Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Téma 6: Multihypergeometrický model

Veronika Bendová

bendova.veroonika@gmail.com

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

- 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
- $\mathbf{X} = (X_1, \ldots, X_k)^T \sim \mathrm{MultiHyperGeom}(N, \mathbf{p})$, 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

$$p(\mathbf{x}) = p(x_1, \ldots, x_k) = \frac{\prod_{j=1}^{k} \binom{M_j}{x_j}}{\binom{N_{\mathrm{pop}}}{N}}, \qquad x_j = 0, 1, \ldots; \qquad j = 1, \ldots, k$$

- $E[\mathbf{X}] = N\mathbf{p}$

- 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

$$p(x_j) = \frac{\binom{M_j}{x_j}\binom{N_{\mathrm{pop}} - M_j}{N - x_j}}{\binom{N_{\mathrm{pop}}}{N}}, \qquad x = 0, 1, \ldots$$

  $\rightarrow$ hypergeometrické rozdělení $\mathrm{HyperGeom}(N, p_j)$

- extraDistr::dmvhyper(x, M, N), extraDistr::rmvhyper(n, M, N)

- **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 níže uvedené tabulce.

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 $\mathrm{MultiHyperGeom}(N, \mathbf{p})$, kde $\mathbf{p} = (p_1, \ldots, p_k)^T$. Správnost funkce otestujte na výpočtu $p(\mathbf{x})$, pro $X \sim \mathrm{MultiHyperGeom}(N, \mathbf{p})$, 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

```
1  dmultihyper <- function(...){
2    Npop <- ... # celkovy rozsah populace vypocitany pomoci vektoru M
3    N <- ... # rozsah nahodneho vyberu vypocitany pomoci vektoru x
4    px <- prod(choose(...)) / choose(...) # pstni fce MultiHyperGeom()
5    return(...)
6  }
7  M <- ... # vektor M
8  dmultihyper(...) # vypocet (a), (b), (c), (d)
9  extraDistr::dmvhyper(rbind(...), ...) # vypocet (a), (b), (c), (d)
```

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

Výsledné hodnoty pravděpodobnosntí 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
(a)

```
12  library(...) # nacteni knihovny extraDistr
13  Npop <- ... # celkovy pocet cokolad
14  M <- ... # pocet horkych, mlecnych a bilych cokolad v cele populaci
15  p1 <- dmvhyper(...) # vypocet pravdepodobnosti
```

```
[1] 0.07461885                                                          16
```

(b)

```
17  p2 <- dmvhyper(...)  # vypocet pravdepodobnosti
```

```
[1] 0.006662397
```
18

(c)

```
19  p3 <- sum(dmvhyper(rbind(...), ...))
20  p3 <- phyper(...)
```

```
[1] 0.4468076
```
21

(d)

```
22  p4 <- 1 - sum(dmvhyper(rbind(...), ...))
23  p4 <- 1 - phyper(...)
```

```
[1] 0.04738324
```
24

