Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Aplikovaná statistika I

Téma 4: Diskrétní náhodné veličiny

Veronika Bendová

bendova.veroonika@gmail.com

Úvod a motivace

- motivace: reálná situace (data) $\rightarrow$ popíšeme ji nějakým známým rozdělením $\rightarrow$ z dat odhadneme parametry rozdělení $\rightarrow$ stanovíme nové závěry na základě vlastností rozdělení
- různé typy dat $\rightarrow$ různé typy rozdělení
  - diskrétní data $\rightarrow$ diskrétní rozdělení
    - binomické rozdělení ... $\mathrm{Bin}(N, p)$
    - alternativní rozdělení ... $\mathrm{Alt}(p)$
    - Poissonovo rozdělení ... $\mathrm{Poiss}(\lambda)$
  - spojitá data $\rightarrow$ spojité rozdělení
    - normální rozdělení ... $N(\mu, \sigma^2)$
    - standardizované normální $N(0, 1)$
    - dvourozměrné normální $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $+$ (spojitá) rozdělení testovacích statistik
    - Pearsonovo chi-kvadrátové rozděení ... $\chi^2(n)$
    - Studentovo t-rozdělení ... $t(n)$
    - Fisherovo-Snedecorovo F-rozdělení ... $F(n_1, n_2)$

Základy pravděpodobnosti

- experiment $\rightarrow$ založen na *náhodném pokusu*
  - porodní hmotnost: náhodný pokus = zvážení 1 novorozence
  - vzdělání matky: náhodný pokus = dotaz na jednu matku
  - číslo na kostce: náhodný pokus = hod kostkou
- základní prostor Ω = množina všech možných výsledků
  - porodní hmotnost: $0 - \infty$; $0 - 6\,000$ g
  - počet starších sourozenců: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 a více
  - kostka . . . 1–6
- jev = výsledek náhodného pokusu
  - hodila jsem kostkou $\rightarrow$ nastal *jev*: (a) padla 5; (b) padlo liché číslo; (c) padlo číslo ≤ 2
  - zvážila jsem novorozence $\rightarrow$ nastal *jev*: (a) vážil 2 654 g; (b) vážil více než 2 500 g, apod.
- pravděpodobnost
  - vyjadřuje, jak velká je naděje, že nějaký jev nastane
  - $\Pr(A) = \Pr(\text{nastal jev A})$
  - $\Pr(A) \in \langle 0; 1 \rangle$; resp. $\langle 0\,\%; 100\,\% \rangle$
  - příklad: hod kostkou
    - $\Pr(\text{padne 1}) = 1/6 \ldots 16.7\,\%$
    - $\Pr(\text{padne liché číslo}) = 1/2 \ldots 50\,\%$
    - $\Pr(\text{padne 3, 4, 5 nebo 6}) = 2/3 \ldots 66.67\,\%$
    - $\Pr(\text{padne 7}) = 0 \ldots 0\,\%$

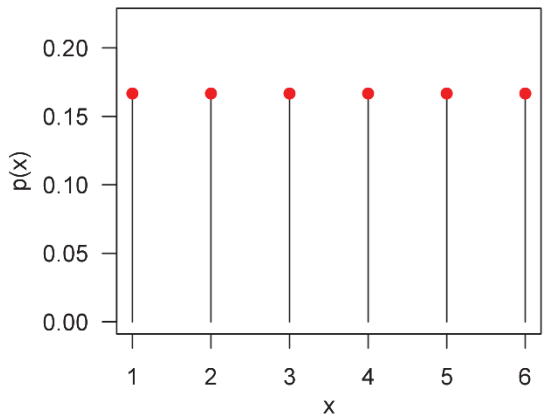
Náhodné veličiny

- víc než výsledek nás často zajímají jeho číselné interpretace
- náhodná veličina X = pravidlo, které zobrazuje základní prostor možných výsledků do množiny reálných čísel

