Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Téma 3: Negativně binomický model

Veronika Bendová

bendova.veroonika@gmail.com

Negativně binomické rozdělení $\mathrm{NegBin}(k, p)$

- Bernoulliho pokusy $X_1, X_2, \ldots$
  - $X_i = 1 \ldots$ událost nastala (úspěch); $X_i = 0 \ldots$ událost nenastala (neúspěch); $i = 1, 2, \ldots$
  - $\Pr(X_i = 1) = p$
  - $\Pr(X_i = 0) = 1 - p = q$
- $X \ldots$ počet nastalých úspěchů předcházejících předem stanovenému počtu k neúspěchů
- $X \sim \mathrm{NegBin}(k, p)$
- $\theta = (k, p)$
- pravděpodobnostní funkce

$$p(x) = \binom{x + k - 1}{x} p^x (1 - p)^k \qquad x = 1, 2, \ldots \tag{3.1}$$

- vlastnosti: $\mathrm{E}[X] = \dfrac{kp}{1 - p}$; $\mathrm{Var}[X] = \dfrac{kp}{(1 - p)^2}$
- dnbinom(n, k, 1 - p), pnbinom(n, k, 1 - p), rnbinom(M, k, 1 - p)

- **Dataset 2: Dělníci v továrně**
  - V rámci studie počtu úrazů v továrnách byl zaznamenán počet úrazů u každého dělníka v jedné vybrané továrně během roku 1920. Celkový počet dělníků zahrnutých do studie $M = 647$. Údaje ze studie jsou uvedeny v následující tabulce.

n	0	1	2	3	4	≥ 5	$\sum$
$m_{observed}$	447	132	42	21	3	2	647

Příklad 3.1. Pravděpodobnostní funkce negativně binomického modelu
Naprogramujte v R funkci **dNegBinom(x, k, p)** počítající hodnoty pravděpodobnostní funkce negativně binomického rozdělení $\mathrm{NegBin}(k, p)$ v hodnotě x. Správnost funkce otestujte na výpočtu $p(x)$, $x = 0, 1, 2$ pro $X \sim \mathrm{NegBin}(k, p)$, kde $k = 5$ a $p = 0.4$. Výsledky ověřte s výsledky funkce dnbinom() dostupné v R. Jaký je rozdíl v syntaxích obou funkcí?

Řešení příkladu 3.1

```
1  dNegBinom <- function(...){ # funkce s povinnymi vstupnimi argumenty x, k, p
2    px <- choose(...) * ... # pstni fce rozdeleni NegBinom(k, p); viz vzorec 3.1
3    return(...) # vystupem funkce bude hodnota ulozena v promenne px
4  }
5  dNegBinom(...) # vypocet pstni fce rozdeleni NegBin(5, 0.4) (fce dNegBinom())
6  dnbinom(...) # vypocet pstni fce rozdeleni NegBin(5, 0.4) (fce dnbinom())
```

```
       p0      p1       p2                                                    7
1 0.07776 0.15552 0.186624                                                    8
```

$p(0) = 0.7776$; $p(1) = 0.1555$; $p(2) = 0.1866$.

Příklad 3.2. Výpočet pravděpodobností na základě negativně binomického modelu
Pravděpodobnost sestřelení terče dosahuje u biatlonisky Koukalové až 95 %. Pravděpodobnost sestřelení terče u biatlonisky Charvátové se pohybuje okolo 65 %. Porovnejte, jaká je pravděpodobnost, že při tréninku sestřelí každá z biatlonistek před prvními dvěma neúspěchy (a) právě pět terčů; (b) nejvýše deset terčů; (c) alespoň sedm terčů; (d) alespoň patnáct terčů; (e) alespoň 25 terčů.