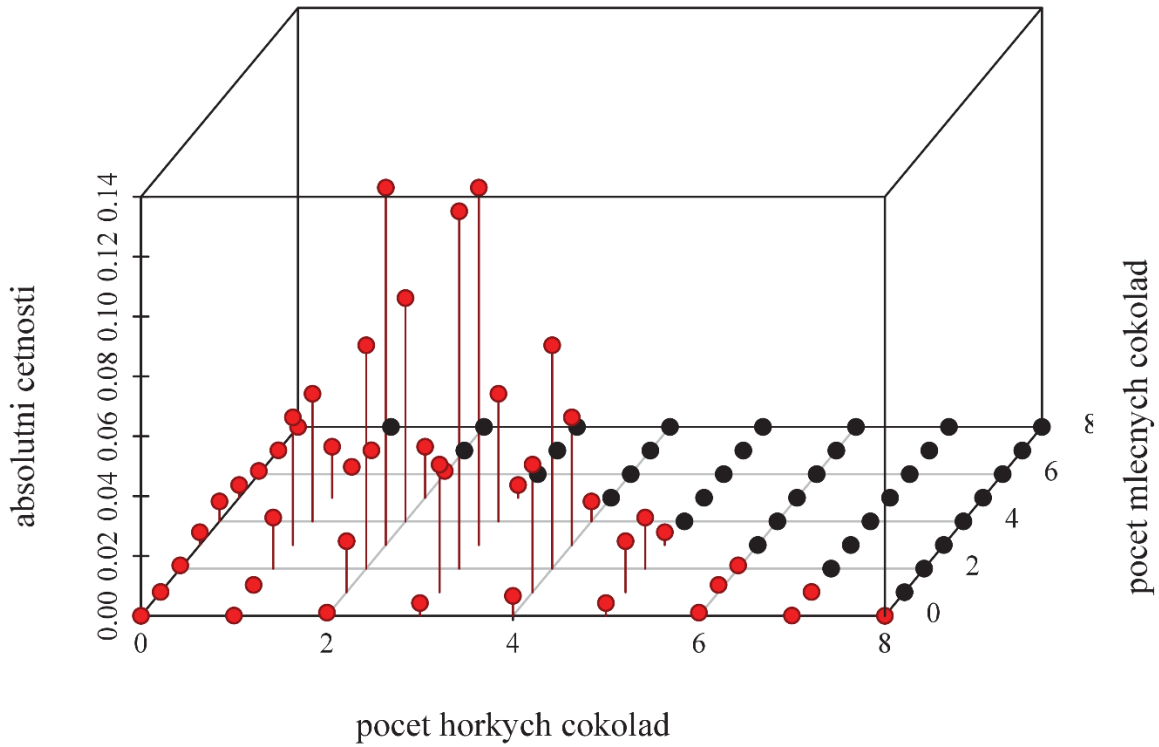
Pravděpodobnost, že Kája bude mít ve svém dílu dvě hořké, dvě bílé a čtyři mléčné čokolády je 7.46 %. Pravděpodobnost, že Kája bude mít čtyři mléčné a čtyři hořké čokolády je 0.67 %. Pravděpodobnost, že Kája bude mít mezi svými čokoládami maximálně dvě hořké, je 44.68 %. Pravděpodobnost, že více než 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 $\mathrm{MultiHyperGeom}(N, \mathbf{p})$, kde $N = 8$ a $\mathbf{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^T$. Vykreslete graf pravděpodobnostní funkce rozdělení $\mathrm{MultiHyperGeom}(N, \mathbf{p})$.

Řešení příkladu 6.3

```
25 M <- ... # pocet horkych, mlecnych a bilych cokolad v cele populaci
26 X <- rep(0:8, 9) # posl. 0, ..., 8, 0, ...,8, ..., 0, ..., 8
27 Y <- rep(0:8, rep(9, 9)) # posl. 0, ..., 0, 1, ..., 1, ... 8, ..., 8
28 fxy <- matrix(NA, ...) # priprava prazdne matice NA hodnot o dimenzi 9x9
29 for(i in 0:8){ # vnorene cykly; kazde dvojici x a y vypocitaji hodnotu p(x, y)
30   for(j in 0:8){
31     fxy[i + 1, j + 1] <- dmvhyper(c(i, j, 8 - (i + j)), M, 8)
32   }
33 }
34
35 s.obs <- cbind(...) # spojeni vektoru X, Y, a c(fxy) po sloupcich
36 barvy <- vypln <- rep('black', 81) # nastaveni cerne barvy vsech bodu
37 barvy[s.obs[, 3] != 0] <- 'darkred' # zmena barvy bodu s nenul. pstni fci na cervenou
38 vypln[s.obs[, 3] != 0] <- 'red' # zmena vyplni bodu s nenul. pstni fci na cervenou
```

```
39  library(...) # nacteni knihovny scatterplot3d
40  par(...) # nastaveni okraju grafu na 5, 4, 0, 1
41  scatterplot3d(s.obs, type = 'h', angle = 71, lwd = ..., pch = ..., color = ...,
42                bg = ..., xlab = ..., ylab = ..., zlab = ...) # graf pstni fce
43                 # rozdeleni MultiHyperGeom(8, (1/3, 1/3, 1/3))
```



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. $\mathbf{X} \sim \mathrm{MultiHyperGeom}(N, \mathbf{p})$, 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

Odhad vektoru parametrů $\mathbf{p}$

```
44  Npop <- ... # rozsah populace
45  M <- ... # pocet obyvatel z jednotlivych okresu
46  N <- ... # rozsah nahodneho vyberu
47  okresy <- ... # nazvy 7 okresu
48  p <- ... # vektor parametru p = (p1, ..., p7)
```

```
[1] 0.09172808 0.32023057 0.18676423 0.09771180 0.13018024 0.07724119 0.09614389
```

