Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Téma 1: Binomický model

Veronika Bendová

bendova.veroonika@gmail.com

Binomické rozdělení $\mathrm{Bin}(N, p)$

- Bernoulliho pokusy $X_1, \ldots, X_N$
  - $X_i = 1 \ldots$ událost nastala; $X_i = 0 \ldots$ událost nenastala; $i = 1, \ldots, N$
  - $\Pr(X_i = 1) = p$
  - $\Pr(X_i = 0) = 1 - p = q$
- $X \ldots$ počet událostí v posloupnosti N nezávislých Bernoulliho pokusů, přičemž pravděpodobnost nastání události v každém pokusu je vyjádřena parametrem p
- $\sum_{i=1}^{N} X_i = X \sim \mathrm{Bin}(N, p)$
- $\theta = (N, p)$
- pravděpodobnostní funkce

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

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

- **Dataset 1: 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 1.1. Výpočet očekávaných početností za předpokladu binomického modelu
Vezměte údaje z **datasetu 1**. Vypočítejte očekávané početnosti výskytu chlapců $m_{expected}$ za předpokladu, že početnosti chlapců X v rodinách mají binomické rozdělení $\mathrm{Bin}(N, p)$ s parametry $N = 12$ a

$$\widehat{p} = \frac{\sum_{n=0}^{N} n m_{observed}}{NM}. \tag{1.1}$$

Řešení příkladu 1.1

```r
N <- ... # hodnota parametru N
n <- ... # posloupnost 0, 1, ..., N
m.obs <- ... # posloupnost m.obs, tj. 3, 24, ..., 7
M <- sum(...) # pocet vsech rodin M (soucet vsech cisel v m.obs)
p <- sum(...) / (...)  # odhad parametru p podle vzorce 1.1
```

```
          p
1 0.519215
```

```r
m.exp <- round(dbinom(...) * ...) # vektor ocek. abs. cetnosti
```

```
      0  1  2   3   4    5    6    7   8   9  10 11 12
m.exp 1 12 72 258 628 1085 1367 1266 854 410 133 26  2
```

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

```
[1] 6114
```

```r
p <- round(...) # odhad parametru p zaokrouhleny na 4 des. mista
m.exp <- round(dbinom(...) * ...) # novy vypocet ocek. abs. cetnosti
```

```
      0  1  2   3   4    5    6    7   8   9  10 11 12
m.exp 1 12 72 259 628 1085 1367 1266 854 410 133 26  2
```

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

```
[1] 6115
```

Odhad parametru p, tj. $\widehat{p} = 0.5192$. Tabulka očekávaných početností $m_{expected}$ je

n	0	1	2	3	4	5	6	7	8	9	10	11	12	$\sum$
$m_{expected}$	1	12	72	259	628	1085	1367	1266	854	410	133	26	2	6115

Z tabulky očekávaných početností vidíme, že součet očekávaných početností dává výsledek $M = 6114$. Odchylka od původní hodnoty $M = 6115$ je způsobena přesným výpočtem očekávaných početností a jejich následným zaokrouhlením. Konkrétně v tomto případě, pokud bychom očekávané početnosti spočítali na základě binomického rozdělení $\mathrm{Bin}(N, p)$, kde $N = 12$ a $\widehat{p} = 0.5192$ (hodnotu $\widehat{p}$ bychom pro další výpočet zaokrouhlili na čtyři desetinná místa) dostali bychom následující tabulku očekávaných početností.

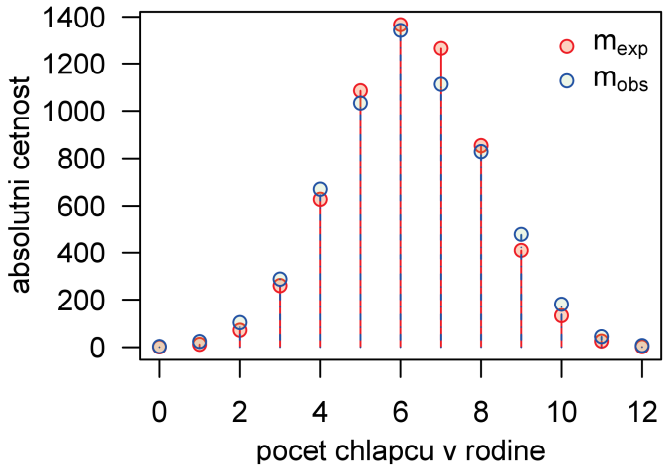
n	0	1	2	3	4	5	6	7	8	9	10	11	12	$\sum$
$m_{expected}$	1	12	72	259	628	1085	1367	1266	854	410	133	26	2	6115

