Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Téma 2: Poissonův model

Veronika Bendová

bendova.veroonika@gmail.com

Poissonovo rozdělení $\mathrm{Poiss}(\lambda)$

- X ... počet událostí, které nastanou v jednotkovém časovém intervalu, přičemž k událostem dochází náhodně, jednotlivě a vzájemně nezávisle. Střední počet těchto událostí je vyjádřen parametrem $\lambda > 0$
- $X \sim \mathrm{Poiss}(\lambda)$
- $\theta = \lambda$
- pravděpodobnostní funkce

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0, 1, \ldots$$

- vlastnosti: $E[X] = \lambda$; $\mathrm{Var}[X] = \lambda$
- dpois(x, lambda), ppois(x, lambda), rpois(M, lambda)

- **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 2.1. Výpočet očekávaných početností za předpokadu Poissonova modelu
Vezměte údaje z **datasetu 2**. Vypočítejte očekávané početnosti výskytu úrazů u dělníků v továrně za předpokladu, že početnosti úrazů X mají Poissonovo rozdělení $\mathrm{Poiss}(\lambda)$ s parametrem

$$\widehat{\lambda} = \frac{\sum_{n=0}^{N} n m_{observed}}{\sum_{n=0}^{N} m_{observed}}. \tag{2.1}$$

Řešení příkladu 2.1

```
n <- ... # posloupnost 0, 1, ..., 5
m.obs <- ... # posloupnost m.obs, tj. 447, 132, ..., 2
M <- sum(...) # celkovy pocet vsech jednotek M (soucet vsech cisel v m.obs)
lambda <- sum(...) / sum(...) # odhad parametru lambda podle vzorce 2.1
```

```
[1] 0.4652
```

```
m.exp <- round(c(dpois(...), 1 - sum(dpois(...))) * ...) # vektor ocek. abs. cetnosti
```

```
        0   1  2 3 4 5
m.exp 406 189 44 7 1 0
```

```
sum(...) # kontrolni soucet ocek. abs. cetnosti m.exp
```

```
[1] 647
```

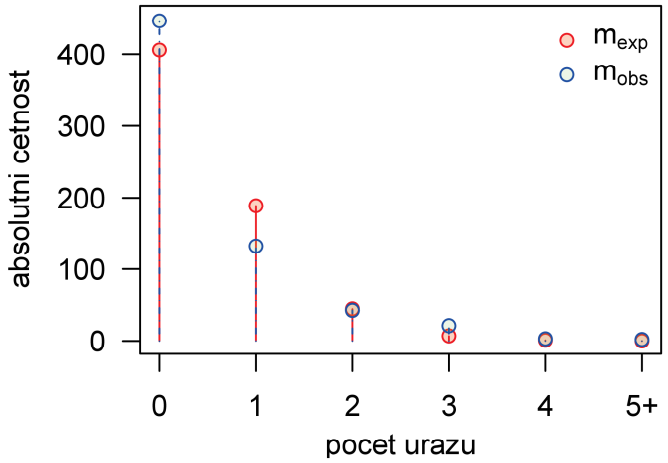
Odhad parametru λ, tj. $\widehat{\lambda} = 0.4652$. Tabulka očekávaných početností $m_{expected}$ je

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

Příklad 2.2. Overdispersion a underdispersion v Poissonově modelu
V předchozím příkladu 2.1 jsme stanovili očekávané početnosti výskytu úrazů u dělníků v továrně. Do jednoho grafu zaneste nyní hodnoty pozorovaných početností $m_{observed}$ a hodnoty očekávaných početností $m_{expected}$. Pozorované a očekávané početnosti od sebe barevně odlište. Na základě výsledného grafu stanovte, zda došlo k overdisperzi nebo underdisperzi. Závěr podložte srovnáním rozptylu vypočítaného z pozorovaných dat s rozptylem vypočítaným z očekávaných dat.

