

Příklady otázek – Pravděpodobnost a statistika

Reálné otázky ze SZZ 2023–2026 (M5)

Jakub Klesa

Pokryté předměty: NDMI002 (Diskrétní matematika), NMAI059 (Pravděpodobnost a statistika 1)

Obsah

1	Pravděpodobnost	1
2	Centrální limitní věta	2

Poznámka. Tento dokument obsahuje reálné otázky z minulých termínů SZZ (2023–2026) zařazené do okruhu M5. U otázek s obrázkem je grafika oříznutá z originálního PDF (viz odkaz na zdroj). Text, číselná data a tabulky jsou přepsané věrně. Zadání i oficiální náčrty řešení pocházejí ze sad zveřejněných na webu MFF UK:

<https://www.mff.cuni.cz/cs/studenti/bakalarske-studium/statni-zaverecne-zkousky/bakalarske-statni-zkousky-studijniho-programu-informatika>

Náčrty označené jako doplněné jsou vlastní, oficiální sada je k dané otázce neobsahuje. Vlastní doplnění k oficiálním náčrtům jsou uvedena v odděleném bloku *Doplňující vysvětlení*.

1 Pravděpodobnost

(Reálná otázka: léto 2023, č. 4 – Pravděpodobnost)

Zadání.

1. Napište větu o linearitě střední hodnoty (nedokazujte ji).
2. Ve skupině n lidí každý hodí (férovou) šestistěnnou kostkou. Každá dvojice lidí, kteří hodili stejné číslo, získá jeden bod. Označme Y počet bodů získaných jednou předem vybranou dvojicí. Určete střední hodnotu veličiny Y .
3. Označme X celkový počet získaných bodů za všechny dvojice (jeden člověk může skórovat ve více dvojicích). Určete střední hodnotu veličiny X .

Náčrt řešení.

(V PDF originálu bez náčrtu; následující řešení doplněno, numericky ověřeno simulací.)

1. *Linearita střední hodnoty.* Pro libovolné náhodné veličiny $X_1, \dots, X_n$ definované na též pravděpodobnostním prostoru a konstanty $a_1, \dots, a_n, b \in \mathbb{R}$ platí

$$E\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i E(X_i) + b.$$

Speciálně $E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n)$ bez ohledu na nezávislost jednotlivých veličin.

2. Veličina Y nabývá hodnot 0 nebo 1 – je to indikátor jevu „vybraná dvojice hodila stejné číslo“. Z 36 stejně pravděpodobných uspořádaných výsledků dvou hodů kostkou má 6 shodu (oba 1, oba 2, ..., oba 6), tedy

$$P(Y = 1) = \frac{6}{36} = \frac{1}{6}, \quad E(Y) = 1 \cdot P(Y = 1) + 0 \cdot P(Y = 0) = \boxed{\frac{1}{6}}.$$

3. Označme pro každou dvojici $\{i, j\} \subseteq \{1, \dots, n\}$, $i < j$, indikátor Y_{ij} shody dvojice. Pak $X = \sum_{\{i,j\}} Y_{ij}$, součet je přes všech $\binom{n}{2}$ dvojic. Pro každou dvojici stejně jako v bodě (2) platí $E(Y_{ij}) = 1/6$. Veličiny Y_{ij} nejsou nezávislé jako celek (po dvou sice ano, ale když $Y_{ij} = Y_{jk} = 1$, je nutně i $Y_{ik} = 1$), ale *linearita střední hodnoty* nezávislost „nepotřebuje“. Tedy

$$E(X) = \sum_{\{i,j\}} E(Y_{ij}) = \binom{n}{2} \cdot \frac{1}{6} = \boxed{\frac{n(n-1)}{12}}.$$

Kontrola: pro $n = 2$ je $E(X) = 1/6$, což sedí s bodem (2); pro $n = 10$ je $E(X) = 90/12 = 7,5$.