Při použití hodnoty odhadu $\widehat{p}$ zaokrouhlené na čtyři desetinná místa došlo k mírné úpravě binomického rozdělení a očekávaná početnost výskytu tří chlapců v rodině se z hodnoty 258.47 přehoupla na hodnotu 258.52, čímž jsme po zaokrouhlení získali absolutní početnost 259 namísto 258. Ostatní očekávané početnosti zůstaly nezměněny, jelikož jejich nezaokrouhlená hodnota byla dostatečně blízko k celému číslu.

Příklad 1.2. Overdispersion a underdispersion v binomickém modelu
V předchozím příkladu 1.1 jsme stanovili očekávané početnosti výskytu chlapců v rodinách s dvanácti dětmi. 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 v tomto případě k overdisperzi nebo underdisperzi. Závěr podložte srovnáním rozptylu vypočítaného z pozorovaných dat s rozptylem vypočítaným z očekávaných dat.

Řešení příkladu 1.2

```
19  par(mar = c(3, 4, 1, 1)) # nastaveni okraju grafu (viz napoveda funkce par())
20  plot(n, m.exp, type = 'h', col = ..., xlab = '', ylab = ...,
21       las = 1) # graf s vertikalnimi cervenymi carami ocek. abs. cetnosti m.exp
22  points(n, m.exp, , pch = 21, col = ...,
23         bg = rgb(1, 0, 0, 0.2)) # cervene body abs. ocek. cetnosti m.exp
24  lines (n, m.obs, col = ..., type = ...,
25         lty = 4) # modre vertikalni cary pozor. abs. cetnosti m.obs
26  points(..., ..., pch = ..., bg = rgb(...),
27         col = ...) # modre body pozor. abs. cetnosti m.obs
28  mtext('...', side = 1, line = 2.1) # doplneni popisku osy x pod graf
29  legend('topright', pch = c(21, 21), col = c(..., ...),
30         pt.bg = c(rgb(...), rgb(...)),
31         legend = c(expression(m[exp]), expression(m[obs])), bty = 'n') # legenda
```



Obrázek: Porovnání pozorovaných a očekávaných početností v binomickém modelu

Oproti očekávaným početnostem jsou pozorované početnosti krajních případů vyšší a naopak pozorované početnosti nejčastějších případů jsou nižší. Tato situace ukazuje na vyšší rozptyl pozorovaných početností než teoretických. Jde tedy o .....disperzi.

```
32 observed <- rep(n, m.obs) # vektor pozor. dat: 0,0,0,1,...,1,...,12,...,12
33 expected <- rep(..., ...) # vektor ocek. dat: 0,1,...,1,...,11,...,11,12,12
34 var.obs <- var(...) # odhad rozptylu na zaklade pozor. dat
35 var.exp <- var(...) # odhad rozptylu na zaklade ocek. dat
36 (tab <- data.frame(var.obs, var.exp)) # tabulka vysledku
```

```
  var.obs  var.exp
1 3.48984 2.995539
```

37
38

Rozptyl skutečných počtů chlapců v rodinách s dvanácti dětmi je 3.4898, rozptyl očekávaných počtů chlapců je 2.9956. Rozptyl skutečných počtů je tedy vyšší než rozptyl očekávaných počtů chlapců v rodinách s dvanácti dětmi.