Řešení příkladu 3.2
(a)

```
9   p1K <- dnbinom(...) # vypocet pravdepodobnosti - Koukalova
10  p1C <- dnbinom(...) # vypocet pravdepodobnosti - Charvatova
```

(b)

```
11  p2K <- pnbinom(...) # vypocet pravdepodobnosti - Koukalova
12  p2C <- pnbinom(...) # vypocet pravdepodobnosti - Charvatova
```

(c), (d), (e)

```
13  p3K <- 1 - pnbinom(c(6, 14, 24), ...) # vypocet pravdepodobnosti - Koukalova
14  p3C <- 1 - pnbinom(...) # vypocet pravdepodobnosti - Charvatova

15  Koukalova <- c(...) # vektor promennych p1K, p2K a p3K
16  Charvatova <- c(...) # vektor promennych p1C, p2C a p3C
17  tab <- data.frame(rbind(...)) # tabulka vysledku
18  names(tab) <- c(...) # doplneni nazvu sloupcu do tabulky vysledku
19  round(...) # vypis tabulky vysledku (zaokrouhleny na 4 des. mista)
```

```
           prave 5 nejvyse 10 alespon 7 alespon 15 alespon 25   20
Koukalova   0.0116     0.1184    0.9428     0.8108     0.6241   21
Charvatova  0.0853     0.9576    0.1691     0.0098     0.0002   22
```

Biatlonistka Koukalová při tréninku před prvními dvěma neúspěchy sestřelí s pravděpodobností 1.16 % právě pět terčů, s pravděpodobností 11.84 % nejvýše 10 terčů, s pravděpodobností 94.28 % alespoň 7 terčů, s pravděpodobností 81.08 % alespoň 15 terčů a s pravděpodobností 62.41 % alespoň 25 terčů. Biatlonistka Charvátová při tréninku před prvními dvěma neúspěchy sestřelí s pravděpodobností 8.53 % právě pět terčů, s pravděpodobností 95.76 % nejvýše 10 terčů, s pravděpodobností 16.91 % alespoň 7 terčů, s pravděpodobností 0.98 % alespoň 15 terčů a s pravděpodobností 0.02 % alespoň 25 terčů.

Příklad 3.3. Výpočet očekávaných početností za předpokladu negativně binomického modelu
V příkladu 2.2 jsme popsali počet úrazů u dělníků v továrně pomocí náhodné veličiny X, která pocházela z Poissonova rozdělení s parametrem $\lambda = 0.4652$ odhadnutým na základě dat. Srovnáním pozorovaných početností m_{obs} s očekávanými početnostmi m_{exp}, za podmínky, že $X \sim \mathrm{Poiss}(0.4652)$ jsme diagnostikovali overdisperzi v datech oproti předpokládanému Poissonovu modelu. Hodnoty pozorovaných a očekávaných početností jsou pro připomenutí uvedeny v následující tabulce.

n	0	1	2	3	4	≥ 5	$\sum$
$m_{observed}$	447	132	42	21	3	2	647
$m_{exp.Poiss}$	406	189	44	7	1	0	647

V případě, že data vykazují po aproximaci Poissonovým rozdělením overdisperzi, bývá vhodné popsat data náhodnou veličinou X pocházející z negativně binomického rozdělení $X \sim \mathrm{NegBin}(k, p)$. Vypočítejte očekávané početnosti výskytu úrazů u dělníků v továrně za předpokladu, že početnosti úrazů pocházejí z negativně binomického rozdělení $\mathrm{NegBin}(k, p)$, kde $\widehat{k} = \frac{\widehat{E[X]}^2}{\widehat{\mathrm{Var}[X]} - \widehat{E[X]}}$ a $\widehat{p} = 1 - \frac{\widehat{E[X]}}{\widehat{\mathrm{Var}[X]}}$. Výsledné početnosti porovnejte s očekávanými početnostmi za předpokladu Poissonova rozdělení.

