rozdělení není triviální. S využitím momentové vytvořující funkce je ale poměrně jednoduchý.

Nechť $Z \sim N(0, 1)$. Pak

$$\begin{aligned}
 M_Z(t) &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2 - 2tz + t^2 - t^2}{2}\right) dz = \\
 &= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-t)^2}{2}\right) dz = \exp\left(\frac{t^2}{2}\right).
 \end{aligned}$$

Poslední integrál je roven 1 díky tomu, že na místě integrované funkce je funkce s vlastnostmi hustoty.

Střední hodnota a rozptyl normálního rozdělení

S využitím předchozího výpočtu $M_Z(t) = \exp\left(\frac{t^2}{2}\right)$ snadno spočítáme, že

$$M'_Z(t) = t \exp\left(\frac{t^2}{2}\right),$$

$$M''_Z(t) = t^2 \exp\left(\frac{t^2}{2}\right) + \exp\left(\frac{t^2}{2}\right).$$

Dosazením $t = 0$ pak dostaneme

$$E(Z) = 0, D(Z) = 1.$$

Pro transformovanou náhodnou veličinu $Y = \mu + \sigma Z \sim N(\mu, \sigma^2)$ pak snadno odvodíme z vlastností střední hodnoty, resp. rozptylu, že $E(Y) = \mu, D(Y) = \sigma^2$ (což zpětně zdůvodňuje zápis $N(\mu, \sigma^2)$).
Momentová vytvořující funkce pro Y má tvar

$$M_Y(t) = \exp\left(\mu t + \sigma^2 \frac{t^2}{2}\right).$$

Příklad

Určete rozdělení součtu nezávislých náhodných veličin

$$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2).$$

Řešení

Z vlastností momentové vytvořující funkce dostáváme

$$\begin{aligned} M_{X+Y}(t) &= \exp(\mu_X t + \sigma_X^2 \frac{t^2}{2}) \exp(\mu_Y t + \sigma_Y^2 \frac{t^2}{2}) = \\ &= \exp((\mu_X + \mu_Y)t + (\sigma_X^2 + \sigma_Y^2) \frac{t^2}{2}). \end{aligned}$$

Proto $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Γ (gamma) rozdělení

Γ rozdělení se často používá u modelů čekání (např. v pojistné matematice je čas dožití často modelován pomocí gamma rozdělení).

Příklad

Určete konstantu c tak, aby funkce $cx^{a-1}e^{-bx}$ pro $x > 0$ a nulová jinde ($a, b > 0$ jsou parametry) byla hustotou náhodné veličiny.

Řešení

Hustota musí splňovat

$$\begin{aligned} 1 &= \int_0^{\infty} cx^{a-1}e^{-bx} dx = \int_0^{\infty} c\left(\frac{t}{b}\right)^{a-1} e^{-t} \frac{1}{b} dt = \\ &= \frac{c}{b^a} \int_0^{\infty} t^{a-1} e^{-t} dt = \frac{c}{b^a} \Gamma(a), \end{aligned}$$

proto $c = \frac{b^a}{\Gamma(a)}$.

Poznámka

Funkce Γ je zobecnění faktoriálu ($\Gamma(n) = (n-1)!$ pro $n \in \mathbb{N}$), definované předpisem $\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx$. Často počítáme hodnoty této funkce s využitím vlastností

$$\Gamma(1/2) = \sqrt{\pi}, \quad \Gamma(a+1) = a \cdot \Gamma(a).$$

Definice

Rozdělení náhodné veličiny s hustotou

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

spočítanou v předchozím příkladu nazýváme **gamma rozdělení** s parametry a, b a značíme $\Gamma(a, b)$. Momentová vytvořující funkce je pak $M(t) = \left(\frac{b}{b-t}\right)^a$, střední hodnota $E(X) = a/b$ a rozptyl $D(X) = a/b^2$.

Příklad (rozdělení χ^2 podruhé)

Nechť Z má normované normální rozdělení. Určete hustotu transformované náhodné veličiny $X = Z^2$.

Řešení

Již dříve jsme vypočetli přímým výpočtem přes distribuční funkci, že hustota

$$f_X(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$$

a řekli jsme si, že jde o (Pearsonovo) χ^2 rozdělení s jedním stupněm volnosti, které značíme $X \sim \chi^2(1)$. Nyní vidíme, že jde o speciální případ Γ -rozdělení, totiž $\Gamma(1/2, 1/2)$.

