# Matematika III – 10. týden
# Číselné charakteristiky – střední hodnota, rozptyl, kovariance, korelace

Jan Slovák

Masarykova univerzita
Fakulta informatiky

17.–21. 11. 2014

# Obsah přednášky

# Kde je dobré číst?


- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.
- J. Slovák, M. Panák, M. Bulant, Matematika drsně a svižně, Muni Press, Brno 2013, v+773 s., elektronická edice `www.math.muni.cz/Matematika_drsne_svizne`
- 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, Tomáš Lerch, Štěpán Mikoláš, Základní statistické metody, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.
- Riley, K.F., Hobson, M.P., Bence, S.J. Mathematical Methods for Physics and Engineering, second edition, Cambridge University Press, Cambridge 2004, ISBN 0 521 89067 5, xxiii + 1232 pp.

# Střední hodnota

Nechť $X$ je náhodná veličina s diskrétním rozdělením. Jestliže řada $\sum_{k=1}^{\infty} x_i P(X = x_i)$ konverguje absolutně (zejména tedy pro všechny $X$ s konečně mnoha možnými hodnotami $x_i$), pak její součet $\mathrm{E}\, X$ nazýváme **střední hodnotou** $X$.
Je-li $X$ náhodná veličina se spojitým rozdělením s hustotou $f(x)$ a nevlastní integrál $\int_{-\infty}^{\infty} xf(x)dx$ konverguje absolutně, pak jeho hodnota $\mathrm{E}\, X$ se nazývá **střední hodnota** $X$.
Je tedy $\mathrm{E}\, X = np$, je-li $X \sim \mathrm{Bi}(n, p)$, zatímco pro rovnoměrné rozdělení na intervalu $(a, b)$ dostaneme dle očekávání

$$\mathrm{E}\, X = \int_a^b \frac{x}{b-a} dx = \frac{1}{2}\frac{b^2 - a^2}{b-a} = \frac{1}{2}(a+b).$$

# Vlastnosti střední hodnoty

**Theorem**

*Uvažme náhodné veličiny $X$, $Y$, skaláry $a$, $b \in \mathbb{R}$, náhodný vektor $W = (X_1, \ldots, X_n)$ a čtvercovou skalární matici $B$ s $n$ řádky.*

- *Pro konstantní náhodnou veličinu $X = a \in \mathbb{R}$ je $\mathrm{E}\, a = a$.*
- $\mathrm{E}(a + bX) = a + b\,\mathrm{E}\, X$.
- $\mathrm{E}(X + Y) = \mathrm{E}\, X + \mathrm{E}\, Y$.
- $\mathrm{E}(a + BX) = a + B(\mathrm{E}\, X)$.

**Theorem**

*Jsou-li veličiny $X$ a $Y$ nezávislé, pak $\mathrm{E}(XY) = \mathrm{E}\, X\, \mathrm{E}\, Y$.*

# Rozptyl

Další charakteristika popisuje, jak moc se dá čekat, že se hodnoty náhodné veličiny „hemží“ kolem nějaké hodnoty.

## Definition

Nechť $X$ je náhodná veličina s konečnou střední hodnotou. Pak definujeme **rozptyl** veličiny $X$ výrazem

$$\operatorname{var} X = \mathrm{E}(X - \mathrm{E}\, X)^2,$$

pokud taková konečná hodnota existuje.
Odmocnina z rozptylu $\sqrt{\operatorname{var} X}$ se nazývá **směrodatná odchylka** náhodné veličiny $X$.

Jde o zjevnou obdobu definice kvadrátu vzdálenosti vektorů nebo funkcí. Zachycujeme tak „očekávanou vzdálenost“ hodnot $X$ od její střední hodnoty.

**Theorem**

*Jestliže má náhodná veličina $X$ konečný rozptyl, pro libovolné skaláry $a$, $b \in \mathbb{R}$ platí*

- $\operatorname{var} X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2$
- $\operatorname{var}(a + bX) = b^2 \operatorname{var} X$
- $\sqrt{\operatorname{var}(a + bX)} = |b| \sqrt{\operatorname{var} X}$.

Občas přiřazujeme k $X$ **normovanou** veličinu $Z$,

$$Z = \frac{X - \mathrm{E}\, X}{\sqrt{\operatorname{var} X}},$$

která má zjevně nulovou střední hodnotu a jednotkový rozptyl.

