Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Téma 5: Hypergeometrický model

Veronika Bendová

bendova.veroonika@gmail.com

Hypergeometrické rozdělení $\mathrm{HyperGeom}(N, p)$

- Nechť N_{pop} je rozsah populace, M je počet statistických jednotek se sledovanou charakteristikou CH vyskytujících se v populaci N_{pop} a N je rozsah náhodného výběru vybraného z populace N_{pop} bez vrácení.

- X . . . počet statistických jednotek se sledovanou charakteristikou CH, vyskytujících se v náhodném výběru o rozsahu N.
- $X \sim \mathrm{HyperGeom}(N, p)$, kde $p = \frac{M}{N_{pop}}$
- $\theta = p$
- pravděpodobnostní funkce

$$p(x) = \frac{\binom{M}{x}\binom{N_{\mathrm{pop}}-M}{N-x}}{\binom{N_{\mathrm{pop}}}{N}}, \qquad x = 0, 1, \ldots \tag{3.1}$$

- vlastnosti: $\mathrm{E}[X] = Np$, $\mathrm{Var}[X] = Np(1-p)r$, kde

$$r = \frac{N_{\mathrm{pop}} - N}{N_{\mathrm{pop}} - 1} = 1 - \frac{N-1}{N_{\mathrm{pop}} - 1} > 1 - f_S, \qquad f_S = \frac{N}{N_{\mathrm{pop}}}$$

- f_S ... *výběrový poměr* (*sampling fraction*)
- je-li $f_S < 0.1$ (resp. $f_S < 0.05$), potom r zanedbáváme ($r \to 1$) a dochází k aproximaci náhodného výběru bez vrácení náhodným výběrem s vrácením, tedy k aproximaci hypergeometrického rozdělení binomickým rozdělením.

- dhyper(x, M, $\mathrm{N_{pop}}$ - M, N), phyper(x, M, $\mathrm{N_{pop}}$ - M, N), rhyper(N, M, $\mathrm{N_{pop}}$ - 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 5.1. Pravděpodobnostní funkce hypergeometrického modelu
Naprogramujte v R funkci dhypergeom(x, Npop, M, N) počítající hodnoty pravděpodobnostní funkce hypergeometrického rozdělení $\mathrm{HyperGeom}(N, p)$ v hodnotě x. Správnost funkce otestujte na výpočtu $p(x)$, $x = 45, 50, 53$, pro $X \sim \mathrm{HyperGeom}(N, p)$, kde $N = 70$ a $p = \frac{M}{N_{pop}} = \frac{240}{350}$. Výsledky ověřte s výsledky funkce dhyper().

Řešení příkladu 5.1

```
1  dhypergeom <- function(...){ # fce s povinnymi vstupnimi argumenty x, Npop, M, N
2    px <- ... # pstni fce rozdeleni HyperGeom(Npop, p); viz vzorec 5.1
3    return(...)
4  }
5  dhypergeom(...) # pstni fce rozdeleni HyperGeom(70, 240/350) (fce dhypergeom())
6  dhyper(...) # pstni fce rozdeleni HyperGeom(70, 240/350) (fce dhyper())
```

```
       p(45)      p(50)      p(53)                                          7
1 0.07752302 0.09869249 0.04164395                                          8
```

$p(45) = 0.0775$; $p(50) = 0.0987$; $p(53) = 0.0416$.

Příklad 5.2. Výpočet pravděpodobností na základě hypergeometrického modelu
Jana s Bárou a Kájou dostali hořko-mléčný adventní kalendář, ve kterém je polovina čokolád hořkých a polovina čokolád mléčných, přičemž 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, který vůbec nemá rád hořkou čokoládu, bude mít ve svém dílu (a) všechny čokolády mléčné; (b) maximálně dvě čokolády hořké; (c) více než polovinu čokolád mléčných.