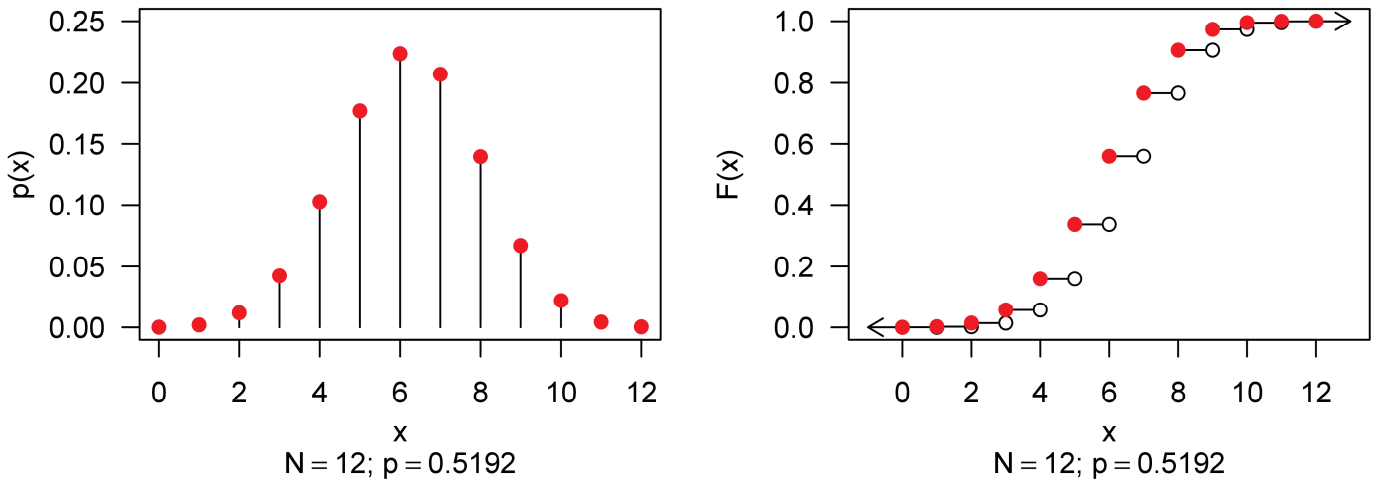
Příklad 1.3. Graf pravděpodobnostní a distribuční funkce binomického modelu
V příkladu 1.1 jsme odhadli hodnotu parametru p binomického rozdělení $\mathrm{Bin}(N, p)$ jako $\widehat{p} = 0.5192$. Hodnota parametru $N = 12$. Nakreslete graf pravděpodobnostní a distribuční funkce binomického rozdělení $\mathrm{Bin}(N, p)$, kde $N = 12$ a $p = 0.5192$.

Řešení příkladu 1.3
Graf pravděpodobností funkce $p(x)$

```
39  N <- ... # hodnota parametru N
40  x <- ... # posloupnost 0, 1, ..., N
41  p <- ... # zaokrouhleny odhad p
42  px <- dbinom(...) # pstni fce Bin(N, p) v hodnotach 0, 1, ..., N
43  Fx <- pbinom(...) # distr. fce Bin(N, p) v hodnotach 0, 1, ..., N
44
45  par(mar = c(...)) # nastaveni okraju grafu (4, 4, 1, 1)
46  plot(x, px, type = ..., ylim = c(0, 0.25),
47       ylab = '...', xlab = '', las = 1) # graf pstni fce (cerne vertikalni cary)
48  points(..., ..., col = ..., pch = 19) # cervene body
49  mtext('x', side = ..., line = ...) # popisek osy x
50  mtext(bquote(paste(N == .(N), '; ', p == .(p))),
51        side = ..., line = 3.2) # druhy popisek osy x: N = ..., p = ...
```

Graf distribuční funkce $F(x)$

```
52  plot(x, Fx, type = 'n', xlim = c(-1, N + 1), ylim = c(0, 1),
53       xlab = ..., ylab = ..., las = ...) # priprava prazdneho grafu
54  segments(x, Fx, x + 1, Fx) # vodorovne cerne cary
55  arrows(0, 0, -1, 0, length = 0.1) # dolni sipka
56  arrows(..., ..., ..., ..., length = ...) # horni sipka
57  points(x, c(0, Fx[1:N]), pch = ..., col = ..., bg = ...) # prazdne body
58  points(..., ..., col = ..., pch = ...) # plne cervene body
59  mtext(..., side = ..., line = ...) # popisek osy x
60  mtext(..., side = ..., line = ...) # druhy popisek osy x: N = ..., p = ...
```



Obrázek: Pravděpodobnostní a distribuční funkce binomického modelu

Příklad 1.4. Výpočet pravděpodobností na základě binomického modelu
Za předpokladu, že náhodná veličina X, udávající počet chlapců v rodině s dvanácti dětmi, pochází z binomického rozdělení, tj. $X \sim \mathrm{Bin}(N, p)$, s parametry $N = 12$ a $p = 0.5192$, 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 1.4
(a)

```
61  N <- ... # hodnota parametru N
62  p <- ... # zaokrouhleny odhad p
63  dbinom(...) # vypocet pravdepodobnosti
```

```
[1] 0.06703911                                                    64
```