Řešení příkladu 3.3

```r
N <- ... # maximalni uvazovany pocet urazu u jednoho delnika
m.obs <- ... # posloupnost m.obs
M <- ... # celkovy pocet vsech delniku M
observed <- ... # vektor pozor. dat: 0, ..., 0, 1, ..., 1, ..., 5, 5
m <- mean(...) # odhad str. hodnoty E[X] ziskany na zaklade vektoru pozor. dat
v <- var(...) # odhad rozptylu Var[X] ziskany na zaklade vektoru pozor. dat
k <- ... # odhad parametru k dopocitany podle vzorce na slidu 6
p <- ... # odhad parametru p dopocitany podle vzorce na slidu 6
tab <- data.frame(...) # tabulka vysledkych parametru k a p
```

```
          k         p
1 0.9548138 0.3276139
```

```r
m.exp.p <- c(...) # vektor ocek. abs. cetnosti za predpokladu X ~ Poiss(0.4652)
m.exp.n <- round(c(dnbinom(...), 1 - pnbinom(...)) * ...) # vektor ocek. abs.
  # cetnosti za predpokladu rozdeleni NegBin(k, p)
tab <- data.frame(rbind(...)) # tabulka vysledku (pozor. i ocek. abs. cetnosti)
names(tab) <- 0:5
```

```
          0   1  2  3 4 5
m.obs   447 132 42 21 3 2
m.exp.p 406 189 44  7 1 0
m.exp.n 443 139 44 14 5 2
```

Odhad $\widehat{k} = 0.9548$, odhad $\widehat{p} = 0.3276$.

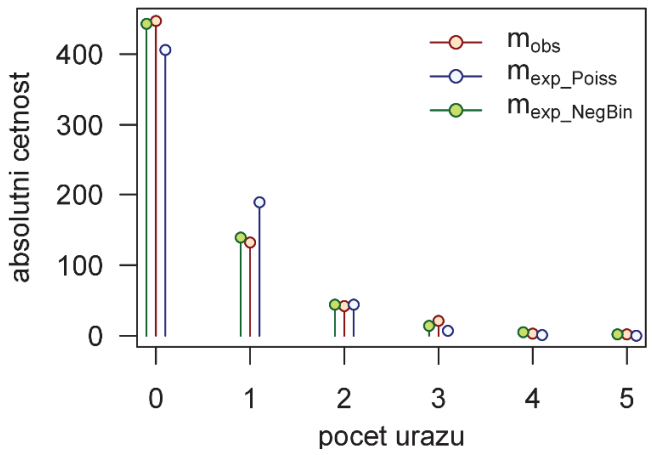
n	0	1	2	3	4	≥ 5	$\sum$
$m_{observed}$	447	132	42	21	3	2	647
$m_{exp.Poiss}$	406	189	44	7	1	0	647
$m_{exp.NegBin}$	443	139	44	14	5	2	647

Z tabulky vidíme, že očekávané početnosti vypočítané za předpokladu negativně binomického modelu lépe popisují chování pozorovaných početností úrazů u dělníků v továrně než početnosti vypočítané za předpokladu Poissonova modelu.

Příklad 3.4. Overdispersion a underdispersion v negativně binomickém modelu
V příkladu 3.3 jsme vypočítali hodnoty očekávaných absolutních početností úrazů dělníků v továrně za podmínky, že data pochází z $\mathrm{NegBin}(k, p)$, kde $\widehat{k} = 0.9548$, $\widehat{p} = 0.3276$. Do jednoho grafu zaneste nyní hodnoty pozorovaných početností $m_{observed}$, hodnoty očekávaných početností za předpokladu Poissonova rozdělení $m_{exp.Poiss}$ a hodnoty očekávaných početností za předpokladu negativně binomického rozdělení $m_{exp.NegBin}$. Na základě výsledného grafu stanovte, zda v případě použití negativně binomického modelu dochází k overdisperzi nebo underdisperzi. Závěr podložte srovnáním rozptylu vypočítaného z pozorovaných dat s rozptylem vypočítaným z očekávaných dat.