Řešení příkladu 2.2

```
11  par(mar = c(...)) # okraje grafu 3, 4, 1, 1
12  plot(n, m.exp, type = 'h', col = ..., xlab = ..., ylab = ..., las = 1,
13       axes = F) # graf s vertikalnimi cervenymi carami ocek. abs. cetnosti m.exp
14  box(bty = 'o') # ramecek okolo grafu
15  axis(1, 0:5, labels = c(0:4, '5+')) # osa x
16  axis(2, las = ...) # osa y
17  points(n, m.exp, pch = 21, col = ...,
18         bg = rgb(1, 0, 0, 0.2)) # cervene body abs. ocek. cetnosti m.exp
19  lines(n, m.obs, type = ..., col = ...,
20        lty = 4) # modre vertikalni cary pozor. abs. cetnosti m.obs
21  points(..., ..., pch = ..., col = ...,
22         bg = rgb(...)) # modre body pozor. abs. cetnosti m.obs
23  mtext(..., side = 1, line = 2.1) # popisek osy x
24  legend('topright', pch = c(..., ...), col = c(..., ...),
25         pt.bg = c(rgb(...), rgb(...)),
26         legend = c(expression(...), expression(...)), bty = 'n') # legenda
```



Obrázek: Porovnání pozorovaných a očekávaných početností v Poissonově modelu

Oproti očekávaným početnostem jsou pozorované početnosti krajních případů vyšší a naopak pozorované početnosti nejčastějších případů (0, 1) jsou nižší. Tato situace ukazuje na vyšší rozptyl pozorovaných početností než očekávaných. Jde tedy o .....disperzi.

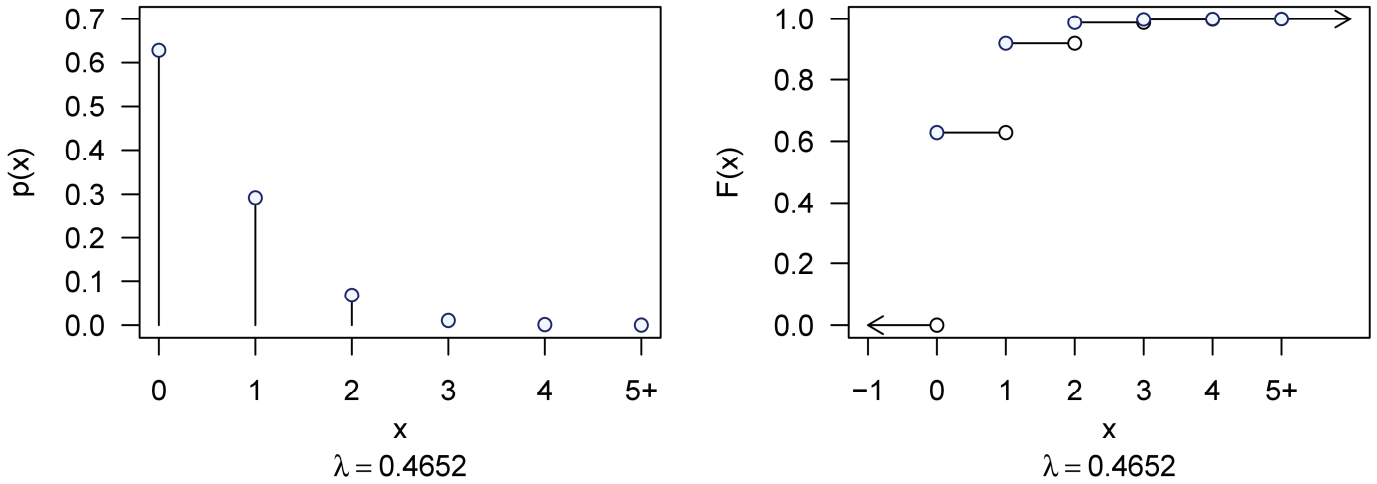
```
observed <- rep(...) # vektor pozor. dat: 0, ..., 0, 1, ..., 5, 5
expected <- ... # vektor ocek. dat: 0, ..., 0, 1, ..., 1, ..., 4
Var.obs <- ... # odhad rozptylu na zaklade pozor. dat
Var.exp <- ... # odhad rozptylu na zaklade ocek. dat
tab <- data.frame(...) # tabulka vysledku
```

```
    Var.obs   Var.exp
1 0.6919002 0.4690953
```

Hodnota rozptylu získaného z pozorovaných dat vyšla 0.6919, hodnota rozptylu získaného z očekávaných dat vyšla 0.4691. Vidíme, že hodnoty rozptylů i směrodatných odchylek se liší již v pozici na prvním desetinném místě. Hodnota rozptylu vypočítaného z pozorovaných dat je přibližně 1.5-krát vyšší než hodnota rozptylu vypočítaného z očekávaných dat.

