# Náhodný výběr

Matematická statistika poskytuje metody pro popis veličin náhodného charakteru pomocí jejich pozorovaných hodnot, přesněji řečeno jde o určení důležitých vlastností rozdělení pravděpodobnosti náhodné veličiny nebo náhodného vektoru z jejich hodnot získaných měřením, statistickým šetřením, nepřímým pozorováním apod. Tyto metody jsou v podstatě zaměřeny na řešení dvou základních úloh matematické statistiky:

- *odhady parametrů a rozdělení,*
- *testování statistických hypotéz o parametrech a rozděleních.*

Tyto úlohy se dle potřeby kombinují, když např. odhadujeme nebo testujeme číselné charakteristiky rozdělení, vyšetřujeme závislosti náhodných veličin apod. Metody matematické statistiky jsou založeny na následujících pojmech.

**1. Pojmy** Opakujeme-li $n$-krát nezávisle pokus, jehož výsledkem je hodnota náhodné veličiny $X$ s distribuční funkcí $F(x,\vartheta)$, kde $\vartheta$ je reálný parametr (případně vektor parametrů anebo jejich funkce) daného rozdělení pravděpodobnosti, pozorujeme vlastně náhodný vektor $\mathbf{X} = (X_1,\ldots,X_n)$ a předpokládáme, že jeho složky jsou nezávislé náhodné veličiny $X_i$ se stejnou distribuční funkcí (pravděpodobnostní funkcí anebo hustotou pravděpodobnosti) jako má pozorovaná náhodná veličina $X$. Náhodný vektor $\mathbf{X}$ se nazývá **náhodný výběr** (z náhodné veličiny $X$ nebo z jejího rozdělení pravděpodobnosti) a číslo $n$ je **rozsah** náhodného výběru. Analogicky definujeme náhodný výběr z náhodného vektoru.

**2. Vlastnosti** Náhodný výběr $\mathbf{X} = (X_1,\ldots,X_n)$ má simultánní distribuční funkci

$$F(\mathbf{x};\vartheta) = F(x_1,\ldots,x_n;\vartheta) = \prod_{i=1}^{n} F(x_i;\vartheta)$$

a simultánní pravděpodobnostní funkci

$$p(\mathbf{x};\vartheta) = p(x_1,\ldots,x_n;\vartheta) = \prod_{i=1}^{n} p(x_i;\vartheta),$$

kde $p(x_i;\vartheta)$ je pravděpodobnostní funkce $i$-té složky, $i = 1,\ldots,n$, jestliže pozorovaná náhodná veličina $X$ je diskrétní, resp. simultánní hustotu pravděpodobnosti

$$f(\mathbf{x};\vartheta) = f(x_1,\ldots,x_n;\vartheta) = \prod_{i=1}^{n} f(x_i;\vartheta),$$

kde $f(x_i;\vartheta)$ je hustota pravděpodobnosti $i$-té složky $X_i$, jestliže pozorovaná náhodná veličina $X$ je spojitá.

**3. Pojmy** Číselný vektor $\mathbf{x} = (x_1,\ldots,x_n)$, který získáme při realizaci náhodného výběru, kde $x_i$ je pozorovaná hodnota složky, $i = 1,\ldots,n$, je **statistický soubor** s **rozsahem** $n$. Množina všech hodnot náhodného výběru, tj. množina všech statistických souborů, tvoří **výběrový prostor.**

**4. Poznámka** Statistický soubor je jinak řečeno pozorovaná hodnota náhodného výběru, což znamená, že při opakovaných realizacích náhodného výběru obdržíme obecně (a náhodně) různé statistické soubory. Zpracování statistického souboru je popsáno v kapitole Popisná statistika.