Řešení příkladu 3.4

```
43  par(...) # nastaveni okraju grafu 3, 4, 1, 1
44    plot (0:N, ..., type = ..., col = ..., xlab = ..., ylab = ...,
45          las = ...) # graf s vertikalnimi cervenymi carami ocek. abs. cetnosti
46  lines(0:N - 0.1, m.exp.n, ...) # zelene vert. cary pozor. abs. cetnosti (NegBin)
47  lines(0:N + 0.1, m.exp.p, ...) # modre vert. cary pozor. abs. cetnosti (Poiss)
48  points(...) # cervene body
49  points(...) # zelene body
50  points(...) # modre body
51  mtext(...) # popisek osy x
52  legend('topright', lty = ..., pch = ..., col = c(...), pt.bg = c(...),
53         legend = c(...), bty = ...) # legenda
```



Obrázek: Pozorované četnosti a očekávané četnosti za předpokladu negativně binomického a Poissonova modelu

Z vykresleného grafu je zřetelně viditelná blízkost pozorovaných početností a očekávaných početností za předpokladu negativně binomického modelu.

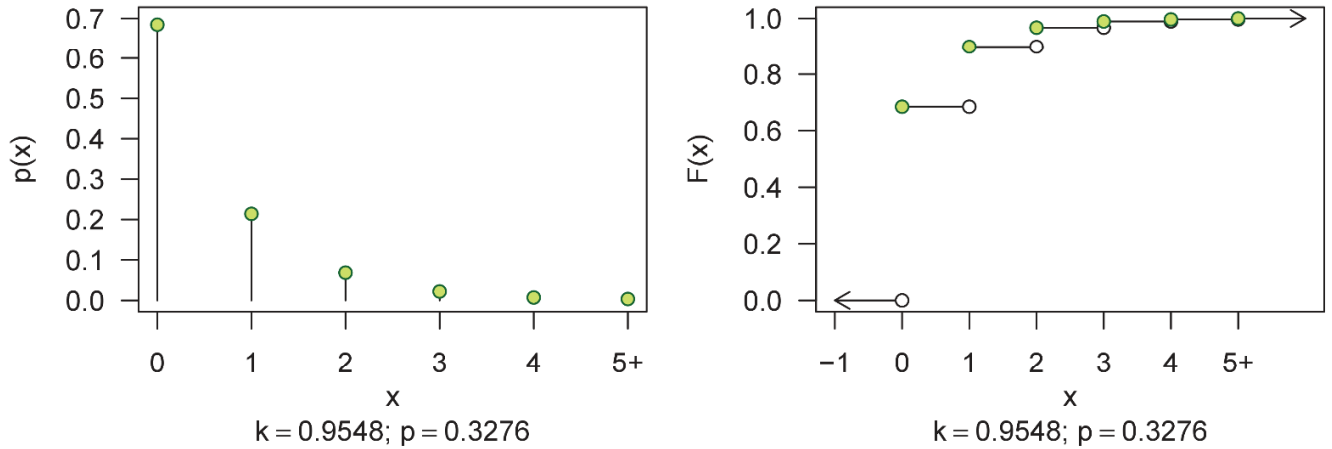
```
observed <- rep(...) # vektor pozor. dat: 0, ..., 0, 1, ..., 1, 5, 5
expected.n <- rep(...) # vektor ocek. dat (NegBin): 0, ..., 0, ..., 5, 5
(tab <- data.frame(Var.obs = var(...), Var.exp.n = ...)) # tabulka rozptylu
```

```
    Var.obs Var.exp.n
1 0.6919002 0.6700035
```

Hodnota rozptylu získaného z pozorovaných dat vyšla 0.6919, hodnota rozptylu získaného z očekávaných dat za předpokladu negativně binomického modelu vyšla 0.6700. Vidíme, že hodnoty rozptylů jsou si velmi blízké. V tomto případě tedy nedochází ani k overdisperzi ani k underdisperzi. Pro srovnání připomeňme, že hodnota rozptylu získaného z očekávaných dat za předpokladu Poissonova modelu vyšla 0.4691 (overdisperze; viz příklad 2.2).