- i-tá realizace náh. veličiny X se značí x_i
  - X ... počet puntíků na vrchní straně kostky: $x_1 = 4$, $x_2 = 1$ ...
  - Y ... dokončené vzdělání; $y_1 = 1$ (ZŠ), $y_2 = 3$ (SŠm) ...
  - Y ... počet starších sourozenců $y_1 = 0$, $y_2 = 2$ ...
  - X ... porodní hmotnost v g; $x_1 = 3470$, $x_2 = 3240$ ...
  - Y ... největší šířka mozkovny v mm; $y_1 = 145$, $y_2 = 139$ ...
- dva typy náhodných veličin
  - diskrétní náhodné veličiny
  - spojité náhodné veličiny

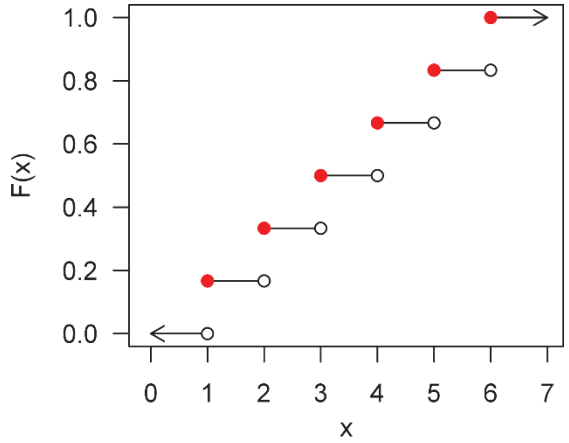
Diskrétní náhodné veličiny

- ze své podstaty nabývají převážně celých hodnot
  - počet sourozenců: 0, 3, 2, . . . ; novorozenec nemůže mít 2.4 sourozence
  - hod kostkou: padne 1, 2, 3, 4, 5, 6; nemůže padnout 3.5
    - $\Pr(X = 4) = \ldots$
    - $\Pr(X \leq 4) = \ldots$
    - $\Pr(X > 4) = \Pr(X \geq 5) = \ldots$
    - $\Pr(3 < X \leq 5) = \ldots$
- pravděpodobnostní funkce $p(x)$
  - $p(x) = \Pr(X = x)$
  - pravděpodobnostní funkce pro případ hodu kostkou



  - nezáporná: $\Pr(x) \geq 0$; normovaná: $\sum_{i=1}^{\infty} \Pr(X = x_i) = 1$

- distribuční funkce $F(x)$
  - $F(x) = \Pr(X \leq x)$
  - distribuční funkce pro případ hodu kostkou



  - **komplementarita**: $\Pr(X > x) = 1 - \Pr(X \leq x) = 1 - F(x)$

Binomické rozdělení

- Bernoulliho pokusy $X_1, \ldots, X_N$
  - $X_i = 1 \ldots$ událost nastala; $X_i = 0 \ldots$ událost nenastala; $i = 1, \ldots, N$
  - $\Pr(X_i = 1) = p$
  - $\Pr(X_i = 0) = 1 - p$
- Binomické rozdělení
  - $X \ldots$ počet událostí v posloupnosti N nezávislých Bernoulliho pokusů, přičemž pravděpodobnost nastání události v každém pokusu je vyjádřena parametrem p
    - počet chlapců v rodině s 12 dětmi
    - celkový počet prstů (na obou rukou), na nichž se alespoň jednou objevil vzor *vír*
  - $\sum_{i=1}^{N} X_i = X \sim \text{Bin}(N, p)$
  - $\theta = (N, p)$
  - pravděpodobnostní funkce

$$p(x) = \binom{N}{x} p^x (1-p)^{N-x} \qquad x = 0, 1, \ldots, N$$

  - vlastnosti: $\text{E}[X] = Np$; $\text{Var}[X] = Np(1-p)$
  - dbinom(x, N, p), pbinom(x, N, p)

Dataset: Počet chlapců v rodinách s 12 dětmi
V rámci studie poměru pohlaví u lidí z roku 1889 bylo na základě záznamů z nemocnic v Sasku zaznamenáno rozdělení počtu chlapců v čtrnáctičlenných rodinách. Mezi $M = 6115$ rodinami s $N = 12$ dětmi byla pozorována početnost chlapců. Údaje ze studie jsou uvedeny v následující tabulce.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	$\sum$
$m_{observed}$	3	24	104	286	670	1033	1343	1112	829	478	181	45	7	6115

Příklad 4.1. Výpočet parametru p binomického rozdělení
Předpokládejme, že náhodná veličina X popisující počet chlapců v rodinách s dvanácti dětmi pochází z binomického rozdělení s parametrem $N = 12$. Vypočítejte odhad pravděpodobnosti výskytu chlapců v rodinách s dvanácti dětmi.

