6 Multihypergeometrické rozdělení $\text{MultiHyperGeom}(N, \mathbf{p})$

- Nechť N_{pop} je rozsah populace, M_j, $j = 1, \ldots, k$ je počet statistických jednotek s j-tou sledovanou charakteristikou CH_j vysytujících se v populaci N_{pop} a N je rozsah náhodného výběru vybraného z popupace N_{pop} bez vrácení.
- X_j ... počet statistických jednotek se sledovanou charakteristikou CH_j, vyskytujících se v náhodném výběru o rozsahu N.
- $\mathbf{X} = (X_1, \ldots, X_k)^T \sim \text{MultiHyperGeom}(N, \mathbf{p})$, kde $\mathbf{p} = (p_1, \ldots, p_k)^T = \left(\frac{M_1}{N_{pop}}, \ldots, \frac{M_k}{N_{pop}}\right)^T$
- $\theta = \mathbf{p}$
- pravděpodobnostní funkce
$$p(\mathbf{x}) = p(x_1, \ldots, x_k) = \frac{\prod_{j=1}^{k} \binom{M_j}{x_j}}{\binom{N_{\text{pop}}}{N}}, \qquad x_j = 0, 1, \ldots; \quad j = 1, \ldots, k$$
- $E[\mathbf{X}] = N\mathbf{p}$
- marginální rozdělení
  - M_j vs. $N_{pop} - M_j$, tj. počet statistických jednotek s j-tou charakteristikou vs. počet všech ostatních statistických jednotek
  - pravděpodobnostní funkce j-tého marginálního rozdělení
$$p(x_j) = \frac{\binom{M_j}{x_j}\binom{N_{\text{pop}}-M_j}{N-x_j}}{\binom{N_{\text{pop}}}{N}}, \qquad x = 0, 1, \ldots$$
    $\rightarrow$ hypergeometrické rozdělení $\text{HyperGeom}(N, p_j)$
- extraDistr::dmvhyper(x, M, N), extraDistr::rmvhyper(n, M, N)
- Data:
  - **Dataset 5: Počet obyvatel Jihomoravského kraje**
  - Podle údajů o počtu obyvatelstva v ČR získaných z webových stránek statistického úřadu www.czso.cz má Jihomoravský kraj ke dni 30.6.2018 celkem 1 184 381 obyvatel. Rozmístění obyvatel v jednotlivých okresích Jihomoravského kraje je k dispozici v tabulce 1.

Tabulka 1: Počet obyvatel v okresích Jihomoravského kraje k datu 30.6.2018

Okres	Blansko	Brno-město	Brno-venkov	Břeclav	Hodonín	Vyškov	Znojmo	$\sum$
Počet obyvatel	108 641	379 275	221 200	115 728	154 183	91 483	113 871	1 184 381

Příklad 6.1. Pravděpodobnostní funkce multihypergeometrického modelu
Naprogramujte v R funkci dmultihypergeom(x, M) počítající hodnoty pravděpodobnostní funkce multihypergeometrického rozdělení $\text{MultiHyperGeom}(N, \mathbf{p})$, kde $\mathbf{p} = (p_1, \ldots, p_k)^T$. Správnost funkce otestujte na výpočtu $p(\mathbf{x})$, pro $X \sim \text{MultiHyperGeom}(N, \mathbf{p})$, kde $\mathbf{p} = \left(\frac{M_1}{N_{pop}}, \ldots, \frac{M_k}{N_{pop}}\right)^T = \left(\frac{5}{30}, \frac{10}{30}, \frac{15}{30}\right)^T$. Vektor $\mathbf{x}$ zvolte (a) $\mathbf{x} = (3, 6, 9)$; (b) $\mathbf{x} = (4, 5, 9)$; (c) $\mathbf{x} = (5, 6, 7)$ (d) $\mathbf{x} = (7, 6, 5)$. Výsledky ověřte s výsledky funkce dmvhyper() z knihovny extraDistr. Jaký je v tomto případě rozsah populace N_{pop} a jaký je rozsah reprezentativního vzorku N?