Příklad 2.3. Graf pravděpodobnostní a distribuční funkce Poissonova modelu
V příkladu 2.1 jsme odhadli hodnotu parametru λ Poissonova rozdělení $\mathrm{Poiss}(\lambda)$ jako $\widehat{\lambda} = 0.4652$. Nakreslete graf pravděpodobnostní a distribuční funkce Poissonova rozdělení $\mathrm{Poiss}(\lambda)$, kde $\lambda = 0.4652$, v hodnotách $x = 0, 1, 2, 3, 4$ a $x \geq 5$.

Řešení příkladu 2.3



Obrázek: Pravděpodobnostní a distribuční funkce Poissonova modelu

Příklad 2.4. Výpočet pravděpodobností na základě Poissonova modelu
Za předpokladu, že náhodná veličina X, udávající počet úrazů u dělníků v továrně, pochází z Poissonova rozdělení, tj. $X \sim \mathrm{Poiss}(\lambda)$, s parametrem $\lambda = 0.4652$, 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.

Řešení příkladu 2.4
(a)

```
34  lambda <- ... # hodnota parametru lambda
35  dpois(...) # vypocet pravdepodobnosti
```

```
[1] 0.6280095
```
36

(b)

```
37  sum(dpois(...)) # vypocet pravdepodobnosti - prvni zpusob
38  ppois(...) - ppois(...) # vypocet pravdepodobnosti - druhy zpusob
```

```
[1] 0.01176292
```
39

(c)

```
40 ppois(...) # vypocet pravdepodobnosti
```

```
[1] 0.9881136                                                    41
```

(d)

```
42 1 - ppois(...) # vypocet pravdepodobnosti
```

```
[1] 0.3719905                                                    43
```

Pravděpodobnost, že u vybraného dělníka nedojde během roku k žádnému úrazu, je 62.80 %.
Pravděpodobnost, že u vybraného dělníka dojde během roku k třem nebo čtyřem úrazům, je 1.18 %.
Pravděpodobnost, že u vybraného dělníka dojde během roku k nejvýše dvěma úrazům, je 98.81 %.
Pravděpodobnost, že u vybraného dělníka dojde během roku k alespoň jednomu úrazu, je 37.20 %.

Příklad 2.5. Střední hodnota a rozptyl náhodné veličiny z Poissonova modelu
Za předpokladu, že náhodná veličina X, udávající počet úrazů u dělníků v továrně, pochází z Poissonova rozdělení, tj. $X \sim \mathrm{Poiss}(\lambda)$, s parametrem $\lambda = 0.4652$, 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 2.1).

Řešení příkladu 2.5

```
44  E.X <- ... # vypoct stredni hodnoty E[X] rozdeleni Poiss(lambda)
45  Var.X <- ... # vypoct rozptylu Var[X] rozdeleni Poiss(lambda)
46
47  expected <- rep(...) # vektor ocek. dat: 0, ..., 0, ..., 4, 4, 4, 5, 5
48  E.exp <- mean(...) # odhad stredni hodnoty na zaklade ocek. dat
49  Var.exp <- var(...) # odhad rozptylu na zaklade ocek. dat
50
51  observed <- ... # vektor pozor. dat: 0, ..., 0, ..., 4
52  E.obs <- ... # odhad stredni hodnoty na zaklade pozor. dat
53  Var.obs <- ... # odhad rozptylu na zaklade pozor. dat
54
55  (tab <- data.frame(E.X, Var.X, ...)) # tabulka vysledku
```

```
     E.X  Var.X     E.exp   Var.exp     E.obs   Var.obs                    56
1 0.4652 0.4652 0.4667697 0.4690953 0.4652241 0.6919002                    57
```

Střední hodnota počtu úrazů dělníků v továrně je 0.4652 s rozptylem 0.4652, odhad střední hodnoty počtu úrazů dělníků v továrně vypočítaný na základě očekávaných hodnot je 0.4668 s rozptylem 0.4691. Odhad střední hodnoty počtu úrazů dělníků v továrně vypočítaný na základě pozorovaných dat je 0.4652 s rozptylem 0.6919.

Příklad 2.6. Simulační studie: Součet náhodných veličin z Poissonova modelu
Věta 1: Nechť $X_1, X_2, \ldots, X_n$, jsou vzájemně nezávislé náhodné veličiny pocházenící z Poissonova rozdělení, tj. $X_i \sim \mathrm{Poiss}(\lambda_i)$, $i = 1, \ldots, n$. Potom náhodná veličina $X = \sum_{i=1}^{n} X_i \sim \mathrm{Poiss}(\sum_{i=1}^{n} \lambda_i)$.