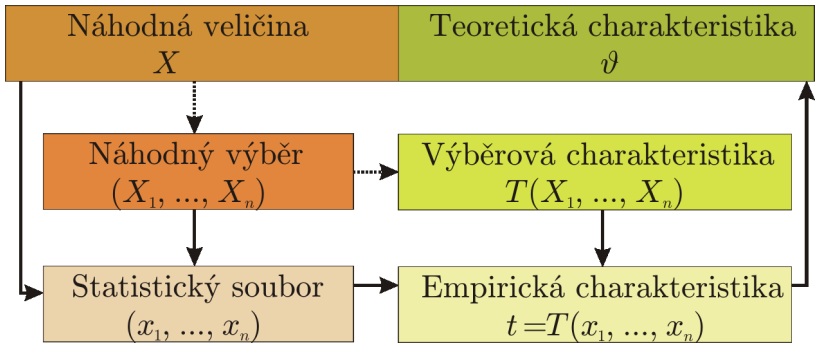
**5. Příklad** Jestliže náhodná veličina $X$ má binomické rozdělení pravděpodobnosti $\mathrm{Bi}(1,p)$ s parametrem $p \in (0;1)$, má pravděpodobnostní funkci $p(x) = p^x(1-p)^{1-x}$, kde $x \in \{0;1\}$. Náhodný výběr z tohoto rozdělení pravděpodobnosti má simultánní pravděpodobnostní funkci

$$p(x_1,\ldots,x_n) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i},$$

kde $x_i \in \{0;1\}$. Výběrový prostor je množina všech statistických souborů $\mathbf{x} = (x_1, \ldots, x_n)$, tj. množina $\{0;1\}^n$.

**6. Pojmy** Funkce náhodného výběru $T(X_1, \ldots, X_n)$ se nazývá **výběrová charakteristika** nebo **statistika**. Její hodnota na statistickém souboru $t = T(x_1, \ldots, x_n)$ je **empirická charakteristika** nebo **pozorovaná hodnota statistiky** $T$.

**7. Poznámka** Výběrovou charakteristiku (statistiku) $T$ (a tím také empirickou charakteristiku $t$) volíme tak, nabývala na výběrovém prostoru s velkou pravděpodobností hodnot blízkých neznámé nebo předpokládané teoretické charakteristice, např. parametru $\vartheta$ pozorované náhodné veličiny $X$. Z toho vyplývá základní princip statistické indukce v matematické statistice, který je schematicky vyjádřen na Obrázku 1.




Obrázek 1: Základní princip statistické indukce

**8. Pojmy** Používáme zejména tyto výběrové charakteristiky:

1. **výběrový průměr** $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,
2. **výběrový rozptyl** $S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$,
3. **výběrová směrodatná odchylka** $S = \sqrt{S^2}$,
4. **výběrový koeficient korelace** $R = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{S(X)\,S(Y)}$ pro náhodný výběr z náhodného vektoru $(X, Y)$, kde $S(X)$ a $S(Y)$ jsou výběrové směrodatné odchylky náhodných veličin $X$ a $Y$.

**9. Vlastnosti** Základní vlastnosti výběrového průměru $\overline{X}$ a výběrového rozptylu $S^2$jsou:

1. Jestliže pozorovaná náhodná veličina $X$ má střední hodnotu $E(X)$, pak
$$E\left(\overline{X}\right) = E(X).$$
2. Jestliže pozorovaná náhodná veličina $X$ má rozptyl $D(X)$, pak
$$D\left(\overline{X}\right) = \frac{D(X)}{n}, \quad \sigma\left(\overline{X}\right) = \frac{\sigma(X)}{\sqrt{n}}, \quad E\left(S^2\right) = \frac{n-1}{n}D(X).$$

Hodnoty výběrových charakteristik jsou empirické charakteristiky, které získáme po zpracování statistického souboru. Např. aritmetický průměr $\overline{x}$ je pozorovaná hodnota výběrového průměru apod. Tyto

hodnoty jsou však náhodné, jinak řečeno empirické charakteristiky se při opakovaných realizacích náhodného výběru náhodně mění. Avšak z předcházejícího plyne, že např. pro $n \to \infty$ rozptyl výběrového průměru $D\left(\overline{X}\right) \to 0$, takže pro dostatečně velké $n$ je takřka jistě aritmetický průměr blízký neznámé střední hodnotě. Přitom ale $\sigma\left(\overline{X}\right) \to 0$ pouze s rychlostí $n^{1/2}$, což znamená, že např. pro dosažení dvojnásobné přesnosti aproximace neznámé střední hodnoty $E(X)$ aritmetickým průměrem $\overline{x}$ musíme zvýšit rozsah náhodného výběru čtyřikrát atd. Ve statistické literatuře se hovoří o tzv. *statistické kletbě*.