(b)

```
65  pbinom(...) # vypocet pravdepodobnosti
```

```
[1] 0.1588736                                                     66
```

(c)

```
1 - pbinom(...) # vypocet pravdepodobnosti - prvni zpusob
sum(dbinom(...)) # vypocet pravdepodobnosti - druhy zpusob
```

```
[1] 0.2330869
```

(d)

```
sum(dbinom(...)) # vypocet pravdepodobnosti - prvni zpusob
pbinom(...) - pbinom(...) # vypocet pravdepodobnosti - druhy zpusob
```

```
[1] 0.7107605
```

Pravděpodobnost, že v rodině bude právě devět chlapců, je 6.7 %. Pravděpodobnost, že v rodině budou nejvýše čtyři chlapci, je 15.89 %. Pravděpodobnost, že v rodině bude alespoň osm chlapců, je 23.31 %. Pravděpodobnost, že v rodině bude čtyři, pět, šest, nebo sedm chlapců, je 71.08 %.

Příklad 1.5. Střední hodnota a rozptyl náhodné veličiny z binomického modelu
Za předpokladu, že náhodná veličina X, udávající počet chlapců v rodině s dvanácti dětmi, pochází z binomického rozdělení, tj. $X \sim \mathrm{Bin}(N, p)$, s parametry $N = 12$ a $p = 0.5192$, 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 1.2).

Řešení příkladu 1.5

```
73  N <- ... # hodnota parametru N
74  p <- ... # zaokrouhleny odhad p
75  E.X <- ... # vypocet stredni hodnoty E[X] rozdeleni Bin(N, p)
76  Var.X <- ... # vypocet rozptylu Var[X] rozdeleni Bin(N, p)
77
78  expected <- rep(n, m.exp) # vektor ocek. dat: 0,0,0,1,...,1,...,12,...,12
79  E.exp <- mean(expected) # odhad stredni hodnoty na zaklade ocek. dat
80  Var.exp <- var(expected) # odhad rozptylu na zaklade ocek. dat
81
82  observed <- ... # vektor pozor. dat: 0,1,...,1,...,11,...11,12,12
83  E.obs <- ... # odhad stredni hodnoty na zaklade pozor. dat
84  Var.obs <- ... # odhad rozptylu na zaklade pozor. dat
85
86  tab <- data.frame(E.X, Var.X, ...) # tabulka vysledku
```

```
     E.X    Var.X    E.exp   Var.exp    E.obs Var.obs
1 6.2304 2.995576 6.229926 2.995539 6.230581 3.48984
```

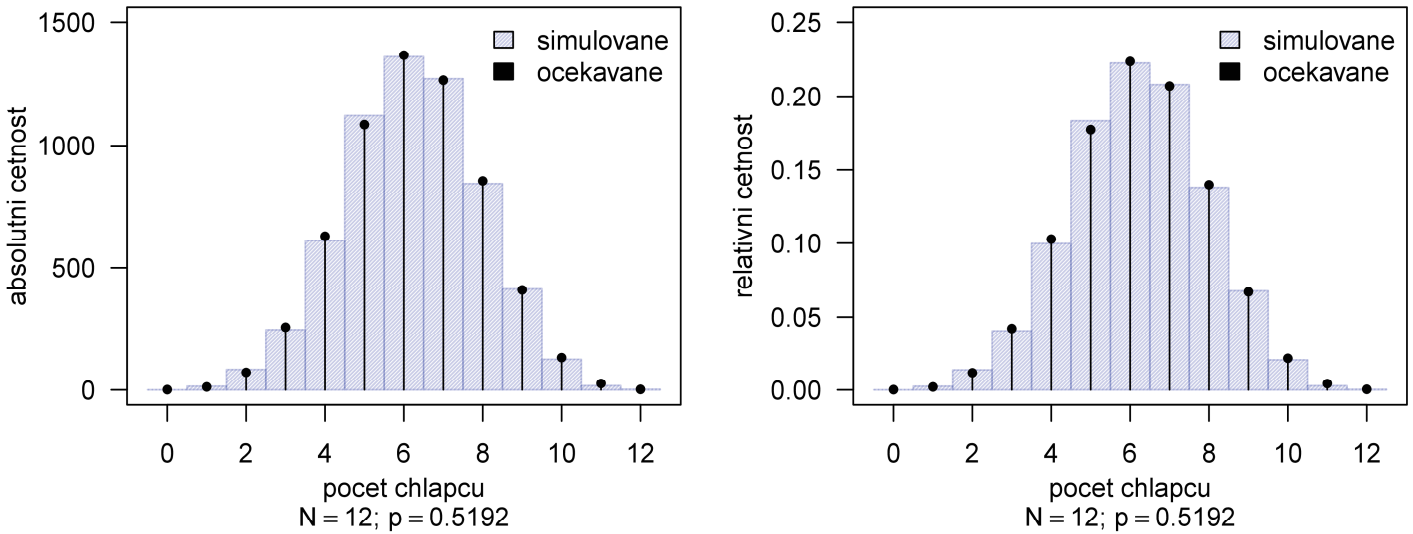
Střední hodnota počtu chlapců v rodině s dvanácti dětmi je 6.2304 s rozptylem 2.9956. Odhad střední hodnoty počtu chlapců v rodině vypočítaný na základě očekávaných hodnot je 6.2299 s rozptylem 2.9955. Odhad střední hodnoty počtu chlapců v rodině vypočítaný na základě pozorovaných dat je 6.2306 s rozptylem 3.4898.