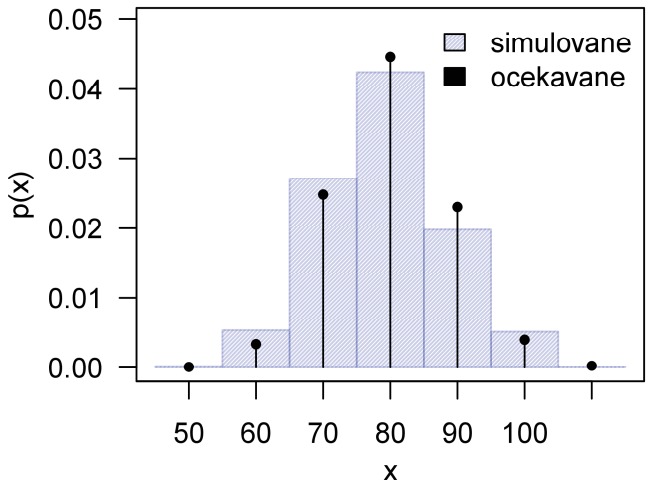
Na základě simulační studie ověřte platnost věty 1. Vygenerujte tři nezávislé náhodné veličiny $X_i \sim \mathrm{Poiss}(\lambda_i)$, $i = 1, 2, 3$, kde $\lambda_1 = 10$, $\lambda_2 = 20$, $\lambda_3 = 50$. Pomocí simulační studie ($M = 1000$) ukažte, že $\sum_{i=1}^{3} X_i \sim \mathrm{Poiss}(\lambda_1+\lambda_2+\lambda_3)$. Součty $M = 1000$ náhodných veličin zobrazte pomocí histogramu (hranice třídicích intervalů nastavte tak, aby šířka každého intervalu byla rovná 10) a superponujte jej hodnotami pravděpodobnostní funkce Poissonova rozdělení s parametrem $\lambda = \lambda_1 + \lambda_2 + \lambda_3$. Hodnoty pravděpodobnostní funkce vykreslete vždy ve středu každého třídicího intervalu.

Řešení příkladu 2.6

```
58  M <- ... # pocet simulaci M
59  lambda1 <- ... # parametr lambda1
60  lambda2 <- ... # parametr lambda2
61  lambda3 <- ... # parametr lambda3
```

```r
X <- replicate(M, sum(rpois(1, lambda1), rpois(1, lambda2), rpois(1, lambda3)))
  # 1000 souctu nah. velicin X1 + X2 + X3, kde X1 ~ Poiss(lambda1), ...
b <- seq(...) # hranice tridicich intervalu; posl. od 45 do 115 po kroku 10
centr <- seq(...) # stredy tridicich intervalu; posl. od 40 do 110 po kroku 10
y <- dpois(...) # pstni fce rozd. Poiss(lam1 + lam2 + lam3) ve stredech trid.int.

par(mar = ...) # nastaveni okraju 3, 4, 1, 1
hist(X, prob = ..., axes = ..., breaks = ..., density = ...,
     col = ..., ylim = c(0, max(y) + 0.005), main = ...,
     xlab = ..., ylab = ...) # histogram nah. vyberu X (v rel. skale)
box(...) # ramecek okolo grafu
axis(..., centr) # osa x
axis(...) # osa y
points(...) # cerne plne body
lines(...) # vertikalni cerne cary
mtext(...) # popisek osy x
legend(..., fill = c(...), density = c(...),
       legend = c(...), bty = ...) # legenda
```



Obrázek: Porovnání rozdělení součtu $\sum_{i=1}^{3} X_i$, kde $X_i \sim \mathrm{Poiss}(\lambda_i)$, $i = 1, 2, 3$, s rozdělením $\mathrm{Poiss}(\lambda_1 + \lambda_2 + \lambda_3)$

Příklad 2.7. Aproximace binomického modelu Poissonovým modelem
Poissonův model $\mathrm{Poiss}(\lambda)$ je limitním případem binomického modelu $\mathrm{Bin}(N, p)$, tj. pro $N \to \infty$, $p \to 0$, $Np \to \lambda$ platí

$$X \sim \mathrm{Bin}(N, p) \to X \sim \mathrm{Poiss}(\lambda).$$

Odvoďte vztah aproximace binomického modelu Poissonovým rozdělením.