**10. Poznámka** Protože $\frac{n-1}{n} < 1$, je $E\left(S^2\right) < D(X)$, takže empirické hodnoty $s^2$ se vzhledem ke skutečnému (a obvykle neznámému) rozptylu častěji vychylují doleva (do menších hodnot) od $D(X)$. Proto se mnohdy definuje výběrový rozptyl $\hat{S}^2$ ve tvaru

$$\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

a pro tento výběrový rozptyl je $E\left(\hat{S}^2\right) = D(X)$. Odpovídající rozptyl statistického souboru pak je

$$\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Statistika $\hat{S}^2$ má však větší rozptyl než statistika $S^2$, ale pro velká $n$ (řádově 100 a více) je rozdíl mezi těmito statistikami zanedbatelný. Analogicky definujeme výběrovou směrodatnou odchylku $\hat{S}$ a směrodatnou odchylku statistického souboru $\hat{s}$. Různé definice uvedených charakteristik je nutno respektovat při zpracování statistického souboru na PC pomocí statistických programů a také ve vzorcích jak pro odhady parametrů, tak i pro testování statistických hypotéz.

Stochastické vlastnosti nejčastěji používaných výběrových charakteristik vyjadřují jejich následující tzv. **statistická rozdělení pravděpodobnosti**. Potřebné hodnoty těchto rozdělení jsou tabelovány anebo se počítají na PC pomocí statistických programů (např. Statistica, S-Plus, Statgraphics, QCExpert, Minitab, Adstat aj.) nebo statistických funkcí (např. Excel).

1. **Normální rozdělení** pravděpodobnosti $\mathrm{N}\left(\mu, \sigma^2\right)$, kde $\mu$, $\sigma^2$ jsou reálná čísla, $\sigma^2 > 0$, náhodné veličiny $X$ (viz kapitolu Rozdělení pravděpodobnosti pro aplikace), zejména pak normované normální rozdělení pravděpodobnosti $\mathrm{N}(0;1)$ náhodné veličiny $U = \frac{X-\mu}{\sigma}$ s distribuční funkcí $\Phi(u)$, jejíž hodnoty jsou tabelovány v tabulce **T1**. Pro kvantily $u_P$ je $u_P = -u_{1-P}$, kde$P \in (0;1)$. Tabulka **T1** také obsahuje nejčastěji používané kvantily pro $P = 0,95;\ 0,975;\ 0,99;\ 0,995$. Normální rozdělení má řadu velmi důležitých vlastností. Např. jestliže nezávislé náhodné veličiny $X_i$ mají rozdělení $\mathrm{N}(\mu_i; \sigma_i^2)$ pro , pak náhodná veličina $\sum_{i=1}^{n} X_i$ má normální rozdělení $\mathrm{N}\left(\sum_{i=1}^{n} \mu_i; \sum_{i=1}^{n} \sigma_i^2\right)$.