Řešení příkladu 5.2
(a)

```
9  Npop <- ... # celkovy pocet cokolad cokolad
10 M <- ... # pocet mlecnych cokolad
11 N <- ... # rozsah nah. vyberu cokolad
12 p1 <- dhyper(...) # vypocet pravdepodobnosti
```

```
[1] 0.0006730381                                                    13
```

(b)

```
14  p2 <- 1 - phyper(...)  # vypocet pravdepodobnosti
```

```
[1] 0.09651366                                                    15
```

(c)

```
16  p3 <- sum(dhyper(...)) # vypocet pravdepodobnosti
17  tab <- data.frame(...) # souhrnna tabulka vysledku
```

```
[1] 0.3334231                                                     18
```

Pravděpodobnost, že všechny Kájovy čokolády budou mít mléčnou příchuť, je 0.07 %. Pravděpodobnost, že Kája bude mít mezi svými čokoládami maximálně dvě hořké, je 9.65 %. Pravděpodobnost, že více než polovina Kájových čokolád bude mléčných, je 33.34%.

Příklad 5.3. Odhad parametru p hypergeometrického modelu
Podle údajů uvedených v datasetu 5 má Jihomoravský kraj ke dni 30.6.2018 celkem 1 184 381 obyvatel, přičemž 379 275 obyvatel náleží do okresu Brno-město. Předpokládejme, že chceme sestavit reprezentativní vzorek 10-ti obyvatel pocházejících z Jihomoravského kraje. Pomocí hypergeometrického modelu charakterizujte chování náhodné veličiny X popisující počet obyvatel z okresu Brno-město v reprezentativním vzorku. Stanovte hodnoty parametrů N_{pop}, M a N, dopočítejte hodnotu parametru p rozdělení $\mathrm{HyperGeom}(N, p)$.

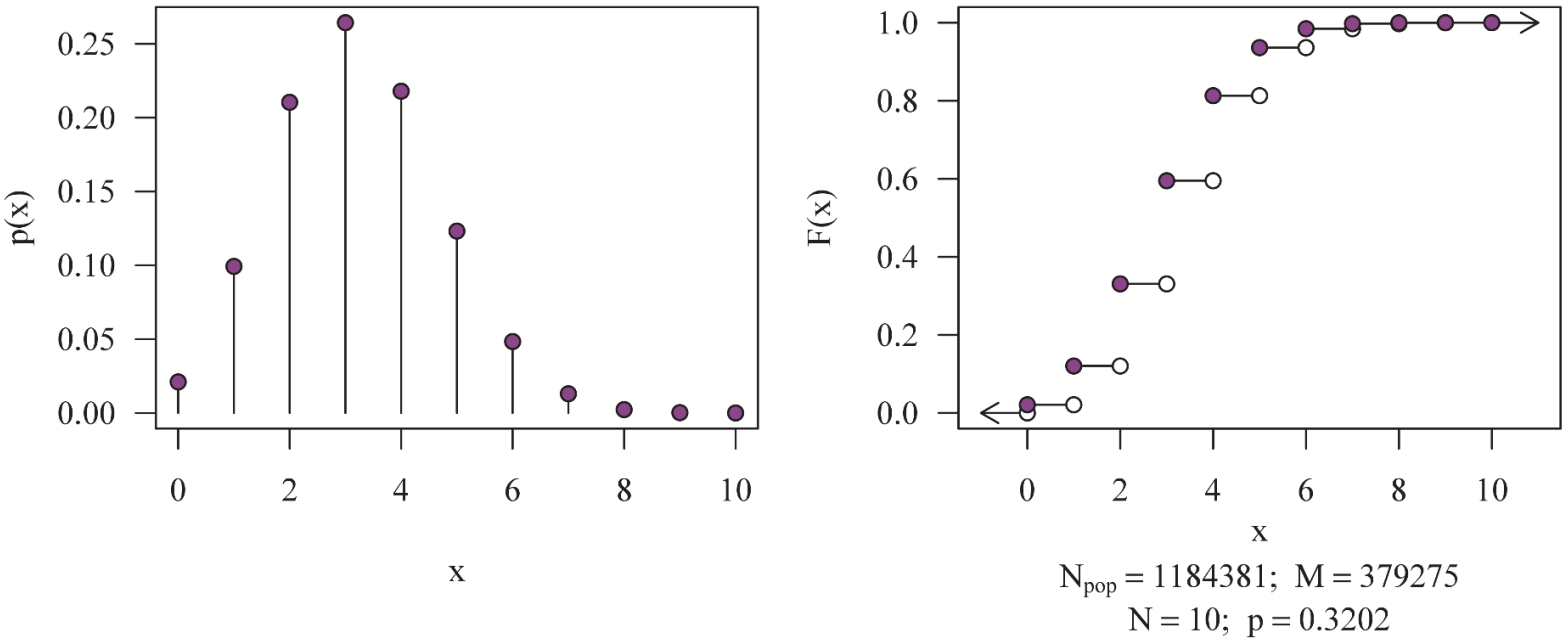
Řešení příkladu 5.3

```
19  Npop <- ... # rozsah populace
20  M <- ... # pocet statistickych jednotek se sledovanou charakteristikou
21  N <- ... # parametr N
22  p <- ... # odhad parametru p
```