Příklad 1.6. Simulační studie pro binomický model
Sekci o binomickém rozdělení zakončíme simulační studií modelující chování očekávaných početností náhodné veličiny X . . . počet chlapců v rodinách s dvanácti dětmi za předpokladu, že $X \sim \mathrm{Bin}(N, p)$, $N = 12$, $p = 0.5192$.

Vygenerujte pseudonáhodná čísla X (početnosti úspěchů) opakovaná M-krát ($M = 6115$) z $\mathrm{Bin}(N, p)$, kde $N = 12$ a $p = 0.5192$. Vytvořte histogram vygenerovaných pseudonáhodných čísel superponovaný hodnotami očekávaných (teoretických) početností za předpokladu, že $X \sim \mathrm{Bin}(N, p)$.

Řešení příkladu 1.6

```
89  M <- ... # pocet vsech rodin M
90  N <- ... # hodnota parametru N
91  n <- ... # posloupnost 0, 1, ..., N
92  p <- ... # zaokrouhleny odhad p
93  X <- rbinom(M, N, p) # vygenerovani nahodneho vyberu o delce M z rozd. Bin(N, p)
94  p.exp <- dbinom(...) # pstni fce Bin(N, p) v hodnotach 0, 1, ..., N
95  m.exp <- round(... * ...) # vektor ocek. abs. cetnosti
96
97  par(mar = ...) # nastaveni okraju 4, 4, 1, 1
98  hist(X, prob = F, breaks = seq(-0.5, 12.5, by = 1), density = 60,
99       col = rgb(...), ylim = c(0, 1500), xlab = '', ylab = ..., main = '',
100      las = ...) # histogram nahodneho vyberu X (v abs. cetnostech)
101 box(bty = 'o') # ramecek okolo grafu
102 lines (n, m.exp, type = ...) # vertikalni cary ocek. abs. cetnosti m.exp
103 points(..., ..., pch = 20) # body ocek. abs. cetnosti m.exp
104 mtext(...) # popisek osy x
105 mtext(...) # druhy popisek osy x: N = ..., p = ...
106 legend('topright', fill = c(rgb(...), 'black'), density = c(60, 200),
107        legend = c('simulovane', 'ocekavane'), bty = 'n') # legenda
108
109 hist(X, prob = T, breaks = ..., density = ..., col = rgb(...),
110      ylim = c(0, 0.25), xlab = ..., ylab = ..., main = ...,
111      las = ...) # histogram nahodneho vyberu X (v rel. cetnostech)
112 box(...) # ramecek okolo grafu
113 lines (...) # vertikalni cary ocek. rel. cetnosti p.exp
114 points(...) # body ocek. rel. cetnosti p.exp
115 mtext(...) # popisek osy x
116 mtext(...) # druhy popisek osy x: N = ..., p = ...
117 legend(..., fill = c(..., ...), density = c(..., ...),
118        legend = c(..., ...), bty = ...) # legenda
```



Obrázek: Porovnání pozorovaných a očekávaných početností v binomickém modelu