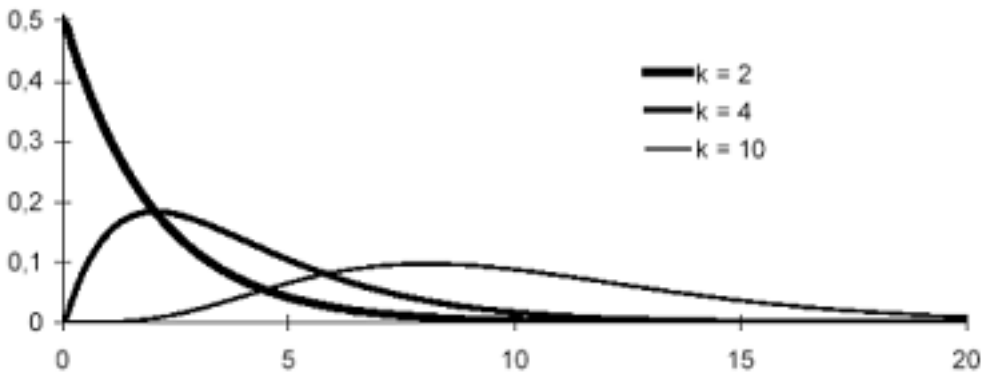
2. **Pearsonovo rozdělení** (*chí-kvadrát rozdělení*) $\chi^2(k)$ s $k$ stupni volnosti, kde $k$ je přirozené číslo, má hustotu pravděpodobnosti

$$f(x) = \begin{cases} \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \mathrm{e}^{-\frac{x}{2}} x^{\frac{k}{2}-1} & \text{pro } x \in (0;\infty), \\ 0 & \text{pro } x \in (-\infty;0\rangle, \end{cases}$$

kde $\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\mathrm{d}t$, $z > 0$, je tzv. *gama funkce*. Graf hustoty pravděpodobnosti Pearsonova rozdělení, které je kladně asymetrické, je znázorněn na Obrázku 2 a jeho základní číselné charakteristiky jsou:

$$E(X) = k, \quad D(X) = 2k, \quad A(X) = 4/\sqrt{2k} > 0.$$

Jestliže $U_1, \ldots, U_k$ jsou nezávislé náhodné veličiny s normovaným normálním rozdělením, pak náhodná veličina $\sum_{i=1}^{k} U_i^2$ má Pearsonovo rozdělení. Kvantily $\chi_P^2$ tohoto rozdělení jsou tabelovány v tabulce **T3**.

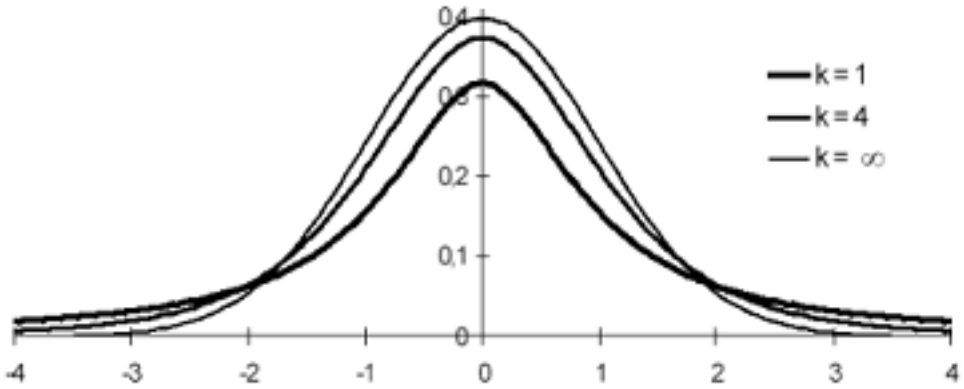



Obrázek 2: Grafy hustoty pravděpodobnosti Pearsonova rozdělení $\chi^2(k)$

3. **Studentovo rozdělení** (*t rozdělení*) $\mathrm{S}(k)$ s $k$ stupni volnosti, kde $k$ je přirozené číslo, má hustotu pravděpodobnosti

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad x \in (-\infty; \infty).$$

Graf hustoty pravděpodobnosti Studentova rozdělení, které je symetrické vzhledem k $x = 0$, je znázorněn na Obrázku 3 a jeho základní číselné charakteristiky jsou:
$E(X) = 0$ pro $k > 1$, $D(X) = k/(k-2)$ pro $k > 2$, $A(X) = 0$ pro $k > 3$, $x_{0,5} = 0$.