Odhad parametru p je 0.3202, tj. náhodná veličina X pochází z rozdělení $\mathrm{HyperGeom}(10, 0.3202)$.

Příklad 5.4. Graf pravděpodobnostní a distribuční funkce hypergeometrického rozdělení
V příkladu 5.3 jsme stanovili, že počet obyvatel z okresu Brno-město v reprezentativním vzorku 10-ti obyvatel Jihomoravského kraje se bude řídit hypergeometrickým modelem $\mathrm{HyperGeom}(N, p)$, kde $N = 10$ a $p = 0.3202$. Vykreslete (a) graf pravděpodobnostní funkce; (b) graf distribuční funkce rozdělení $\mathrm{HyperGeom}(10, 0.3202)$. Na základě grafů určete, kolik obyvatel z reprezentativního vzorku bude s nějvětší pravděpodobností pocházet z okresu Brno-město a stanovte přesnou hodnotu této pravděpodobnosti.

Řešení příkladu 5.4



Obrázek: Pravděpodobnostní a distribuční funkce hypergeometrického modelu

S největší pravděpodobností (26.43%) budou v reprezentativním vzorku právě tři obyvatelé okresu Brno-město.

Příklad 5.5. Výpočet pravděpodobností na základě hypergeometrického modelu
Za předpokladu, že náhodná veličina X, udávající počet obyvatel z okresu Brno-město v reprezentativním vzorku 10-ti obyvatel Jihomoravského kraje, pochází z hypergeometrického rozdělení s parametry $N = 10$ a $p = 0.3202$, tj. $X \sim \mathrm{HyperGeom}(10, 0.3202)$ vypočítejte pravděpodobnost, že v reprezentativním vzorku budou (a) nejvýše tři obyvatelé z okresu Brno-město; (b) alespoň šest obyvatel z okresu Brno-město; (c) žádný obyvatel z okresu Brno město; (d) alespoň sedm obyvatel z jiného okresu; (e) nejvýše čtyři obyvatelé z jiného okresu; (f) všichni obyvatelé z jiného okresu.

Řešení příkladu 5.5
(a)

```
23 Npop <- ... # rozsah populace
24 M <- ... # pocet statistickych jednotek se sledovanou charakteristikou
25 N <- ... # parametr N
26 p1 <- ... # vypocet pravdpodobnosti
```

```
[1] 0.5950101
```
27

(b)

```
28 p2 <- ... # vypocet pravdpodobnosti
```

```
[1] 0.06392275
```
29

(c)

```
p3 <- ... # vypocet pravdpodobnosti
```

```
[1] 0.02106728
```

(d)

```
p4 <- ... # vypocet pravdpodobnosti
```

```
[1] 0.5950101
```

(e)

```
p5 <- ... # vypocet pravdpodobnosti
```

```
[1] 0.06392275
```

(f)

```
36  p6 <- ... # vypocet pravdpodobnosti
37  tab.B <- data.frame(...) # souhrnna tabulka vysledku - Brno-mesto
38  tab.O <- data.frame(...) # souhrnna tabulka vysledku - Ostatni okresy
```

```
[1] 0.02106728                                                          39
```

	nejvýše tři	alespoň šest	žádný
Brno-město	0.5950	0.0639	0.0211

	alespoň sedm	nejvýše čtyři	všichni
Ostatní okresy	0.5950	0.0639	0.0211

Nejvýše tři obyvatelé z okresu Brno-město budou v náhodném vzorku s pravděpodobností 59.50%. Alespoň šest obyvatel z okresu Brno-město budou v náhodném vzorku s pravděpodobností 6.39%. Pravděpodobnost, že v náhodném vzorku nebude žádný obyvatel okresu Brno město je 2.11%.

