5.1 Náhodné veličiny

Náhodnou veličinu na X pravděpodobnostním prostoru $\langle \Omega, \mathcal{F}, \mathcal{P} \rangle$ jste si definovali jako zobrazení $X : \Omega \to \mathbb{R}$, které splňuje pro každé $a \in \mathbb{R}$:

$$\{\omega \in \Omega \,|\, X(\omega) \leq a\} \in \mathcal{F}$$

▷ **Příklad 5.1.** Nechť $\Omega = \{1, 2, 3, 4\}$ a $\mathcal{F} = \{\emptyset, \Omega, \{1\}, \{2, 3, 4\}\}$. Je $\mathcal{F}$ σ-algebra?
Nechť $X(\omega) = \omega + 1$. Je takto definovaná funce náhodnou veličinou na $\langle \Omega, \mathcal{F}, \mathcal{P} \rangle$?
Pro $a = 2$ je podmínka z definice splněna, platí totiž $\{\omega \in \Omega \,|\, X(\omega) \leq 2\} = \{1\} \in \mathcal{F}$. Ale vezmeme-li $a = 3$, dostaneme $\{\omega \in \Omega \,|\, X(\omega) \leq 3\} = \{1, 2\} \notin \mathcal{F}$. Protože podmínka má platit pro každé $a \in \mathbb{R}$, není takto definované zobrazení náhodnou veličinou na daném pravděpodobnostním prostoru.
Pokud $X(1) = 0$ a $X(2) = X(3) = X(4) = 1$, pak takto definované zobrazení bude náhodnou veličinou na $\langle \Omega, \mathcal{F}, \mathcal{P} \rangle$. Ověřte.

Náhodná veličina je zobrazení, které každému výsledku náhodného pokusu přiřadí jednoznačně nějaké reálné číslo. Jak takovou náhodnou veličinu přidat do našeho již existujícího data framu?

▷ **Příklad 5.2.** Uvažujte náhodný pokus hod dvěma kostkami. Již víme, že množinu možných výsledků tohoto náhodného pokusu můžeme uložit do data framu. Na našem pravděpodobnostním prostoru budeme uvažovat tři náhodné veličiny, X bude součet teček na obou kostkách, Y bude průměr hodnot a Z bude minimum z hodnot. V R můžeme definovat náhodnou veličinu pro daný data frame pomocí fukce `addrv(space, FUN = NULL, invars = NULL, name = NULL, ...)`, kde `space` je data frame, `FUN` je funkce, která má být použita na každý řádek v data framu, `invars` je vektor sloupců, které mají být vstupem do `FUN`, `name` je jméno definované náhodné veličiny, `...` je výraz, který definuje náhodnou veličinu, pokud není použito parametru `FUN`.

```
> DF<-rolldie(2,makespace="TRUE")
> DF<-addrv(DF,FUN=sum,name="X")
> DF<-addrv(DF,FUN=mean,invars=c("X1","X2"),name="Y")
> DF<-addrv(DF,FUN=min,invars=c("X1","X2"),name="Z")
> DF
   X1 X2  X   Y Z      probs
1   1  1  2 1.0 1 0.02777778
2   2  1  3 1.5 1 0.02777778
3   3  1  4 2.0 1 0.02777778
4   4  1  5 2.5 1 0.02777778
5   5  1  6 3.0 1 0.02777778
6   6  1  7 3.5 1 0.02777778
.............................
33  3  6  9 4.5 3 0.02777778
34  4  6 10 5.0 4 0.02777778
35  5  6 11 5.5 5 0.02777778
36  6  6 12 6.0 6 0.02777778
```

Nyní můžeme snadno spočítat např. pravděpodobnost, že součet teček bude větší než 7?

```
> prob(DF,X>7)
[1] 0.4166667
```

Všechny tři definované náhodné veličiny jsou diskrétní. Jednou ze základních charakteristik diskrétních náhodných veličin je pravděpodobnostní funkce $f_X : \mathbb{R} \to [0, 1]$ definovaná jako $f_X(x) = P(\{X = x\})$. Chceme-li určit pravděpodobnostní funkci pro veličinu X, můžeme si postupně určit

```
> prob(DF,X==2)
[1] 0.02777778
> prob(DF,X==3)
[1] 0.05555556
.................
> prob(DF,X==12)
[1] 0.02777778
```