Řešení příkladu 6.1

```
         p1         p2         p3 p4
1 0.1215182 0.07291091 0.01562377  0
```

Výsledné hodnoty pravděpodobnostní funkce jsou (a) $p(3, 6, 9) = 0.1215$; (b) $p(4, 5, 9) = 0.0729$; (c) $p(5, 6, 7) = 0.0156$; (d) $p(7, 6, 5) = 0$. Rozsah celé populace $N_{pop} = 30$. Rozsah reprezentativního vzorku $N = 18$. ★

Příklad 6.2. Výpočet pravděpodobností na základě multihypergeometrického modelu
Jana s Bárou a Kájou dostali adventní kalendář, ve kterém je třetina čokolád hořkých, třetina čokolád mléčných a třetina čokolád bílých. Příchutě čokolád jsou v kalendáři rozmístěny náhodně. O čokolády se děti rozhodly podělit rovným dílem, ale protože je Kája nejmenší, dovolily mu sestry, aby svůj díl čokolád snědl jako první. Vypočítejte, jaká je pravděpodobnost, že Kája bude mít ve svém dílu (a) dvě hořké, dvě bílé a čtyři mléčné čokolády; (b) čtyři mléčné a čtyři hořké čokolády; (c) maximálně dvě čokolády hořké; (d) více než polovinu čokolád mléčných.

Řešení příkladu 6.2

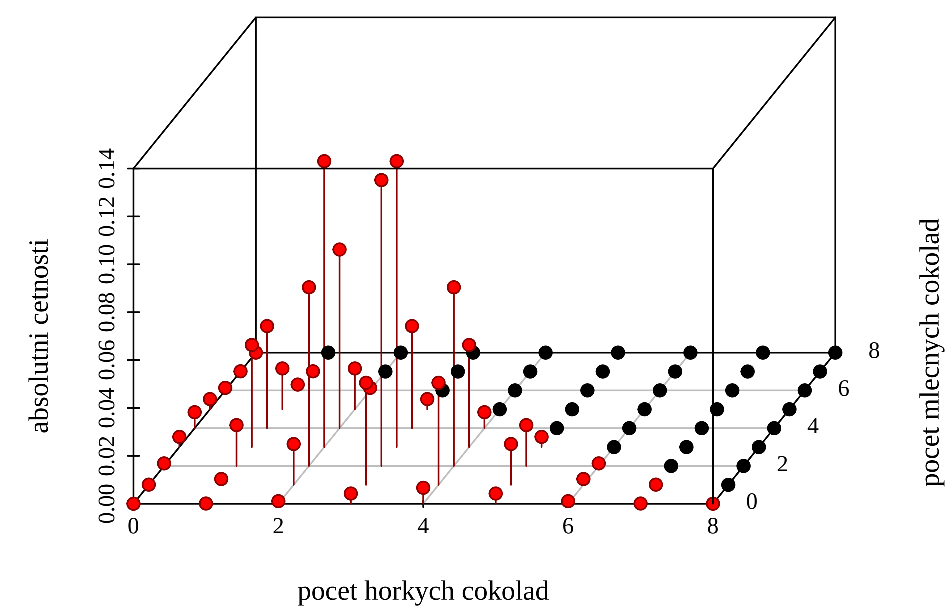
```
             p1          p2         p3          p4
Kaja 0.07461885 0.006662397 0.4468076 0.04738324
```

	2 hořké, 2 bílé, 4 mléčné	4 mléčné, 4 hořké	maximálně 2 hořké	více než 1/2 mléčných
Kája	0.0746	0.0067	0.4468	0.0474

Pravděpodobnost, že Kája bude mít ve svém dílu dvě hořké, dvě bílé a čtyři mléčné čokolády je 7.46 %. Pravděpodobnost, že Kája bude mít čtyři mléčné a čtyři hořké čokolády je 0.67 %. Pravděpodobnost, že Kája bude mít mezi svými čokoládami maximálně dvě hořké, je 44.68 %. Pravděpodobnost, že více než polovina Kájových čokolád bude mléčných, je 4.74 %. ★

Příklad 6.3. Pravděpodobnostní funkce multihypergeometrického modelu
V příkladu 6.2 jsme stanovili, že počet hořkých, mléčných a bílých čokolád v Kájově dílu se bude řídit multihypergeometrickým modelem MultiHyperGeom$(N, \mathbf{p})$, kde $N = 8$ a $\mathbf{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^T$. Vykreslete graf pravděpodobnostní funkce rozdělení MultiHyperGeom$(N, \mathbf{p})$.

Řešení příkladu 6.3



★

Příklad 6.4. Multihypergeometrický model
Podle údajů uvedených v datasetu 5 má Jihomoravský kraj ke dni 30.6.2018 celkem 1 184 381 obyvatel, přičemž 108 641 obyvatel náleží do okresu Blansko, 379 275 obyvatel náleží do okresu Brno-město, atd. Předpokládejme, že chceme sestavit reprezentativní vzorek N obyvatel pocházejících z Jihomoravského kraje. Náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)$ popisující rozložení počtu obyvatel z jednotlivých okresů Jihomoravského kraje v reprezentativním vzorku má potom multihypergeometrické rozdělení, tj. $\mathbf{X} \sim \text{MultiHyperGeom}(N, \mathbf{p})$, kde $\mathbf{p}$ je vektor pravděpodobností výskytu obyvatel z jednotlivých okresů Jihomoravského kraje.