Naopak: Alespoň sedm obyvatel z jiného okresu než Brno-město budou v náhodném vzorku s pravděpodobností 59.50%. Nejvýše čtyři obyvatelé z jiného okresu než Brno-město budou v náhodném vzorku s pravděpodobností 6.39%. Pravděpodobnost, že v náhodném vzorku budou všichni obyvatelé z jiného okresu než Brno město je 2.11%.

Příklad 5.6. Aproximace hypergeometrického modelu binomickým – stanovení max. rozsahu
Prolog: Mějme populaci statistických jednotek o rozsahu N_{pop}*, přičemž pravděpodobnost výskytu statistické jednotky se sledovanou charakteristikou je* p*. Z populace* N_{pop} *vybereme náhodný výběr (bez vrácení) o rozsahu* N*. Náhodná veličina* X *popisuje počet statistických jednotek se sledovanou charakteristikou vyskytujících se v náhodném výběru. Potom* $X \sim \mathrm{HyperGeom}(N, p)$ *se střední hodnotou* $\mathrm{E}[X] = Np$ *a rozptylem* $\mathrm{Var}[X] = Np(1-p)r$*, kde* $r = \frac{N_{\mathrm{pop}}-N}{N_{\mathrm{pop}}-1} > 1 - f_S$*, přičemž* $f_S = \frac{N}{N_{\mathrm{pop}}}$ *je tzv. výběrový poměr. Je-li* $f_S < 0.1$ *(resp.* $f_S < 0.05$*),* r *zanedbáváme a hypergeometrické rozdělení* $\mathrm{HyperGeom}(N, p)$ *aproximujeme binomickým rozdělením* $\mathrm{Bin}(N, p)$*.*

Podle údajů uvedených v datasetu 5 má Jihomoravský kraj ke dni 30.6.2018 celkem 1 184 381 obyvatel, přičemž 379 275 obyvatel náleží do okresu Brno-město. Za předpokladu, že vybereme z populace Jihomoravského kraje náhodný výběr o rozsahu N, má náhodná veličina X popisující počet obyvatel z okresu Brno-město v náhodném výběru hypergeometrické rozdělení $\mathrm{HyperGeom}(N, p)$, $p = \frac{379\,275}{1\,184\,381} = 0.3202$. Zaměřme se nyní na aproximaci hypergeometrického modelu binomickým modelem. Stanovte nejprve maximální rozsah náhodného výběru $N_{\max}$, při kterém je ještě možné aproximovat data hypergeometrického modelu binomickým modelem ($f_S = 0.05$).

Řešení příkladu 5.6

```
40  fS <- ... # hodnota vyberoveho pomeru fs
41  N.opt <- ... # vypocet maximalniho rozsahu reprezentativniho vzorku
```

```
[1] 59219.05
```

42

Maximální rozsah reprezentativního vzorku, při kterém je ještě možné aproximovat hypergeometrický model binomickým modelem je 59 219.

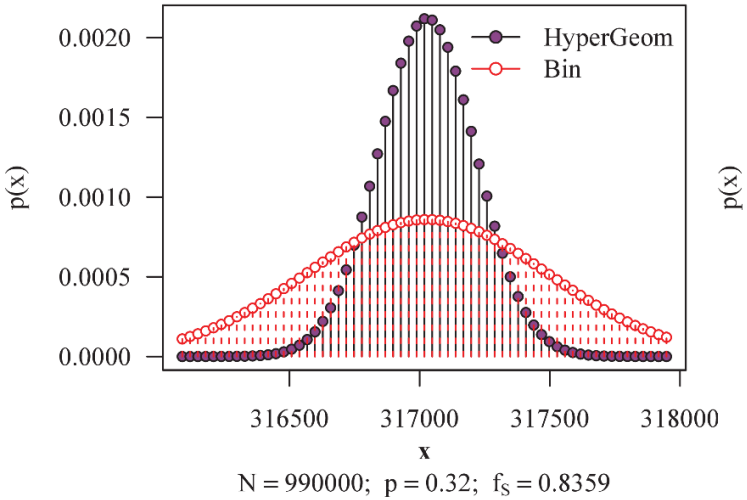
Příklad 5.7. Aproximace hypergeometrického modelu binomickým – graf pravěpodobnostní funkce V návaznosti na příklad 5.6 vykreslete nyní graf zachycující kvalitu aproximace pravděpodobnostní funkce hypergeometrického rozdělení pravděpodobnostní funkcí binomického rozdělení s parametry N a p, kde $p = 0.3202$. Hodnotu N zvolte (a) 990 000; (b) 590 000; (c) 59 219; (d) 5 900. Do grafu doplňte popisek zachycující hodnotu parametru N (rozsah reprezentativního vzorku), hodnotu parametru p a hodnotu výběrového poměru f_S.