Příklad 3.5. Graf pravděpodobnostní a distribuční funkce negativně binomického modelu
V příkladu 3.3 jsme odhadli hodnoty parametrů k a p negativně binomického rozdělení $\mathrm{NegBin}(k, p)$ jako $\widehat{k} = 0.9548$, $\widehat{p} = 0.3276$. Na základě tohoto rozdělení jsme vypočítali očekávané početnosti výskytu úrazů u dělníků v továrně. Nakreslete graf pravděpodobnostní a distribuční funkce negativně binomického rozdělení $\mathrm{NegBin}(k, p)$, kde $k = 0.9548$, $p = 0.3276$, v hodnotách $x = 0$, 1, 2, 3, 4 a $x \geq 5$.

Řešení příkladu 3.5



Obrázek: Pravděpodobnostní a distribuční funkce negativně binomického modelu

Příklad 3.6. Výpočet pravděpodobností na základě negativně binomického modelu
Za předpokladu, že náhodná veličina X, udávající počet úrazů u dělníků v továrně, pochází z negativně binomického modelu s parametry $k = 0.9548$ a $p = 0.3276$, tj. $X \sim \mathrm{NegBin}(k, p)$ vypočítejte pravděpodobnost, že u náhodně vybraného dělníka dojde během jednoho roku k (a) nula úrazům; (b) třem nebo čtyřem úrazům; (c) nejvýše dvěma úrazům; (d) alespoň jednomu úrazu. Výsledky porovnejte s pravděpodobnostmi vypočítanými v příkladu 2.6 za předpokladu Poissonova modelu.

Řešení příkladu 3.6
(a)

```
59  k <- ... # parametr k
60  p <- ... # parametr p
61  dnbinom(...) # vypocet pravdepodobnosti
```

```
[1] 0.6845545                                                    62
```

(b)

```
63  sum(dnbinom(...)) # vypocet pravdepodobnosti - prvni zpusob
64  pnbinom(...) - pnbinom(...) # vypocet pravdepodobnosti - druhy zpusob
```

```
[1] 0.02929255                                                   65
```

(c)

```
66 pnbinom(...) # vypocet pravdepodobnosti - prvni zpusob
67 sum(dnbinom(...)) # vypocet pravdepodobnosti - druhy zpusob
```

```
[1] 0.9672589
```
68

(d)

```
69 1 - pnbinom(...) # vypocet pravdepodobnosti
```

```
[1] 0.3154455
```
70

Pravděpodobnost, že u vybraného dělníka nedojde během roku k žádnému úrazu, je 68.46 %.
Pravděpodobnost, že u vybraného dělníka dojde během roku k třem nebo čtyřem úrazům, je 2.93 %.
Pravděpodobnost, že u vybraného dělníka dojde během roku k nejvýše dvěma úrazům, je 96.73 %.
Pravděpodobnost, že u vybraného dělníka dojde během roku k alespoň jednomu úrazu, je 31.54 %.

počet úrazů	žádný	tři nebo čtyři	nejvýše dva	alespoň jeden
NegBin	0.6846	0.0293	0.9673	0.3154
Poiss	0.6280	0.0118	0.9881	0.3720

Příklad 3.7. Střední hodnota a rozptyl náhodné veličiny z negativně binomického modelu
Za předpokladu, že náhodná veličina X, udávající počet úrazů u dělníků v továrně, pochází z negatovně binomického rozdělení, tj. $X \sim \mathrm{NegBin}(k, p)$, s parametry $k = 0.9548$, $p = 0.3276$, vypočítejte střední hodnotu $\mathrm{E}[X]$ a rozptyl $\mathrm{Var}[X]$ náhodné veličiny X. Střední hodnotu a rozptyl porovnejte s jejich odhady vypočítanými na (a) základě očekávaných dat; (b) na základě pozorovaných dat (viz příklad 3.3). Výsledky porovnejte s pravděpodobnostmi vypočítanými za předpokladu Poissonova modelu $\mathrm{Poiss}(\lambda)$, kde $\lambda = 0.4652$.