Obecně pro součet Y čtverců n nezávislých náhodných veličin s rozdělením $N(0, 1)$ obdobně odvodíme, že má rozdělení $\Gamma(n/2, 1/2)$ a říkáme, že Y má rozdělení $\chi^2(n)$ (*chí kvadrát s n stupni volnosti*). Toto rozdělení se ve statistice používá velmi často.

Další důležitá rozdělení

F-rozdělení

Jsou-li X, Y nezávislé náhodné veličiny s rozděleními

$X \sim \chi^2(k), Y \sim \chi^2(m)$, pak má transformovaná náhodná veličina

$$U = \frac{X/k}{Y/m}$$

takzvané Fisher-Snedecorovo F-rozdělení $F(k, m)$ s k a m stupni volnosti.

Studentovo t-rozdělení

Jsou-li $Z \sim N(0, 1)$ a $X \sim \chi^2(n)$ nezávislé náhodné veličiny, pak má veličina

$$T = \frac{Z}{\sqrt{X/n}}$$

tzv. Studentovo t-rozdělení $t(n)$ s n stupni volnosti.

Přehled rozdělení odvozených od normálního

- $Z_1, \dots, Z_k \sim N(0, 1)$ **nezávislá** normovaná normální
- $\chi^2_k = \sum_{i=1}^k Z_i^2 \sim \chi^2(k)$ chí-kvadrát o k stupních volnosti
- $F_{k,m} = \frac{X_k^2/k}{X_m^2/m} \sim F(k, m)$. . . F-rozdělení s k a m stupni volnosti
- $T_k = \frac{Z}{\sqrt{X_k^2/k}} \sim t(k)$ t-rozdělení s k stupni volnosti

Odtud zejména $Z^2 \sim \chi^2(1)$ a $T_k^2 \sim F(1, k)$.

rozdělení	střední hodnota	rozptyl
$N(\mu, \sigma^2)$	μ	σ^2
$\chi^2(k)$	k	$2k$
$t(k)$	0	$k/(k - 2)$
$F(k, m)$	$m/(m - 2)$	$2m^2(k + m - 2)/k(m - 2)^2(m - 4)$

Náhodný výběr

Definice

Náhodným výběrem rozsahu n rozumíme n -tici **nezávislých** a **stejně rozdělených** náhodných veličin $X_1, \dots, X_n \sim F_X(x)$ (někdy také hovoříme o n nezávislých kopiích náhodné veličiny X).

Náhodným výběrem rozsahu n z p -rozměrného rozdělení rozumíme n -tici **nezávislých** a **stejně rozdělených** p -rozměrných náhodných vektorů.

V matematické statistice často pracujeme s transformacemi náhodného výběru, takovým náhodným veličinám (příp. vektorům) říkáme **statistiky**. V následujícím zavedeme několik důležitých statistik a ukážeme jejich souvislost s číselnými charakteristikami náhodných veličin.

Základní statistiky

Definice

Nechť X_1, \dots, X_n je náhodný výběr. Statistiku

$$M = \frac{1}{n} \sum_{i=1}^n X_i$$

nazýváme **výběrový průměr**, statistiku

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M)^2$$

výběrový rozptyl a statistiku $S = \sqrt{S^2}$ **výběrová směrodatná odchylka**. Analogicky se definují i výběrová kovariance, příp. výběrový korelační koeficient pro dvourozměrný náhodný výběr.

Vlastnosti statistik

Protože jsou uvedené statistiky náhodnými veličinami, lze se přirozeně ptát po jejich číselných charakteristikách.

Věta

Nechť X_1, \dots, X_n je náhodný výběr rozsahu n z rozdělení se střední hodnotou μ a rozptylem σ^2 . Pak platí:

- $E(M) = \mu,$
- $D(M) = \text{var}(M) = \sigma^2/n,$
- $E(S^2) = \sigma^2.$

Důkaz.

Ukážeme jen (nejsložitější) 3. tvrzení.