Řešení příkladu 5.7

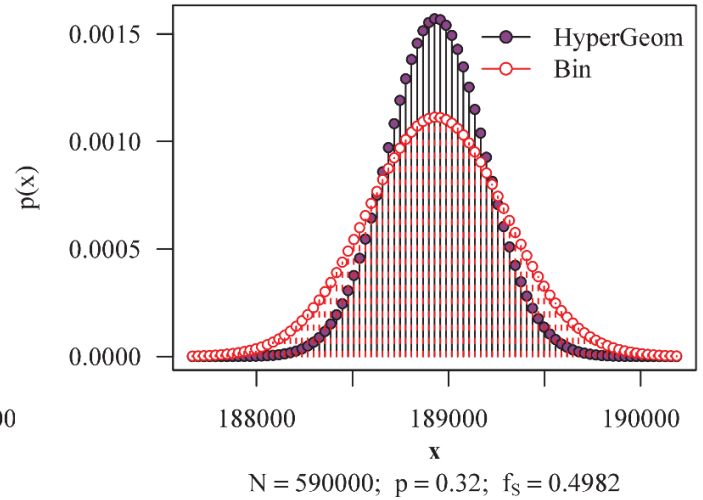
```r
hyper_cz <- function(...){ # fce s povinnymi argumenty M, Npop, N a volitelnymi
    # argumenty k = 100, res = F, plot = T.
  p <- ... # odhad parametru p
  fS <- ... # vypocet vyberoveho pomeru fs
  r <- ... # vypocet r
  EX <- ... # stredni hodnota E[X] rozd. HyperGeom(N, p)
  VarX <- ... # rozptyl Var[X] rozd. HyperGeom(N, p)
  VarY <- ... # rozptyl Var[X] rozd. Bin(N, p)
  x <- seq(from = round(EX - 5 * sqrt(VarX)), to = round(EX + 5 * sqrt(VarX)),
           by = k) # automaticky generovana posl. celych cisel x
  fx <- dhyper(...) # pstni fce rozd. HyperGeom(N, p) v hodnotach x
  fy <- dbinom(...) # pstni fce rozd. Bin(N, p) v hodnotach x

  if(plot == T){ # vykresleni grafu, pokud argument plot == T
  par(...) # nastaveni okraju grafu 5, 5, 1, 1
  plot(x, fx, ...) # cerne vert. cary rel. cetn. (HyperGeom(N, p))
  box(...) # ramecek okolo grafu
  lines(x, fy, ...) # cervene prerusovane vert. cary rel. cetn. (Bin(N, p))
  points(...) # tmave fialove body (HyperGeom(N, p))
  points(...) # cerveno-bile body (Bin(N, p))
  mtext(..., font = 2, ...) # popisek osy x
  mtext(bquote(paste(N == .(format(N, scientific = F )), ';  ', ...)),
        ...) # druhy popisek osy x
  mtext(...) # popisek osy y
  legend(...) # legenda
}
  results <- data.frame(N, ... , VarY)) # souhrnna tabulka vysledku
  if(res == T) return(results) # vypis tabulky vysledku, pokud argument res == T
}
```