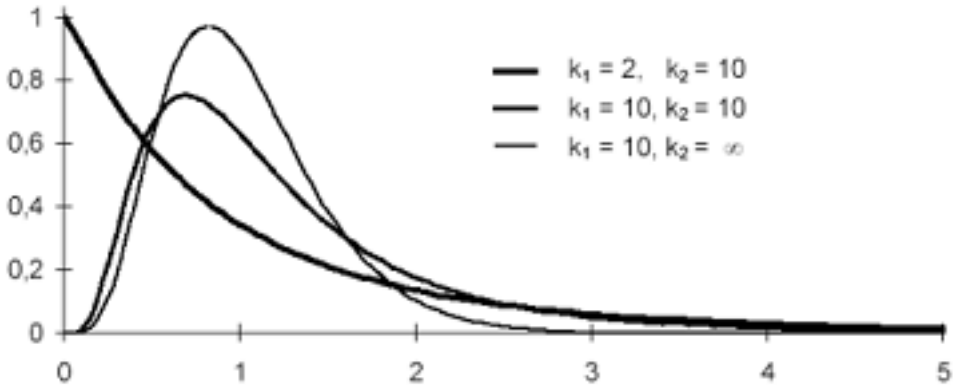

Obrázek 3: Grafy hustoty pravděpodobnosti Studentova rozdělení $\mathrm{S}(k)$

Studentovo rozdělení s jedním stupněm volnosti je tzv. **Cauchyovo rozdělení**. Pro $k \to \infty$ konverguje Studentovo rozdělení k normovanému normálnímu rozdělení $\mathrm{N}(0;1)$. Jestliže $U$ a $V$ jsou nezávislé náhodné veličiny, přičemž $U$ má normované normální rozdělení a $V$ má Pearsonovo rozdělení $\chi^2(k)$, pak náhodná veličina $\frac{U}{\sqrt{V}}\sqrt{k}$ má Studentovo rozdělení $\mathrm{S}(k)$. Kvantily $t_P$ tohoto rozdělení jsou tabelovány v tabulce **T2** a pro je $t_P = -t_{1-P}$.

4. **Fisherovo-Snedecorovo rozdělení** (*F rozdělení*) $\mathrm{F}(k_1, k_2)$ s $k_1, k_2$ stupni volnosti, kde jsou přirozená čísla, má hustotu pravděpodobnosti

$$f(x) = \begin{cases} \frac{1}{B\left(\frac{k_1}{2},\frac{k_2}{2}\right)} \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} x^{\frac{k_1}{2}-1} \left(1+\frac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}} & \text{pro } x \in (0;\infty)\,, \\ 0 & \text{pro } x \in (-\infty;0\rangle, \end{cases}$$

kde $B(z_1,z_2) = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\,\mathrm{d}t = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$, $z_1 > 0$, $z_2 > 0$, je tzv. *beta funkce*. Graf hustoty rozdělení, které je kladně asymetrické, je znázorněn na Obrázku 4 a jeho základní číselné charakteristiky jsou: $E(X) = k_2/(k_2-2)$ pro $k_2 > 2$, $D(X) = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$ pro $k_2 > 4$.




Obrázek 4: Grafy hustoty pravděpodobnosti Fisherova-Snedecorova rozdělení $\mathrm{F}(k_1,k_2)$

Jestliže $V_1$ a $V_2$ jsou nezávislé náhodné veličiny, přičemž $V_1$ má Pearsonovo rozdělení $\chi^2(k_1)$ a $V_2$ má Pearsonovo rozdělení $\chi^2(k_2)$, pak náhodná veličina $\frac{V_1/k_1}{V_2/k_2}$ má Fisherovo-Snedecorovo rozdělení. Kvantily $F_P(k_1,k_2)$ tohoto rozdělení jsou tabelovány v tabulce **T4** a pro $P \in (0;1)$ je $F_P(k_1,k_2) = 1/F_{1-P}(k_2,k_1)$.

Nejčastěji řešené úlohy při aplikacích metod matematické statistiky se týkají pozorovaných náhodných veličin s normálním rozdělením pravděpodobnosti. Využíváme přitom následující vlastnosti tohoto rozdělení.