Normální rozdělení $Z$ má hustotu $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ distribuční funkci $\Phi(z) = \int_{-\infty}^{z} \varphi(t)dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dt$.
Náhodná veličina $Y = \mu + \sigma Z$, $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$ má distribuční funkci

$$F_Y(y) = \int_{-\infty}^{\frac{y-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

$$\{\text{substituce } x = \mu + \sigma z\}$$

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

Takové rozdělení je *normální*, píšeme $Y \sim N(\mu, \sigma^2)$.
Parametry odpovídají střední hodnotě a rozptylu.

Uvažme $Z \sim N(0,1)$ a podívejme se na náhodnou veličinu $X = Z^2$.

$$\begin{aligned} F_X(x) &= P[Z^2 < x] \\ &= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \int_0^x \frac{1}{\sqrt{2\pi}} t^{-1/2} e^{-t/2} dt \end{aligned}$$

s hustotou

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

Říkáme mu rozdělení $\chi^2$, píšeme $X \sim \chi^2(1)$.

# kvantilová funkce

Je-li $F(x)$ distibuční funkce náhodné veličiny $X$, pak

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

je kvantilová funkce náhodné veličiny $X$.
Hodnota $F^{-1}(\alpha)$ se nazývá $\alpha$-kvantil.
Tzv. kritické hodnoty pro veličinu $X$ jsou pak $F^{-1}(1-\alpha)$.

# Čebyševova nerovnost

**Theorem**

*Má-li $X$ rozptyl a $\epsilon > 0$ je libovolné, pak platí*

$$P(|X - \mathrm{E}\, X| \geq \epsilon) \leq \frac{\mathrm{var}\, X}{\epsilon^2}.$$

# Kovariance veličin

Jsou-li $X$ a $Y$ dvě náhodné veličiny, pro které existují jejich konečné royptyly, pak definijeme jejich **kovarianci** vztahem

$$\operatorname{cov}(X, Y) = \mathrm{E}(X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y).$$

Evidentně je $\operatorname{cov}(X, X) = \operatorname{var} X$ a $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$.

**Theorem**

*Nechť existují konečné rozptyly veličin $X$ a $Y$. Pak*

- $\operatorname{cov}(X, Y) = \mathrm{E}(XY) - (\mathrm{E}\,X)(\mathrm{E}\,Y)$
- *pro jakékoliv skaláry $a$, $b$, $c$, $d$ platí* $\operatorname{cov}(a + bX, c + dY) = bd \operatorname{cov}(X, Y)$
- $\operatorname{var}(X + Y) = \operatorname{var} X + \operatorname{var} Y + 2 \operatorname{cov}(X, Y)$.

Od kovariance snadno odvodíme tzv. **korelační koeficient** dvou náhodných veličin $X$ a $Y$. Definujeme jej jako kovarianci příslušných normovaných veličin:

$$\rho_{X,Y} = \operatorname{cov}\left(\frac{X - \mathrm{E}\,X}{\sqrt{\operatorname{var} X}}, \frac{Y - \mathrm{E}\,Y}{\sqrt{\operatorname{var} Y}}\right) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \operatorname{varY}}}.$$

**Theorem**

- $\rho_{a+bX,c+dY} = \operatorname{sign}(bd)\rho_{X,Y}$, *pro* $bd \neq 0$
- $\rho_{X,X} = 1$
- $\rho_{X,Y} = 0$, *pokud jsou veličiny* $X$ *a* $Y$ *nezávislé.*
- *pokud je* $\rho_{X,Y}$ *definován, pak je roven jedné právě, když existují konstanty* $a$, $b$, $c$ *tak, že* $P(aX + bY = c) = 1$.

# Varianční matice

Uvažme náhodný vektor $W = (X_1, \ldots, X_n)$ takový, že pro všechny jeho komponenty existuje rozptyl. Pak **varianční matice** $\operatorname{var} W$ je dána

$$\operatorname{var} W = \begin{pmatrix} \operatorname{var} X_1 & \operatorname{cov}(X_1, X_2) & \ldots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{var} X_2 & \ldots & \operatorname{cov}(X_2, X_n) \\ & & \ldots & \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \ldots & \operatorname{var} X_n \end{pmatrix}.$$

**Theorem**

*Pro náhodný vektor $X$, skaláry $a$, matice skalárů $B$ platí*

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