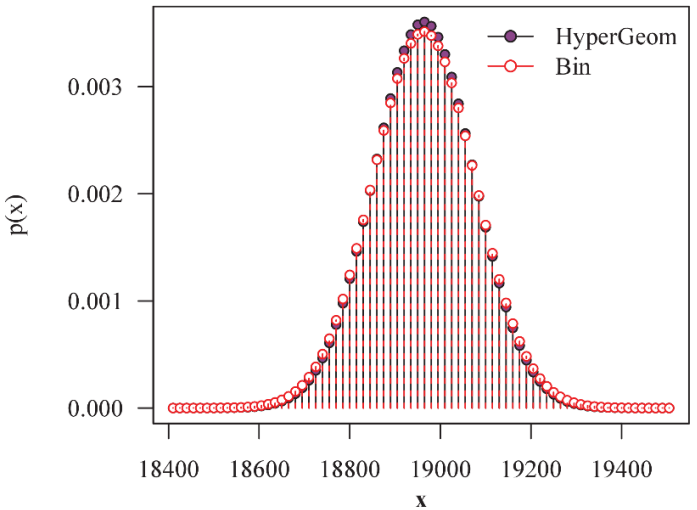
```r
M <- ... # pocet statistickych jednotek se sledovanou charakteristikou
Npop <- ... # rozsah populace
hyper_cz(M = ..., Npop = ..., N = ..., k = 30) # graf pro (a)
hyper_cz(..., k = 30) # graf pro (b)
hyper_cz(..., k = 15) # graf pro (c)
hyper_cz(..., k = 5) # graf pro (d)
```



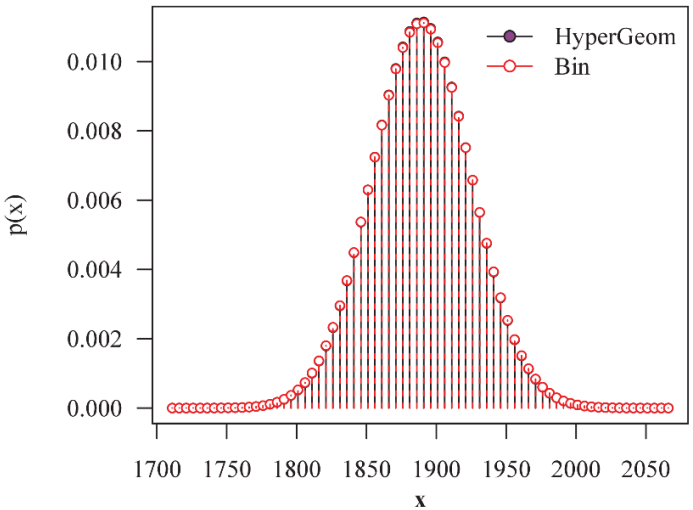
N = 990000; p = 0.32; f_S = 0.8359



N = 590000; p = 0.32; f_S = 0.4982



N = 59219; p = 0.32; f_S = 0.05



N = 5900; p = 0.32; f_S = 0.005

Příklad 5.8. Aproximace hypergeometrického modelu binomickým – výpočet charakteristik
V návaznosti na příklady 5.6 a 5.7 vytvořte přehlednou tabulku obsahující hodnoty N, N_{pop}, M, p, f_S, r, $\mathrm{E}[X]$, $\mathrm{E}[Y]$, $\mathrm{Var}[X]$ a $\mathrm{Var}[Y]$ pro každou variantu (a)–(d). $\mathrm{E}[X]$ a $\mathrm{Var}[X]$ značí střední hodnotu a rozptyl náhodné veličiny X z hypergeometrického rozdělení, $\mathrm{E}[Y]$ a $\mathrm{Var}[Y]$ značí střední hodnotu a rozptyl náhodné veličiny Y z binomického rozdělení.

Řešení příkladu 5.8

```
78  r1 <- hyper_cz(..., res = ..., plot = ...) # souhrnna tabulka vysledku pro (a)
79  r2 <- hyper_cz(...) # souhrnna tabulka vysledku pro (b)
80  r3 <- hyper_cz(...) # souhrnna tabulka vysledku pro (c)
81  r4 <- hyper_cz(...) # souhrnna tabulka vysledku pro (d)
82  tab <- rbind(...) # souhrnna tabulka vysledku pro (a)-(d)
```

N	N_{pop}	M	p	f_S	r	$\mathrm{E}[X]$	$\mathrm{E}[Y]$	$\mathrm{Var}[X]$	$\mathrm{Var}[Y]$
990000	1184381	379275	0.3202	0.8359	0.1641	317028	317028	35369	215506
590000	1184381	379275	0.3202	0.4982	0.5018	188936	188936	64454	128433
59219	1184381	379275	0.3202	0.0500	0.9500	18964	18964	12246	12891
5900	1184381	379275	0.3202	0.0050	0.9950	1889	1889	1278	1284