**11. Vlastnosti** Jestliže pozorovaná náhodná veličina $X$ má normální rozdělení $\mathrm{N}(\mu;\sigma^2)$, pak platí:

1. $\overline{X}$ má normální rozdělení $\mathrm{N}(\mu;\frac{\sigma^2}{n})$,
2. $\frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ má normální rozdělení $\mathrm{N}(0;1)$,
3. $\frac{\overline{X}-\mu}{S}\sqrt{n-1}$ má Studentovo rozdělení $\mathrm{S}(n-1)$,
4. $\frac{nS^2}{\sigma^2}$ má Pearsonovo rozdělení $\chi^2(n-1)$.

**12. Vlastnosti** Jestliže pozorovaná náhodná veličina $X$ má normální rozdělení $\mathrm{N}\left(\mu\left(X\right),\sigma^2\left(X\right)\right)$ a pozorovaná náhodná veličina $Y$ má normální rozdělení $\mathrm{N}\left(\mu\left(Y\right),\sigma^2\left(Y\right)\right)$, $X$ a $Y$ jsou nezávislé a také náhodné výběry $(X_1,\ldots,X_{n_1})$, $(Y_1,\ldots,Y_{n_2})$ jsou nezávislé, pak statistika:

1. $\frac{\overline{X}-\overline{Y}-(\mu(X)-\mu(Y))}{\sqrt{\frac{\sigma^2(X)}{n_1}+\frac{\sigma^2(Y)}{n_2}}}$ má normální rozdělení,
2. $\frac{\overline{X}-\overline{Y}-(\mu(X)-\mu(Y))}{\sqrt{n_1S^2(X)+n_2S^2(Y)}}\sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}$ má pro $\sigma^2(X)=\sigma^2(Y)$ Studentovo rozdělení $\mathrm{S}\left(n_1+n_2-2\right)$,
3. $\frac{\frac{n_1S^2(X)}{n_1-1}}{\frac{n_2S^2(Y)}{n_2-1}}$ má pro $\sigma^2(X)=\sigma^2(Y)$ Fisherovo-Snedecorovo rozdělení $\mathrm{F}(n_1-1,n_2-1)$.

**13. Vlastnosti** Jestliže $X_1,X_2,\ldots$ je posloupnost nezávislých náhodných veličin s libovolným stejným rozdělením pravděpodobnosti (např. i asymetrickým nebo diskrétním), které má střední hodnotu $\mu_0$ a směrodatnou odchylku $\sigma_0$, pak posloupnost náhodných veličin

$$\frac{\frac{1}{n}\sum_{i=1}^{n}X_i-\mu_0}{\sigma_0}\sqrt{n}$$

konverguje (v distribuci) k náhodné veličině $U$ s normovaným normálním rozdělením $\mathrm{N}(0;1)$.

Z předcházející vlastnosti plyne, že při dostatečně velkém rozsahu náhodného výběru $n$ můžeme rozdělení pravděpodobnosti výběrového aritmetického průměru pro libovolnou pozorovanou náhodnou veličinu $X$ se střední hodnotou a rozptylem $\sigma_0^2$ aproximovat normálním rozdělením $\mathrm{N}\left(\mu_0;\frac{\sigma_0^2}{n}\right)$. To také znamená, že při dostatečně velkém rozsahu statistického souboru $n$ má smysl aproximovat např. střední hodnotu $\mu_0$ aritmetickým průměrem $\overline{x}$.

**14. Příklad** Rozdělení pravděpodobnosti výběrového průměru $\overline{X}$ pro náhodný výběr z binomického rozdělení pravděpodobnosti (viz Příklad 5) lze pro dostatečně velký rozsah výběru $n$ dobře aproximovat rozdělením normálním $\mathrm{N}(p;\frac{p(1-p)}{n})$, neboť $\mu_0=p$ a $\sigma_0^2=p(1-p)$. Tato aproximace rozdělení pravděpodobnosti výběrového průměru je dostačující pro $n>\frac{9}{p(1-p)}$.