Řešení příkladu 3.7

```
71  E.X <- ... # stredni hodnota E[X] rozdeleni NegBin(k, p)
72  Var.X <- ... # rozptyl Var[X] rozdeleni NegBin(k, p)
73
74  expected <- ... # vektor ocek. dat: 0, ..., 0, 1, ...,1, ..., 5, 5
75  E.exp <- mean(...) # odhad stredni hodnoty na zaklade ocek. dat ()
76  Var.exp <- var(...) # odhad rozptylu na zaklade ocek. dat ()
77
78  observed.n <- ... # vektor pozor. dat (NegBin): 0, ..., 0, ..., 5, 5
79  E.obs <- ... # odhad stredni hodnoty na zaklade pozor. dat
80  Var.obs <- ... # odhad rozptylu na zaklade pozor. dat
81  (tab <- data.frame(...)) # tabulka vysledku
```

```
        E.X     Var.X     E.exp   Var.exp     E.obs   Var.obs                    82
1 0.4652241 0.6919002 0.4621329 0.6700035 0.4652241 0.6919002                    83
```

Střední hodnota počtu úrazů dělníků v továrně je 0.4652 s rozptylem 0.6919, odhad střední hodnoty počtu úrazů dělníků v továrně vypočítaný na základě očekávaných hodnot je 0.4652 s rozptylem 0.6919. Odhad střední hodnoty počtu úrazů dělníků v továrně vypočítaný na základě pozorovaných dat je 0.4621 s rozptylem 0.6700.

Tabulka: Porovnání odhadu střední hodnoty $E[X]$ a rozptylu $\mathrm{Var}[X]$ (a) z pozorovaných dat; (b) z očekávaných dat za předpokladu, že $X \sim \mathrm{NegBin}(k, p)$; (c) z očekávaných dat za předpokladu, že $X \sim \mathrm{Poiss}(\lambda)$

	$\widehat{\mathrm{E}[X]}$	$\widehat{\mathrm{Var}[X]}$
pozorovaná data	0.4652	0.6919
očekávaná - NegBin	0.4652	0.6700
očekávaná data - Poiss	0.4668	0.4691

Příklad 3.8. Simulační studie pro negativně binomický model a Poissonův model
Vytvořte simulační studii modelující chování očekávaných početností náhodné veličiny X popisující počet úrazů dělníků v továrně za předpokladu, že

(a) $X \sim \mathrm{Poiss}(\lambda)$, kde $\lambda = 0.4652$,

(b) $X \sim \mathrm{NegBin}(k, p)$, kde $k = 0.9548$ a $p = 0.3276$.

Vygenerujte pseudonáhodná čísla X (početnosti úspěchů) opakovaná M-krát ($M = 647$) pocházející (a) z Poissonova rozělení, tj. $X \sim \mathrm{Poiss}(\lambda)$; (b) z negativně binomického rozdělení, tj. $X \sim \mathrm{NegBin}(k, p)$. Pro každé rozdělení vytvořte histogram vygenerovaných pseudonáhodných čísel a superponujte jej hodnotami očekávaných (teoretických početností).