Řešení příkladu 4.1

Pravděpodobnost p výskytu chlapců v rodinách s dvanácti dětmi odhadneme pomocí vzorce

$$\widehat{p} = \frac{\text{počet narozených chlapců}}{\text{celkový počet narozených dětí}} = \frac{\sum_{n=0}^{N} n m_{observed}}{NM}.$$

```r
N <- 12
n <- 0:N
m.obs <- c(3, 24, 104, 286, 670, 1033, 1343, 1112, 829, 478, 181, 45, 7)
M <- sum(m.obs)
p <- sum(n * m.obs) / (N * M) #  0.519215
(p <- round(p, 4))
```

```
[1] 0.5192
```

Interpretace výsledků: Pravděpodobnost výskytu chlapců v rodinách s dvanácti dětmi je ..... (.....%).

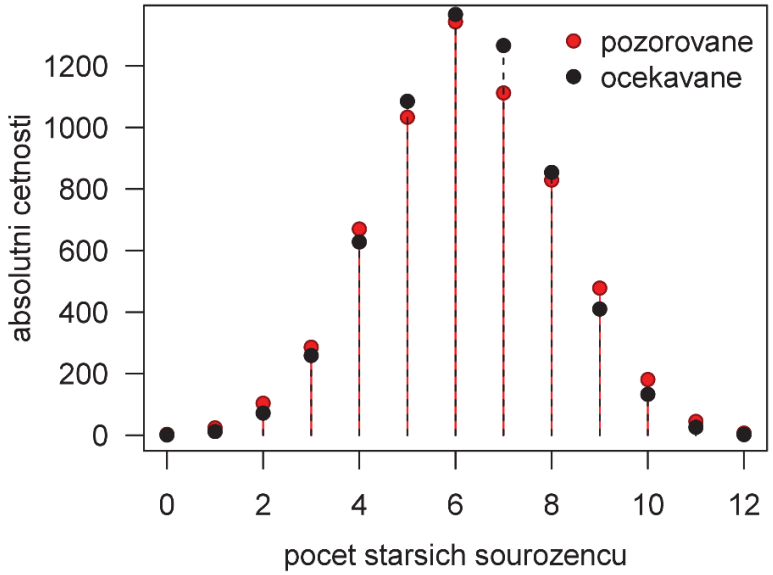
Příklad 4.2. Pozorované a očekávané početnosti v binomickém rozdělení
Za předpokladu, že počet chlapců v rodinách s dvanácti dětmi pochází z binomického rozdělení s parametry $N =$ ..... a $p =$ ..... odhadněte očekávané početnosti chlapců v rodinách s dvanácti dětmi a porovnejte je s pozorovanými početnostmi.

Řešení příkladu 4.2

```r
m.exp <- round(dbinom(0:12, 12, p) * 6115)
tab <- data.frame(rbind(m.obs, m.exp))
names(tab) <- 0:12
```

```
      0  1   2   3   4    5    6    7   8   9  10 11 12
m.obs 3 24 104 286 670 1033 1343 1112 829 478 181 45  7
m.exp 1 12  72 259 628 1085 1367 1266 854 410 133 26  2
```

```
par(mar = c(4, 4, 1, 1))
plot(0:12, m.obs, type = 'h', col = 'red', xlab = '',
     ylab = 'absolutni cetnosti', las = 1)
lines (0:12, m.exp, type = 'h', lty = 2, col = 'black')
points(0:12, m.obs, pch = 21, col = 'darkred', bg = 'red')
points(0:12, m.exp, pch = 21, col = 'black', bg = 'black')
mtext('pocet starsich sourozencu', side = 1, line = 2.4)
legend('topright', pch = c(21, 21), col = c('darkred', 'black'),
       pt.bg = c('red', 'black'), legend = c('pozorovane', 'ocekavane'),
       bty = 'n')
```