Řešení příkladu 2.7

$$\begin{aligned}\binom{N}{x} p^x (1-p)^{N-x} &= \frac{N!}{x!(N-x)!} p^x (1-p)^N (1-p)^{-x} \frac{N^x}{N^x} = \\ &= \frac{(Np)^x}{x!} \left(1 - \frac{Np}{N}\right)^N (1-p)^{-x} \frac{N!}{(N-x)!N^x} = \\ &= \frac{\lambda^x}{x!} e^{-\lambda},\end{aligned}$$

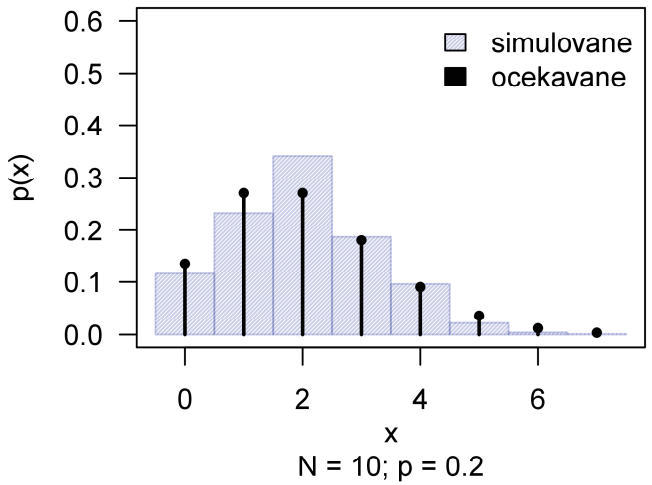
protože $\lim_{Np \to \lambda} (Np)^x = \lambda^x$, $\lim_{Np \to \lambda, N \to \infty} \left(1 - \frac{Np}{N}\right)^N = e^{-\lambda}$, $\lim_{p \to 0} (1-p)^{-x} = 1$, $\lim_{N \to \infty} \frac{N!}{(N-x)!N^x} = 1$.

Příklad 2.8. Aproximace binomického modelu Poissonovým modelem
Jak jsme si ukázali v příkladu 2.7, Poissonovo rozdělení je limitním případem binomického rozdělení pro $N \to \infty$, $p \to 0$ a $Np \to \lambda$. Pomocí simulační studie ověříme toho tvrzení. Vygenerujte pseudonáhodná čísla X (četnosti úspěchů) opakovaná M-krát ($M = 1000$) z $\mathrm{Bin}(N, p)$, kde $N = 10$ a $p = 0.2$. Vykreslete histogram vygenerovaných pseudonáhodných čísel a superponujte jej hodnotami očekávaných početností za předpokladu $X \sim \mathrm{Poiss}(\lambda)$, kde $\lambda = Np = 5 \times 0.2 = 1$.

1. Pomocí animace ukažte, jak se s $p \to 0$ zlepšuje aproximace binomického rozdělení Poissonovým rozdělením. Hodnotu parametru N zvolte pevně $N = 10$ a za p dosazujte hodnoty 0.8, 0.75, 0.7, . . . , 0.05.
2. Pomocí animace ukažte, jak se s $N \to \infty$ zlepšuje aproximace binomického rozdělení Poissonovým rozdělením. Hodnotu parametru p zvolte pevně $p = 0.2$ a za hodnoty parametru N dosazujte $N = 3, 4, 5, \ldots, 23, 24, 25, 30, 40, 50$.