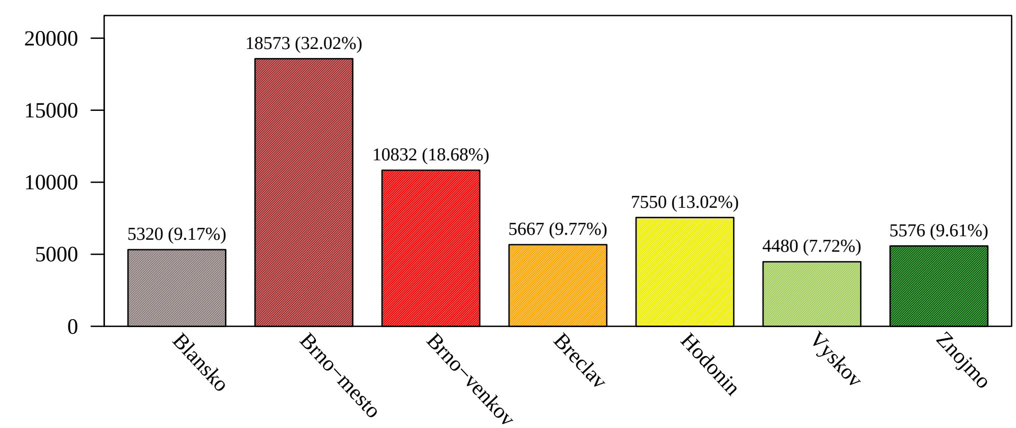
(1) Vypočítejte odhad vektoru $\mathbf{p}$. (2) Stanovte, jaké bude rozložení počtu obyvatel z jednotlivých okresů v reprezentativním vzorku za předpokladu, že rozsah reprezentativního vzorku N bude (a) 580; (b) 58 000; (c) 550 000; (d) 580 000; (e) 900 000; (f) 1 100 000. (3) Pro $N = 58\,000$ a $N = 900\,000$ nakreslete sloupcový diagram očekávaných absolutních četností obyvatel z každého okresu.

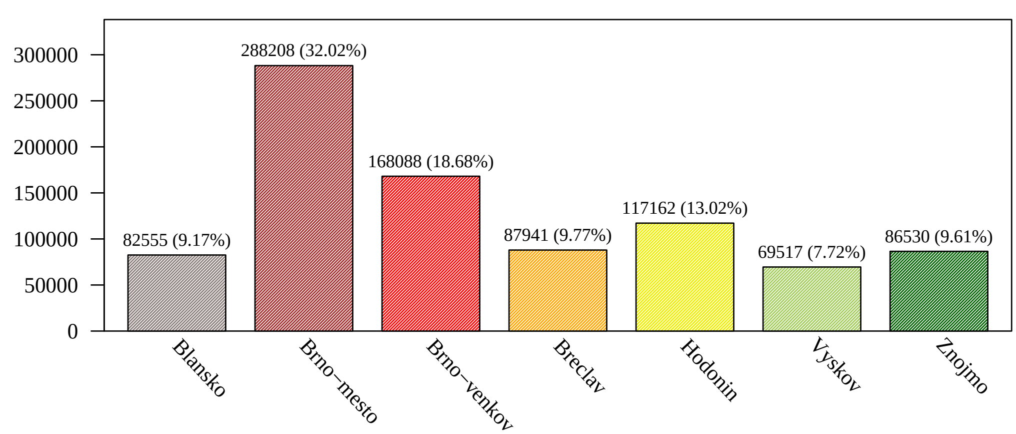
Řešení příkladu 6.4

```
     Blansko Brno-mesto Brno-venkov   Breclav   Hodonin     Vyskov     Znojmo
1 0.09172808  0.3202306   0.1867642 0.0977118 0.1301802 0.07724119 0.09614389
```

Odhad vektoru parametrů $\mathbf{p} = (0.0917, 0.3202, 0.1868, 0.0977, 0.1302, 0.0772, 0.0961)^T$.

```
            Blansko Brno-mesto Brno-venkov Breclav Hodonin Vyskov Znojmo     Sum
N = 580          53        186         108      57      76     45     56     581
N = 58000      5320      18573       10832    5667    7550   4480   5576   57998
N = 550000    50450     176127      102720   53741   71599  42483  52879  549999
N = 580000    53202     185734      108323   56673   75505  44800  55763  580000
N = 9e+05     82555     288208      168088   87941  117162  69517  86530  900001
N = 1100000  100901     352254      205441  107483  143198  84965 105758 1100000
```





★