Příklad 4.3. Výpočet pravděpodobností za předpokladu binomického rozdělení
Za předpokladu, že náhodná veličina X popisující počet chlapců v rodinách s dvanácti dětmi pochází z binomického rozdělení s parametry $N = \ldots\ldots\ldots\ldots$ a $p = \ldots\ldots\ldots\ldots$ vypočítejte pravděpodobnost, že v rodině s dvanácti dětmi bude (a) právě devět chlapců, (b) nejvýše čtyři chlapci, (c) alespoň osm chlapců, (d) čtyři, pět, šest, nebo sedm chlapců.

Řešení příkladu 4.3
(a) pravděpodobnost, že v rodině s 12 dětmi bude právě devět chlapců

```
24  N <- 12
25  p <- 0.5192
26  dbinom(9, N, p) # 0.06703911
```

(b) pravděpodobnost, že v rodině s 12 dětmi budou nejvýše čtyři chlapci

```
27  sum(dbinom(0:4, N, p)) # 0.1588736
28  pbinom(4, N, p) # 0.1588736
```

(c) pravděpodobnost, že v rodině s 12 dětmi bude alespoň osm chlapců

```
1 - pbinom(7, N, p) # 0.2330869
sum(dbinom(8:12, N, p)) # 0.2330869
```

(d) pravděpodobnost, že v rodině s 12 dětmi bude čtyři, pět, šest, nebo sedm chlapců

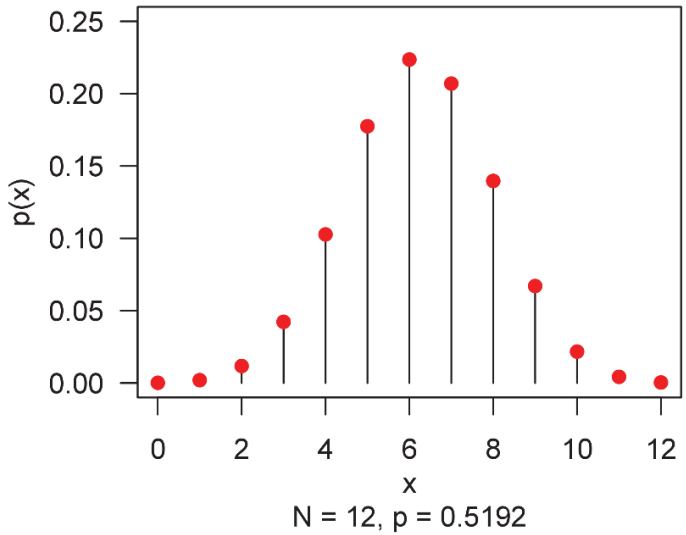
```
sum(dbinom(4:7, N, p)) # 0.7107605
pbinom(7, N, p) - pbinom(3, N, p) # 0.7107605
```