Řešení příkladu 2.8

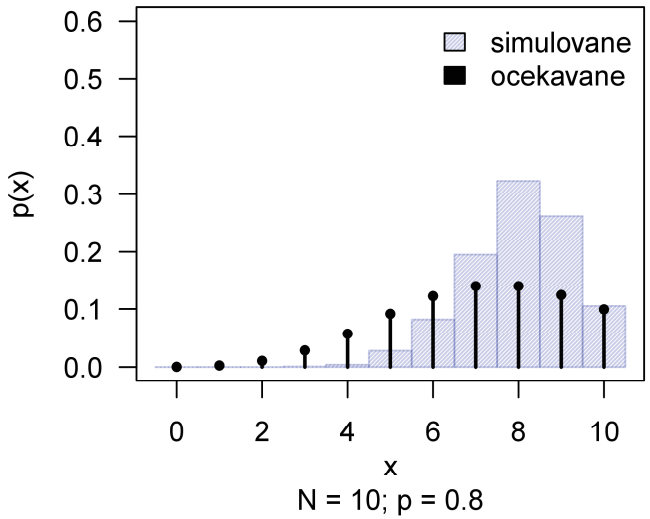
```r
a.poiss  <- function(N, p){
  # N <- 10, p <- 0.2
  X <- rbinom(...) # 1000 pseudonahodnych cisel z rozd. Bin(N, p)
  lambda <- ... # vyjadreni parametru lambda pomoci parametru N a p
  px.pois <- dpois(...) # pstni fce Poiss(lambda) v hodnotach 0, 1, ..., N
  par(...) # nastaveni okraju 4, 4, 1, 1
  hist(X, prob = ..., breaks = seq(-0.5, max(X) + 0.5, by = 1), density = ...,
       col = rgb(...), ylim = c(0, 0.6), las = ..., main = ...,
       xlab = ..., ylab = ...) # histogram nah. vyberu X (v rel. skale)
  box(...) # ramecek okolo grafu
  points(...) # cerne plne body; pstni fce Poiss(lambda)
  lines (...) # # cerne verikalni cary; pstni fce Poiss(lambda)
  mtext(...) # popisek osy x
  mtext(...) # druhy popisek osy x: N = ..., p = ...
  legend(..., fill = c(...), density = c(...),
         legend = c(...), bty = ...) # legenda
}
a.poiss(N = 10, p = 0.2) # kontrola funkcnosti fce a.poiss()
```



Obrázek: Aproximace binomického rozdělení Poissonovým rozdělením; N = 10, p = 0.2

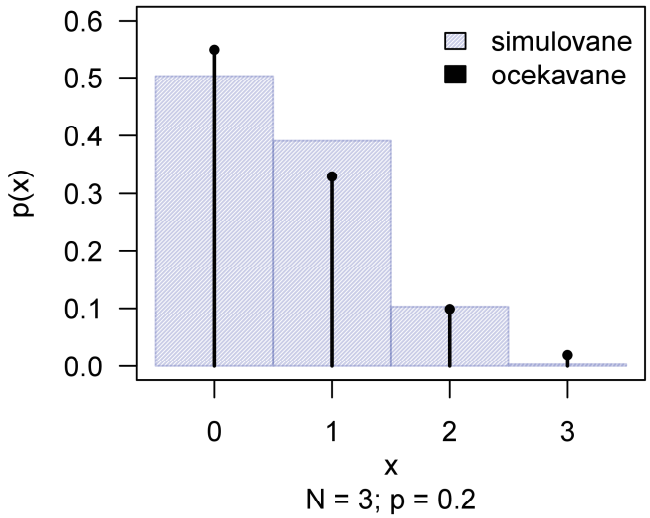
```
p <- seq(...) # posl. parametru p; od 0.8 do 0.01 po kroku -0.05
library(animation) # nacteni knihovny animation
oopts <- ani.options() # ulozeni hodnot defaultniho nastaveni parametru animace
ani.record(reset = T) # vymazani neaktualnich grafu z pameti Rka

saveLatex( # vygeneruje animaci
  for(i in 1:length(p)){ # cyklus; prochazi i od 1 az po delku vektoru p
      a.poiss(N = 10, p[i]) # provede fci a.poiss() pro kazdou hodnotu posl. p
      ani.record(reset = F) # uchova vsechny grafy od zacatku cyklu v pameti
  },
  latex.filename = 'aprox-bin-poiss-p.tex', # nazev .tex (i .pdf) souboru
  img.name = 'Rplot-aprox-bin-poiss-p') # nazev vygenerovanych obrazku
    # (za uvedeny nazev se vzdy pripoji cislo obrazku)
```



Obrázek: Aproximace binomického rozdělení Poissonovým rozdělením pro $p \to 0$

```
111  N <- c(3:25, 30, 40, 50) # posl. parametru N; 3, 4, ..., 25, 30, 40, 50
112  oopts <- ... # ulozeni hodnot defaultniho nastaveni parametru animace
113  ani.record(...) # vymazani neaktualnich grafu z pameti Rka
114  saveLatex(
115    for(i in 1:length(N)){
116      a.poiss(...) # provede fci a.poiss() pro kazdou hodnotu posl. N (p = 0.2)
117      ani.record(...) # uchova vsechny grafy od zacatku cyklu v pameti
118      },
119    latex.filename = 'aprox-bin-poiss-N.tex',
120    img.name = 'Rplot-aprox-bin-poiss-N')
```