2 Centrální limitní věta

(Reálná otázka: léto 2025, č. 4 – Centrální limitní věta)

Zadání.

- Napište znění Centrální limitní věty.
- Předpokládejme, že program jeden vstup zpracovává náhodnou dobu se střední hodnotou 100 ms a směrodatnou odchylkou 20 ms. Použijte Centrální limitní větu pro odhad pravděpodobnosti, že zpracování sto vstupů bude trvat více než 10.3 s. Napište přesnou formuli pomocí Φ a také vyčíslete pomocí tabulky níže.
- Vysvětlete, jaké jsme udělali předpoklady (záleží na tom, jaké je rozdělení doby běhu programu?) a jak jsou realistické.

x	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5
$\Phi(x)$	0.01	0.02	0.07	0.16	0.31	0.5	0.69	0.84	0.93	0.98	0.99

Náčrt řešení.

- Znění věty: Necht $X_1, X_2, \dots$ jsou stejně rozdělené n.n.v. se střední hodnotou μ a rozptylem σ^2 . Označme

$$Y_n = \frac{(X_1 + \dots + X_n) - n\mu}{\sqrt{n} \cdot \sigma}.$$

Pak

$$Y_n \xrightarrow{d} N(0, 1).$$

Neboli, pokud F_n je distribuční funkce Y_n , tak

$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x) \quad \text{pro každé } x \in \mathbb{R}.$$

Říkáme, že posloupnost Y_n konverguje k $N(0, 1)$ v *distribuci*.

2. Označme X_i dobu běhu pro i -tý vstup. To, že $X_1 + \dots + X_{100} > 10.3$ je ekvivalentní tomu, že

$$Y_{100} > \frac{10.3 - 100 \cdot 0.1}{\sqrt{100} \cdot 0.02} = 1.5.$$

Podle CLV je $P(Y_{100} < 1.5) \approx \Phi(1.5)$. Proto hledaná pravděpodobnost je přibližně $1 - \Phi(1.5) \approx 1 - 0.93 = 0.07$, neboli zhruba 7 %.

3. CLV nepředpokládá nic speciálního o rozdělení jednotlivých X_i (pokud jsou veličiny stejně rozdělené s konečnou střední hodnotou a rozptylem). Předpokládáme ale nezávislost – to může být problém; pokud je například počítač přetížen, bude na všech vstupech pomalejší. Také nahrazujeme limitou hodnotu pro $n = 100$, to ale typicky při aplikaci CLV není problém.

Doplňující vysvětlení (nad rámec oficiálního náčrtu):

- *Přesná formule.* Jednotky sjednotíme na sekundy: $\mu = 0.1$, $\sigma = 0.02$, $n = 100$. Pro $S_n = X_1 + \dots + X_n$ je

$$P(S_{100} > 10.3) = P\left(Y_{100} > \frac{10.3 - n\mu}{\sqrt{n}\sigma}\right) \approx 1 - \Phi\left(\frac{10.3 - 100 \cdot 0.1}{\sqrt{100} \cdot 0.02}\right) = 1 - \Phi(1.5).$$

Přesnější hodnota $\Phi(1.5) \approx 0.9332$ dává ≈ 0.067 ; tabulka v zadání zaokrouhluje na 0.07.

- *Záleží na rozdělení doby běhu?* V limitě ne: CLV platí pro libovolné rozdělení s konečnou střední hodnotou a nenulovým konečným rozptylem, stačí znát μ a σ . Pro konečné $n = 100$ ale tvar rozdělení ovlivňuje, jak přesná aproximace je; u silně zešikmeného rozdělení (např. občasné velmi dlouhé běhy) bývá chyba v chvostu, na který se tu ptáme, větší.
- *Stejné rozdělení.* Kromě nezávislosti předpokládáme, že všechny vstupy mají stejné rozdělení doby zpracování. To je rozumné, pokud vstupy nejsou systematicky různě náročné (např. různě velké).