Tento způsob je ale zdlouhavý a obecně nepoužitelný. Ukážeme si, jak lze využit v R tzv. faktorů. Faktor je vektor, podle kterého můžeme hodnoty jiného vektoru třídit do skupin a následně zpracovávat. Skupiny jsou vzájemně různé hodnoty faktoru a zjistíme je pomocí příkazu `levels`. V našem případě je faktor vektor hodnot, kterých nabývá náhodná veličina X, protože chceme sečíst pravděpodobnosti v pravém sloupci dataframu pro každou hodnotu X zvlášť. To, aby funkce `sum` byla na vekor pravděpodobností aplikována dle faktoru `x` zajistí funkce `tapply`.

```
> x<-DF$X
> x
 [1]  2  3  4  5  6  7  3  4  5  6  7  8  4  5  6  7  8  9  5  6  7  8  9 10  6
[26]  7  8  9 10 11  7  8  9 10 11 12
> levels(factor(x))
 [1] "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12"
> pfce<-tapply(DF$prob,x,sum)
> pfce
         2          3          4          5          6          7          8
0.02777778 0.05555556 0.08333333 0.11111111 0.13888889 0.16666667 0.13888889
         9         10         11         12
0.11111111 0.08333333 0.05555556 0.02777778
```

Nyní nám již nic nebrání pravděpodobnostní funkci náhodné veličiny X vykreslit do grafu.

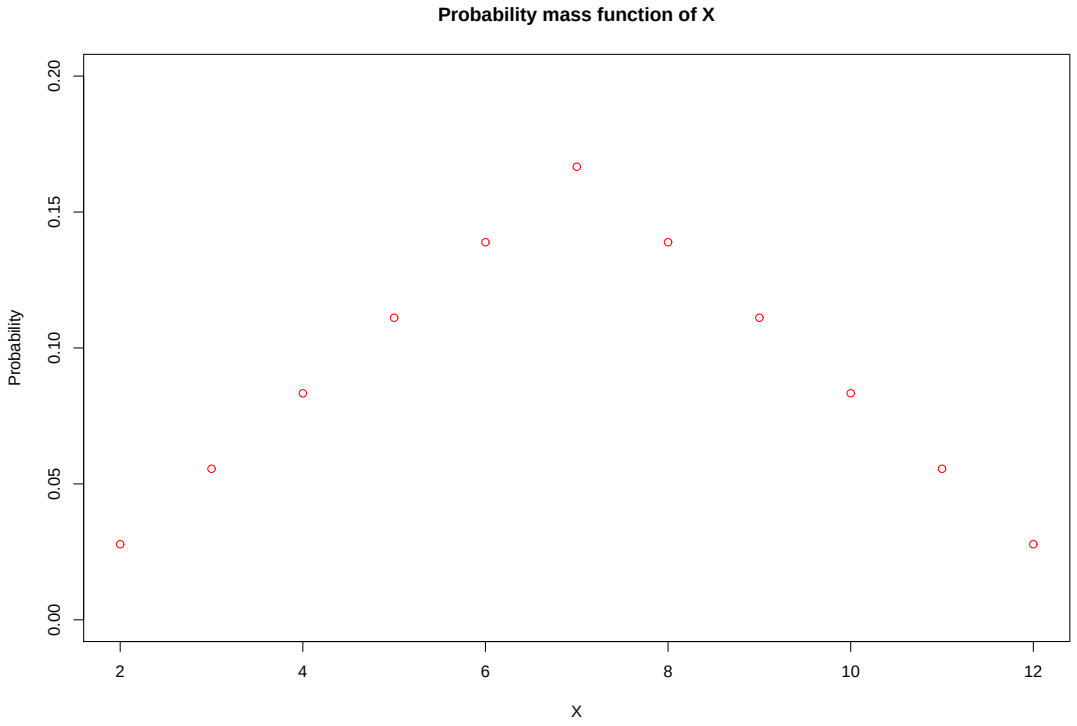
```
> postscript("pfceX.eps")   #uložení grafu do souboru pfceX.eps
> plot(i,pfce[i-1],type="p", ylim=c(0,0.2), main="Probability mass function of X",
  xlab="X", ylab="Probability", col="red")
> dev.off()
```

Vykreslený graf vidíte na obrázku 5.2. Příkaz `dev.off()` zavře grafické zařízení. Výsledek, tedy soubor pfceX.eps najdete v adresáři, ve kterém jste R spustili. Projděte si funkci `plot` v nápovědě, podívejte se na možné parametry a vyzkoušejte různá nastavení.