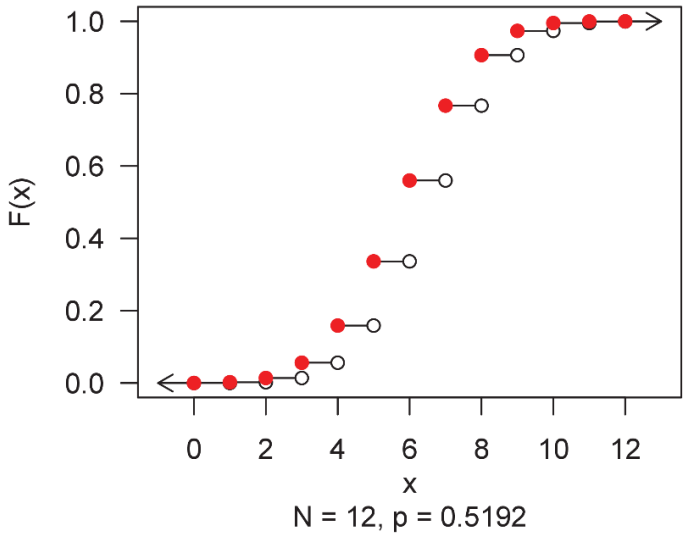
Interpretace výsledků: Pravděpodobnost, že v rodině bude právě devět chlapců, je .....%. Pravděpodobnost, že v rodině budou nejvýše čtyři chlapci, je .....%. Pravděpodobnost, že v rodině bude alespoň osm chlapců, je .....%. Pravděpodobnost, že v rodině bude čtyři, pět, šest, nebo sedm chlapců, je .....%.

Příklad 4.4. Graf pravděpodobnostní a distribuční funkce binomického rozdělení
Nakreslete graf pravděpodobnostní funkce a graf distribuční funkce binomického rozdělení Bin(N, p), kde $N = 12$ a $p = 0.5192$.

Řešení příkladu 4.4

```
33  x <- 0:N
34  px <- dbinom(x, N, p)
35  par(mar = c(4, 4, 1, 1))
36  plot(x, px, type = 'h', ylim = c(0, 0.25), xlab = '', ylab = 'p(x)', las = 1)
37  points(x, px, col = 'red', pch = 19)
38  mtext('x', side = 1, line = 2)
39  mtext('N = 12, p = 0.5192', side = 1, line = 3)
```





Poissonovo rozdělení

- 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 $\lambda > 0$
  - počet starších sourozenců
  - počet úmrtí v důsledku kopnutí koněm v Pruských armádních jednotkách
  - počet revizních operací kolenního koubu
- $X \sim \text{Poiss}(\lambda)$
- $\theta = \lambda$
- pravděpodobnostní funkce

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

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

Příklad 4.5. Výpočet parametru λ Poissonova rozdělení
Načtete datový soubor 17-anova-newborns-2.txt a odstraňte z něj neznámá pozorování. Zaměřte se na znak $X =$*počet starších sourozenců* novorozence. Za předpokladu, že náhodná veličina X popisující počet starších sourozenců novorozence pochází z Poissonova rozdělení parametrem λ odhadněte střední hodnotu počtu starších sourozenců λ.

Řešení příkladu 4.5

Střední hodnotu počtu starších sourozenců odhadneme pomocí vzorce

$$\lambda = \frac{\text{počet starších sourozenců}}{\text{počet novorozenců}} = \frac{\sum_{i=1}^{N} x_i}{N}.$$

```
40  data <- read.delim('17-anova-newborns-2.txt')
41  data <- na.omit(data)
42  prch <- data$prch.N
43  N <- length(prch) # 1381
44  (lambda <- sum(prch) / N) # 0.9427951
```

```
[1] 0.9427951
```
45

Interpetace výsledků: Střední hodnota počtu starších sourozenců novorozenců v datovém souboru $\lambda = \ldots\ldots\ldots\ldots\ldots\ldots\ldots$.

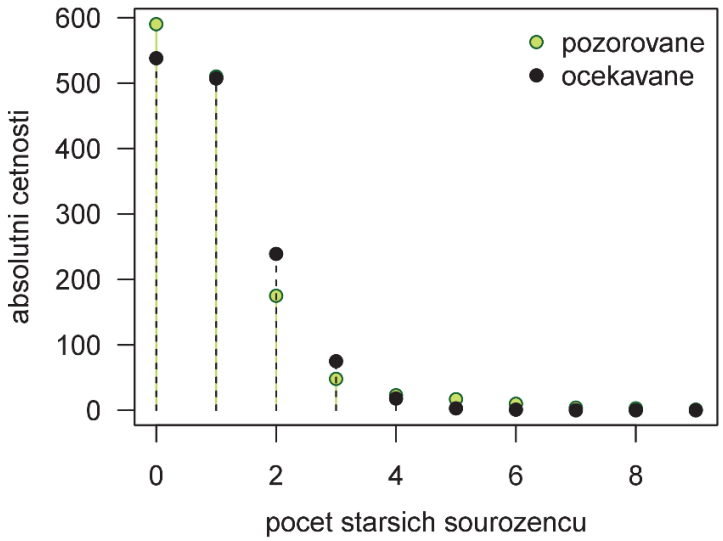
Příklad 4.6. Porovnání pozorovaných a očekávaných početností v Poissonově rozdělení
Za předpokladu, že počet starších sourozenců novorozenců pochází z Poissonova rozdělení s parametrem $\lambda =$..... ..... odhadněte očekávané početnosti starších sourozenců a porovnejte je s pozorovanými početnostmi.