Snadno se odvodí, že platí

$$\sum (X_i - \mu)^2 = \sum (X_i - M)^2 + n(M - \mu)^2.$$

Proto je

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} E\left(\sum (X_i - \mu)^2\right) - \frac{n}{n-1} E(M - \mu)^2 = \\ &= \frac{1}{n-1} \sum D(X_i) - \frac{n}{n-1} D(M) = \\ &= \frac{n}{n-1} \sigma^2 - \frac{1}{n-1} \sigma^2 = \sigma^2. \end{aligned}$$



V předchozí větě jsme ukázali, že výběrový průměr M splňuje $E(M) = \mu$, jeho střední hodnota je tedy rovna odhadovanému parametru μ . V takovém případě říkáme, že statistika M je **nestranným odhadem** parametru μ .

Podobně jsme viděli, že S^2 je nestranným odhadem parametru σ^2 .

Všimněme si rovněž, že „přirozeněji“ definovaná statistika $\frac{1}{n} \sum (X_i - M)^2$ není nestranným odhadem σ^2 , její střední hodnota je totiž $\frac{n-1}{n} \sigma^2$.

Příklad

Rozmyslete si, je-li S nestranným odhadem směrodatné odchylky σ .

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je X_1, \dots, X_n náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- M a S^2 jsou nezávislé náhodné veličiny.
- $M \sim N(\mu, \sigma^2/n)$, a tedy $U = (M - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.
- $T = (M - \mu)/(S/\sqrt{n}) \sim t(n - 1)$.
- $K = (n - 1)S^2/\sigma^2 \sim \chi^2(n - 1)$.
- $\sum(X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.

Poznámka

K odhadu μ , neznáme-li σ^2 , slouží T , v opačném případě U .
K odhadu σ^2 , neznáme-li μ , slouží K , v opačném případě následující (bezejmenná?) statistika, která je vlastně statistikou K , v níž místo odhadu M použijeme přímo μ .

Příklad

V roce 1951 bylo rozsáhlým statistickým průzkumem zjištěno, že střední hodnota výšky desetiletých chlapců je 136,1 cm se směrodatnou odchylkou $\sigma = 6,4$ cm.

V roce 1961 byla zjištěna výška pouze u 15 náhodně vybraných chlapců:

130	140	136	141	139	133	149	151
139	136	138	142	127	139	147	

Otázkou je, zda se v porovnání s rokem 1951 změnila střední výška chlapců, pokud předpokládáme, že variabilita výšek se v různých generacích příliš nemění.

Řešení

Vzhledem k tomu, že základní soubor všech desetiletých chlapců je rozsáhlý, lze zmíněná data považovat za náhodný výběr^a. Zjistíme, že výběrový průměr $M = 139,133$, $n = 15$ a s využitím statistiky U dostáváme, že s 95% pravděpodobností leží hodnota μ v intervalu

$$(M - 1,96\sigma/\sqrt{n}; M + 1,96\sigma/\sqrt{n}) = (135,9; 142,4).$$

Protože i střední hodnota výšek z roku 1951 leží v tomto intervalu, nemáme vážný důvod tvrdit, že se střední výška změnila. Pokud bychom ovšem připustili vyšší možnost omylu a stanovili interval se spolehlivostí pouze 90%, pak bychom na této hladině hypotézu, že se střední výška změnila, „přijali“ – interval je nyní $(136,41; 141,85)$. Podobně, pokud nás zajímá pouze **dolní odhad** střední hodnoty výšek chlapců (a vůbec tedy nepřipouštíme možnost, že by se střední výška snížila), pak s 95% pravděpodobností je střední výška větší než 136,41, a tedy nyní opět „přijímáme“ hypotézu, že se střední výška zvýšila.

Příklady k procvičení

Příklad

Předpokládejme, že velká skupina studentů má ze zápočtové písemky ze statistiky bodové hodnoty normálně rozloženy kolem střední hodnoty 72 se směrodatnou odchylkou 9 bodů. Určete pravděpodobnost, že

- a) náhodně vybraný student bude mít výsledek lepší než 80 bodů,
- b) průměr výsledků náhodného výběru 10 studentů bude lepší než 80 bodů.

Příklad

Rychlost letadla byla určována v 5 zkouškách, jejichž aritmetický průměr byl $m = 870,3 \text{ ms}^{-1}$. Najděte 95% interval spolehlivosti pro μ víte-li, že měření rychlosti se řídí normálním rozdělením se směrodatnou odchylkou $2,1 \text{ ms}^{-1}$.