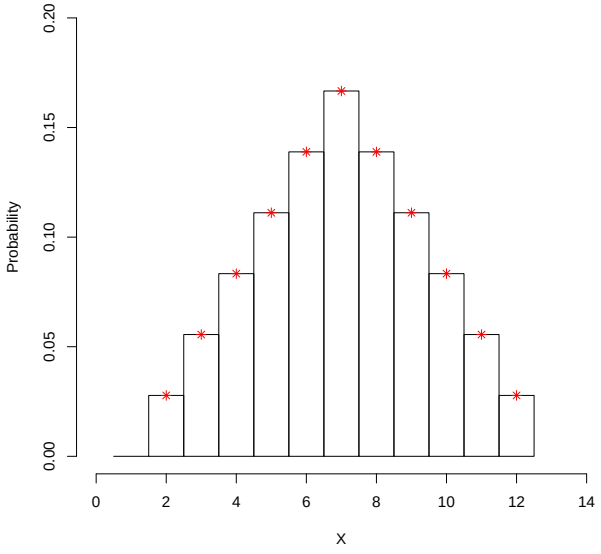
Pravděpodobnostní funkci náhodné veličiny X můžeme získat i jinak - z histogramu hodnot, kterých veličina nabývá.

```
> bins=seq(0.5,12.5,1)
> postscript("pfce2.eps")
> hist(x,freq=FALSE,breaks=bins,xlim=c(0,14),ylim=c(0,0.2),
  xlab="X",ylab="Probability",main="")
> points(i,pfce[i-1],type="p", col="red",pch=8)
> dev.off()
```

Parametrem `breaks` určíme body zlomu mezi jednotlivými sloupci v histogramu. Parametrem nemusí být jen vektor, je možné i přímo zadat počet sloupců.
Mohli bychom takto určit pravděpodobnostní funkci (tj. pomocí četnosti hodnot), kdyby kostka byla falešná? Svou odpověď si promyslete. Na obrázku 5.2 vidíte histogram i hodnoty pravděpodobnostní funkce získané pomocí přechozího přístupu.



Obrázek 1: Pravděpodobnostní fuknce n. v. X



Obrázek 2: Histogram hodnot n.v. X

► **Příklad 5.1.** Určete pravděpodobnostní funkce pro náhodné veličiny Y a Z, vykreslete grafy těchto funkcí. Pro všechny tři náhodné veličiny určete střední hodnotu, rozptyl a sestrojte distribuční funkci.

▷ **Příklad 5.3.** Birthday problem
Narozeninový problém nebo také narozeninový paradox řeší otázku, jaká je pravděpodobnost, že v náhodné skupině n lidí budou mít alespoň dva narozeniny ve stejný den? Kolik si myslíte, že je potřeba lidí, aby pravděpodobnost byla větší než 50%? Tipujete kolem 180? Správná odpověď je 23! Předpokládejme pro jednoduchost, že rok má 365 dní a že pravděpodobnost narození v daný den je

pro všechny dny stejná. Uvažujme skupinku 23 lidí. Vezmeme-li prvního člověka, můžeme jeho datum narození porovnat s dalšími 22 lidmi. Vezmeme-li druhého, porovnáváme jeho datum už jen s 21 lidmi atd. Celkový počet dvojic které ve skutečnosti srovnáváme je tak $22 + 21 + \cdots + 1 = \binom{23}{2} = 253$. Podíváme-li se místo na počet lidí ve skupině na počet různých dvojic, které lze vytvořit, přestává být paradox paradoxem.
Jak spočítáme pravděpodobnost, že ve skupině n lidí mají alespoň dva stejné datum narození? Pravděpodobnost určíme pomocí pravděpodobnosti jevu opačného, tj. že žádní dva lidé nemají narozeniny ve stejný den. To ale znamená, že zatímco první člověk má 365 možností, kdy může mít narozeniny, druhý jich má už jen 364 atd. až n-tý má jen $365 - n + 1$ možností, tj. celkem $\frac{365!}{(365-n)!}$. Počet všech možných kombinací pro n osob je pak 365^n. Pravděpodobnost pak určíme jako $P(A_n) = 1 - \dfrac{365!}{(365-n)!365^n}$.
Vykreslíme si funkci v R pomocí `function(arguments)`:

```
bp<-function(n){
 1-factorial(365)/(factorial(365-n)*365^n)
}
> bp(23)
...  value out of range ...
```

Tento problém vyřešíme snadno pomocí funkce `lfaktorial()`, která vrátí přirozený logaritmus z faktoriálu. Řešení bude vypadat proto následovně:

```
bp<-function(n){
  1-exp(lfactorial(365)-lfactorial(365-n)- n*log(365))
}
> bp(23)
[1] 0.5072972
> bp(50)
[1] 0.9703736
```

Vidíte, že pro 50 lidí je již pravděpodobnost blízká 1!. Pro jaké počty lidí ve skupině, je pravděpodobnost mez 0.9 a 0.95?