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

Rozložení počtu obyvatel z jednotlivých okresů v reprezentativním vzorku

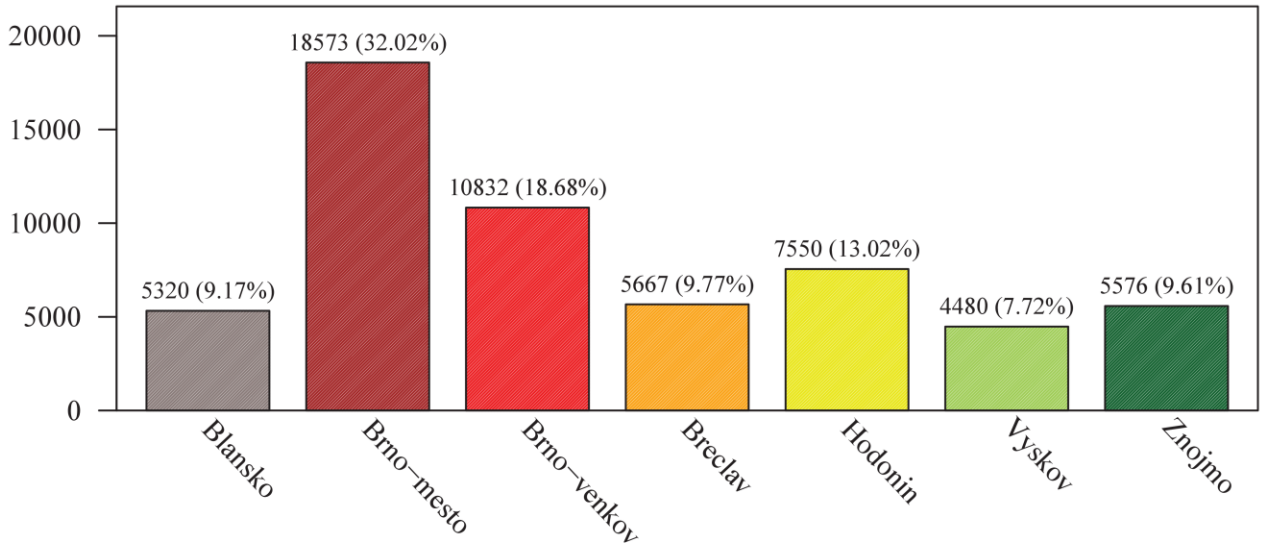
```
jihom.kraj <- function(...){ # funkce s povinnymi vstupnimi argumenty Npop, M, N
  p <- ... # vektor parametru p = (p1, ..., p7)
  pocet <- round(N %*% t(p)) # ocek. abs. cetnosti v repr. vzorku
  return(...)
}

tab <- jihom.kraj(...) # souhrnna tabulka ocek. abs. cetnosti v repr. vzorku
tab <- data.frame(tab, row.names = paste('N =', N)) # prevod tab na data.frame
tab <- cbind(..., apply(tab, 1, sum)) # doplneni sloupce radkovych souctu
names(...) <- c(okresy, ...) # zmena nazvu sloupcu tabulky tab (okresy + 'Sum')
```

	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

Sloupcový diagram očekávaných absolutních četností

```
67  N <- ... # rozsah nahodneho vyberu N (pro graf (a))
68  n <- c(jihom.kraj(...)) # ocek. pocetnosti v repr. vzorku N obyvatel JM kraje
69  col <- c(...) # vektor 7 barev sloupcu
70
71  par(...) # nastaveni okraju grafu na 6, 4, 1, 1
72  barplot(n, col = ..., axes = F, xlab = '', space = 0.3,
73          width = 1, xlim = c(0.4, 9), ylim = c(0, max(n) + 3000),
74          density = 80)# sl. diagram abs. cetnosti
75  text(seq(0.8, 9, by = 1.3), n + 1000,
76       paste(n, ' (', round(n / N, 4) * 100, '%)', sep = ''),
77       cex = 0.8) # popisky nad sloupci
78  box(...) # ramecek okolo grafu
79  text(seq(0.8, 9, by = 1.3), -700, okresy, cex.axis = 0.9, xpd = T, srt = -47,
80       adj = 0) # popisky sloupcu pod osou x otocene o 47° smerem dolu
81  axis(...) # osa y
```




```
82  N <- ... # rozsah nahodneho vyberu N (pro graf (b))
83  n <- c(jihom.kraj(...)) # ocek. pocetnosti v repr. vzorku N obyvatel JM kraje
84  barplot(...) # sl. diagram abs. cetnosti
85  text(...) # popisky nad sloupci
86  box(...) # ramecek okolo grafu
87  text(...) # popisky sloupcu pod osou x otocene o 47°smerem dolu
88  axis(...) # osa y
```