Řešení příkladu 4.6

```
46  m.obs <- data.frame(table(prch))$Freq
47  m.exp <- round(c(dpois(0:9, lambda)) * N)
48  tab <- data.frame(rbind(m.obs, m.exp))
49  names(tab) <- 0:9
```

```
        0   1   2  3  4  5  6 7 8 9
m.obs 590 510 175 48 23 17 10 4 3 1
m.exp 538 507 239 75 18  3  1 0 0 0
```

```
53  par(mar = c(4, 4, 1, 1))
54  plot(0:9, m.obs, type = 'h', ...) # zelene vertikalni cary
55  lines(0:9, m.exp, ...) # cerne vert. cary
56  points(0:9, m.obs, ...) # zelene body
57  points(0:9, m.exp, ...) # cerne body
58  mtext(...) # popisek osy x
59  legend(...) # doplneni legendy
```



Příklad 4.7. Výpočet pravděpodobností za předpokladu Poissonova rozdělení
Za předpokladu, že data pochází z Poissonova rozdělení s parametrem $\lambda = \ldots\ldots\ldots\ldots\ldots\ldots$ určete pravděpodobnost, novorozenec má (a) dva, tři nebo čtyři starší sourozence; (b) alespoň čtyři starší sourozence; (c) nejvýše dva starší sourozence; (d) právě jednoho staršího sourozence.

Řešení příkladu 4.7
(a) pravděpodobnost, že novorozenec má dva, tři nebo čtyři starší sourozence

```
60  sum(dpois(...))
61  ppois(...) - ppois(...)
```

(b) pravděpodobnost, že novorozenec má alespoň čtyři starší sourozence

```
1 - ppois(...) # 0.01567936
```

(c) pravděpodobnost, že novorozenec má nejvýše dva starší sourozence

```
ppois(...) # 0.9299142
sum(dpois(...)) # 0.9299142
```

(d) pravděpodobnost, že novorozenec má právě jednoho staršího sourozence

```
dpois(...) # 0.3672541
```

Interpretace výsledů: Pravděpodobnost, že novorozenec má dva, tři nebo čtyři starší sourozence je .....%. Pravděpodobnost, že novorozenec má alespoň čtyři starší sourozence je .....%. Pravděpodobnost, že novorozenec má nejvýše dva starší sourozence je .....%. Pravděpodobnost, že novorozenec má jednoho staršího sourozence je .....%.

Příklad 4.8. Graf pravděpodobnostní a distribuční funkce Poissonova rozdělení
Nakreslete graf pravděpodobnostní a distribuční funkce Poissonova rozdělení Poiss(0.9428) v hodnotách $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$, a $x \geq 9$.

Řešení příkladu 4.8

```
66  N <- 9
67  x <- 0:N
68  px <- dpois(x, lambda) # pstni fce rozdeleni Poiss(lambda) v hodnotach x
69  par(...) # nastaveni okraju 5, 4, 1, 1
70  plot(x, px, type = ..., ylim = c(0, 0.45), ...) # zelene vertikalni cary
71  points(x, px, ...) # zelene body
72  box(...) # ramecek okolo grafu
73  mtext(...) # popisek osy x
74  mtext(bquote(paste(lambda == 0.9428)), side = 1, line = 3) # druhy popisek osy x
```