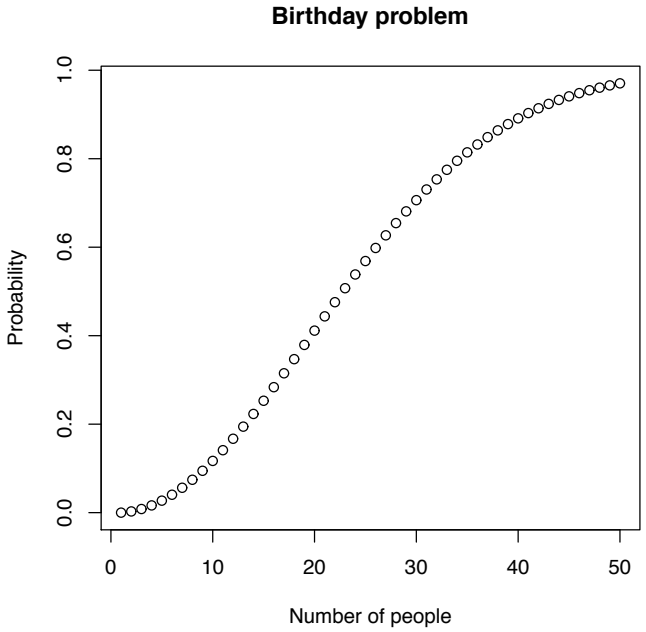
```
> n[which(0.9< bp(n)&bp(n)<0.95)]
[1] 41 42 43 44 45 46
```

Vykreslete si graf závislosi pravděpodobnosti na počtu lidí, např. pro $n = 1 : 50$ vidíte výsledek na obrázku 5.3.

► **Příklad 5.2. Kolik kaprů je v rybníce?**
Jedním z možných způsobů, jak odhadnout počet ryb n v rybníce je vylovit určité množství ryb (např. 50) a ty označit. Po nějaké době opět vylovit 50 ryb a zjistit, kolik z nich je označených - k. Pak lze určit pro jaké n je pravděpodobnost, že chytíte právě k označených a $50 - k$ neoznačených ryb, největší. Předpokládejte, že jste označili 50 ryb a při druhém odchytu bylo mezi 50ti rybami 5 označných. Odhadněte, kolik ryb je v rybníce. [499]
Nápověda: Vykreslete graf pravděpodobnosti (v závislosti na n) a najděte maximum této funkce.

► **Příklad 5.3.** Házíme opakovaně mincí. $\Omega = \{P, O\}$. Uvažujte náhodný jev A: PP padne dřív než OO. Určete $P(A)$ obecným předpisem i pro nevyváženou minci. $[(p^2 + qp^2)/(1 - pq)]$

► **Příklad 5.4.** Při hodu mincí uvažujme $\Omega = \{P, O\}$, $\mathcal{F} = 2^\Omega$. Definujme dvě náhodné veličiny X, Y a to následovně:
$X(P) = 1 \qquad Y(P) = 0$
$X(O) = 0 \qquad Y(O) = 1.$
Určete pro obě dvě nádhoné veličiny pravděpodobnostní a distribuční funkci.



► **Příklad 5.5.** Nechť náhodná veličina X znači počet aut, která projedou myčkou za jednu hodinu. Tato náhodná veličina má pravděpodobnostní funkci dánu tabulkou:

x	4	5	6	7	8	9
$P(X = x)$	1/12	1/12	1/4	1/4	1/6	1/6

Uvažujme dále funkci $g(X) = 2X - 1$, která reprezentuje množství peněz v dolarech, které majitel myčky platí obsluze. Určete očekáváný výdělek obsluhy za jednu hodinu. Určete rozptyl náhodné veličiny X a náhodné veličiny $g(X)$. [12.67; 2.14; 8.55]

Reference

[1] Capinski M., Zastawniak T. J.: Probability Through Problems
Springer 2001, ISBN 978-0-387-95063-1.

[2] Devore J. L.: Probability and Statistics for Engineering and the Sciences
Duxbury Press, 7. vydání 2008, ISBN 978-0-495-55744-9.

[3] Kerns G. J.: Elementary Probability on Finite Sample Spaces, 2009, reference manual package prob, available from: http://CRAN.R-project.org/package=prob

[4] Kerns G. J: Introduction to Probability and Statistics Using R, First Edition
http://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf

[5] Kenneth Baclawski: Introduction to Probability with R
Chapman and Hall/CRC, ISBN 978-1420065213.