Obrázek: Aproximace binomického rozdělení Poissonovým rozdělením pro $N \to \infty$

- **Dataset 3: Pruské armádní jednotky**
  - V rámci studie z roku 1898 byly zpracovávány počty smrtelných úrazů v pruských armádních jednotkách způsobené kopnutím koněm. Údaje o smrtelných úrazech po kopnutí koněm by zaznamenávány po dobu dvaceti let u deseti armádních jednotek. Počty úrazů v každé jednotce za jeden rok jsou uvedeny v následující tabulce.

n	0	1	2	3	4	5+	$\sum$
$m_{observed}$	109	65	22	3	1	0	200

  Rozsah náhodného výběru je $M = 200$ (10 jednotek $\times$ 20 let).

Příklad 2.9. Výpočet očekávaných početností za předpokladu Poissonova modelu
Vezměte údaje z **datasetu 3**. Vypočítejte očekávané početnosti výskytu smrtelných úrazů způsobených kopnutím koněm za předpokladu, že početnosti úrazů X mají Poissonovo rozdělení $\mathrm{Poiss}(\lambda)$ s parametrem

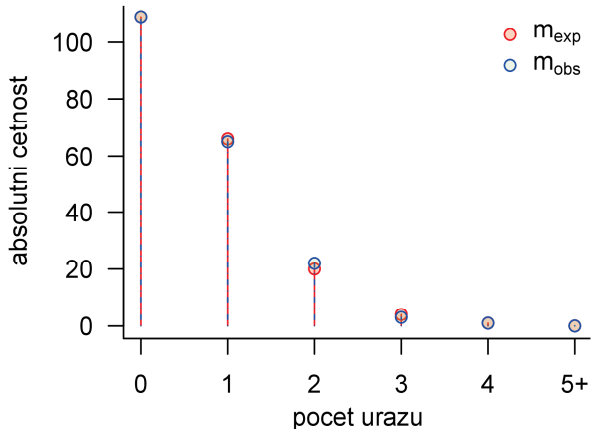
$$\widehat{\lambda} = \frac{\sum_{n=0}^{N} n m_{observed}}{\sum_{n=0}^{N} m_{observed}}. \tag{2.2}$$

Řešení příkladu 2.9
Odhad parametru λ, tj. $\widehat{\lambda} = 0.61$. Tabulka očekávaných početností $m_{expected}$ je

n	0	1	2	3	4	≥ 5	$\sum$
$m_{expected}$	109	66	20	4	1	0	200

Příklad 2.10. Overdispersion a underdispersion v Poissonově modelu
V předchozím příkladu 2.9 jsme stanovili očekávané početnosti výskytu smrtelných úrazů v pruských armádních jednotkách. Do jednoho grafu zaneste hodnoty pozorovaných početností $m_{observed}$ a hodnoty očekávaných početností $m_{expected}$. Pozorované a očekávané početnosti od sebe barevně odlište. Na základě výsledného grafu stanovte, zda došlo k overdisperzi nebo underdisperzi. Závěr podložte srovnáním rozptylu vypočítaného z pozorovaných dat s rozptylem vypočítaným z očekávaných dat.

Řešení příkladu 2.10



Obrázek: Porovnání pozorovaných a očekávaných početností v Poissonově modelu

Z tabulky očekávaných početností a z grafu vidíme, že očekávané početnosti se od pozorovaných příliš neliší. V tomto případě tedy nedochází ani k overdisperzi, ani k underdisperzi. Naše tvrzení podpoříme srovnáním rozptylu vypočítaného z porozovaných dat s rozptylem vypočítaným z očekávaných dat.

```
    Var.obs   Var.exp
1 0.6109548 0.621005
```

Hodnota rozptylu získaného z pozorovaných dat vyšla 0.6110, hodnota rozptylu získaného z očekávaných dat vyšla 0.6210. Vidíme, že hodnoty rozptylů se liší v pozici na druhém desetinném místě, hodnoty směrodatných odchylek se liší v pozici na třetím desetinném místě.