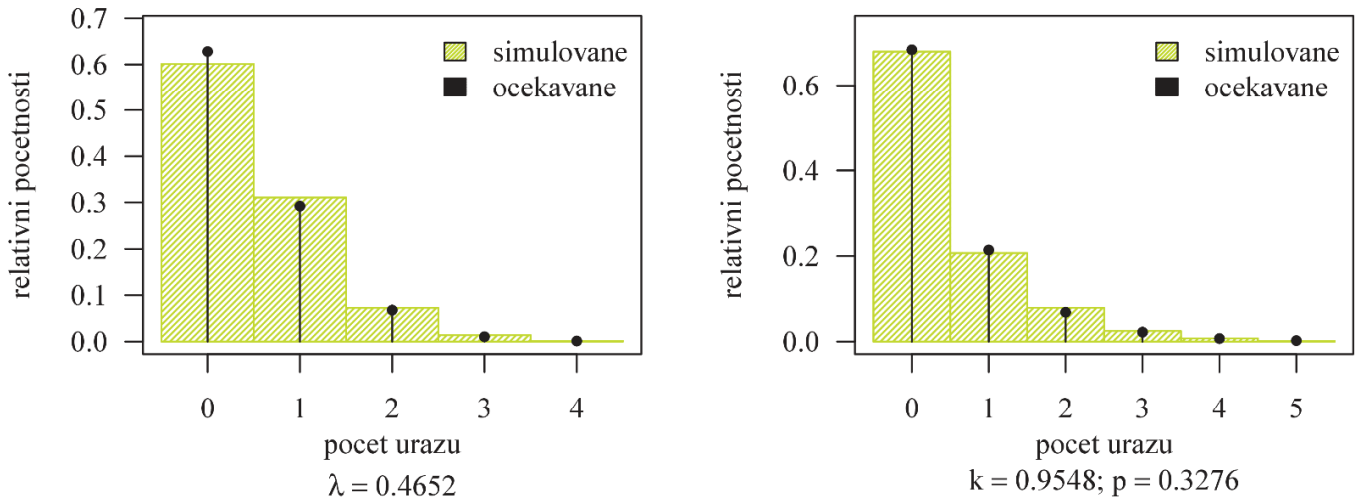
Řešení příkladu 3.8

Poissonovo rozdělení

```
84  N <- ... # maximalni uvazovany pocet urazu u jednoho delnika
85  M <- ... # celkovy pocet delniku M
86  lambda <- ... parametr lambda Poissonova rozdeleni
87
88  par(...) # nastaveni okraju 5, 4, 3, 2
89  X <- rpois(...) # vektor M pseudonahodnych cisel z rozd. Poiss(lambda)
90  px <- dpois(...) # pstni fce rozd. Poiss(lambda) v hodnotach 0, 1, ..., N
91  hist(X, prob = ..., breaks = seq(-0.5, max(X) + 0.5, by = 1), density = ...,
92       col = ..., ylim = c(0, max(px) + 0.05), xlab = ..., ylab = ...,
93       main = ..., las = ...) # histogram nah. vyberu X (v rel. skale)
94  box(...) # ramecek okolo grafu
95  lines (0:N, px, ...) # cerne vertikalni cary; pstni fce Poiss(lambda)
96  points(0:N, px, ...) # cerne plne body; pstni fce Poiss(lambda)
97  mtext(...) # popisek osy x
98  mtext(bquote(paste(lambda == .(lambda))), ...) # druhy popisek osy x: lambda =
99  legend(...) # legenda
```

Negativně binomické rozdělení

```
100  k <- ... # parametr k
101  p <- ... # parametr p
102  Y <- rnbinom(...) # vektor M pseudonahodnych cisel z rozd. NegBin(k, p)
103  py <- dnbinom(...) # pstni fce rozd. NegBin(k, p) v hodnotach 0, 1, ..., N
104  hist(Y, prob = ..., breaks = seq(-0.5, max(Y) + 0.5, by = 1),
105       ylim = c(0, max(py) + 0.05), ...) # histogram nah. vyberu Y (v rel. skale)
106  box(...) # ramecek okolo grafu
107  lines (...) # cerne vertikalni cary; pstni fce Poiss(lambda)
108  points(...) # cerne plne body; pstni fce Poiss(lambda)
109  mtext(...) # popisek osy x
110  mtext(bquote(paste(k == .(round(k, 4)), '; ', p == .(round(p, 4)))), ...)
111    # druhy popisek osy x: k = ...; p = ...
112  legend(...) # legenda
```



Obrázek: Porovnání pozorovaných četností a očekávaných četností za předpokladu (a) Poissonova modelu (vlevo); (b) negativně binomického modelu (vpravo)