6.1 Generování (pseudo)náhodných čísel

Jedna z důležitých součástí statistického zpracování dat je možnost generování (pseudo)náhodných čísel ze zvolené distribuce. Pokud chceme vybírat čísla z konečné populace (ať už s opakováním nebo bez), můžeme použít funkce `sample(x, size, replace = FALSE, prob = NULL)`. Použití je jasné z následujících příkladů:

```
> sample(1:5,size=10,replace=TRUE)
 [1] 5 4 1 2 3 1 2 5 3 2
> sample(5,size=10,replace=TRUE)
 [1] 1 3 3 5 5 1 3 3 1 3
> sample(10,size=5,replace=FALSE)
[1]  8  5  2  1 10
> sample(10,size=5,replace=TRUE)
[1] 7 8 4 5 5
> x<-sample(1:3,size=100,replace=TRUE,prob=c(0.2,0.7,0.1))
> table(x)
x
 1  2  3
15 76  9
```

Jednín z nejznámějších algoritmů pro generování (pseudo)náhodných čísel na intervalu $[0, M]$ je Lineární kongruentní generátor (LCG):

1. Zvolte $a, b, M \in \mathbb{N}$
2. Zvolte $x_0 \in \{0, 1, \ldots, M-1\}$
3. Iterujte $x_i := (ax_{i-1} + b) \bmod M$,

kde mod značí zbytek po celočíselném dělení. Uvědomte si, že algoritmus je deterministický (stejné počáteční hodnoty vytvoří stejnou sekvenci) a periodický. Kvalita algoritmu je hodně ovlivněna volbou počátečních hodnot.

► **Příklad 6.1.** Implementujte LCG v R, zvolte sami a, b, M. Simulujte 10 000 čísel na intervalu $[0, 1]$ a vykreslete histogram. Srovnejte Vámi vytvořený generátor s tím, který je implementovaný v R (pomocí `runif(n, min=0, max=1)` vygenerujete `n` náhodných čísel z intervalu `[min,max]`).

V R je jako defaultní generátor použit „Mersenne twister“, algoritmus navržený dvojicí Matsumoto a Nishimura v roce 1998 [5].

6.2 Diskrétní rozdělení

Pro generování (pseudo)náhodných čísel z často používaných distribucí jsou v R k dispozici funkce, např. pro binomické rozdělení **r**binom. Stejně tak můžeme využít distribuční funkci **p**binom, pravděpodobnostní funkci **d**binom a kvantilovou funkci **q**binom. Stejný způsob tvoření příkazů pro uvedené funkce platí i pro ostatní rozdělení. Jména rozdělení a jejich parametry (společné všem čtyřem funkcím) shrnuje tabulka 6.2. V tabulce nejsou uvedena pouze diskrétní rozdělení, ale i význačná spojitá.

▷ **Příklad 6.1.** Jak vypadá rozdělení pavděpodobností, tj. pravděpodobnostní funkce, pro náhodnou veličinu, která má binomické rozdělení $X \sim Bi(n, p)$ měníme-li p, tj. pravděpodobnost úspěchu?

Rozdělení	Jméno v R	Parametry
Beta	beta	shape1,shape2
Binomické	binom	size,prob
χ^2	chisq	df
Exponenciální	exp	rate
F	f	df1,df2
Bamma	gamma	shape, rate
Beometrické	geom	prob
Negativní binomické	pnbinom	size, prob
Normální	normal	mean, sd
Poissonovo	pois	lambda
Studentovo	t	df
Uniformní	unif	min, max

Tabulka 1: Některá rozdělení a jejich reprezentace v R

```
binom<-function(p){
plot(0:10, dbinom(0:10, 10,p),ylim=c(0,1))
Sys.sleep(1)             #Zajistí pauzu mezi kreslenými grafy
}
sapply((0:20)/20,binom)  #Opakované volání funkce binom s různými parametry
```

Zkuste vykreslit distrubuční funkci s měnící se pravděpodobností úspěchu. Jak vypadá distribuční funkce pro $p = 0$ a pro $p = 1$? Vysvětlete. Zkuste vykreslit pravděpodobnostní i distribuční funkci při měnícím se n a konstantním p. n volte v rozmezí od 1 do 100, vhodně upravte rozsah os.

▷ **Příklad 6.2.** Házíme kostkou a zajímá nás po kolika hodech padne první šestka. Úkolem je simulovat tuto náhodnou veličinu a porovnat výsledky pro N opakování s teoretickým rozdělením, které by tato náhodná veličina měla mít.
Označme X náhodnou veličinu „počet hodů než uvidíme šestku“. Celý pokus je posloupností nezávislých Bernoulliho pokusů, kde pravděpodobnost úspěchu (padne šestka) je $p = 1/6$ a pravděpodobnost neúspěchu je pak $1 - p = 5/6$. Potřebujeme tedy generovat čísla 0 a 1 (neúspěch, úspěch) s pravděpodobností 5/6 a 1/6. Toho můžeme docílit pomocí již zmíněné funkce `sample`, nebo můžeme použít generátor náhodných čísel z binomického rozdělení `rbinom(n, size, prob)`. V posloupnosti vygenerovaných čísel nás zajímá pozice, kdy padla první 1. Náhodná veličina X může obecně nabývat hodnot $\{1, 2, \ldots\}$. My ale potřebujeme konečný počet pozic k. k zvolíme tak, aby pravděpodobnost, že první úspěch nastane až po k hodech byl blízko 0. $P\{X > k\} = 1 - P\{X \leq k\}$. Náhodná veličina X má geometrické rozdělení a k určení $P\{X \leq k\}$ použijeme funkci `pgeom(x, prob)`. Pozor: geometrické rozdělení je v R definováno jako počet neúspěšných pokusů před prvním úspěchem, nikoliv jako počet opakování než nastane první úspěch.

```
> pgeom(49,1/6)
[1] 0.9998901
> pgeom(99,1/6)
[1] 1
```

Vidíme, že $P\{X \leq 50\} = 0.9998901$ a že počet opakování hodu kostkou do padnutí první 6 nepřekročí číslo 100. Zvolíme tedy $k = 100$.

```
rm(list = ls())
graphics.off()
x<-seq(1,100,1)
f=function(){
```

```
  gen<-rbinom(100,1,1/6)     #generováno 100 hodnot z binomického rozdělení
  x[gen==1][1]               #nalezení pozice, kde padla první jednička
}
#pokus opakujeme N-krát a výsledky uložíme do vektoru X
N=10000
X=numeric(N)
for(i in 1:N){
  X[i] = f()
}
#Hodnoty náhodné veličiny X můžeme zobrazit jednoduše v tabulce
val<-tabulate(X,nbins=max(X))
#a zakreslit do grafu
plot(1:max(X),val)
```

Opakované volání funkce lze řešit i jinak než for-cyklem, například pomocí funkce `replicate()`. Vyzkoušejte.
Nyní vytvoříme histogram hodnot náhodné veličiny X a porovnáme s pravděpodobnostní funkcí pro geometrické rozdělení, kterou pro názornost zakreslíme jako spojitou funkci, nikoliv jako diskrétní (parametr type="l").

```
y<-0:(max(X)-1)
bins=seq(0.5,(max(X)+0.5),1)
hist(X,freq=FALSE,breaks=bins,xlab="X")
points(y,dgeom(y,prob=1/6),type="l",col="red")
```

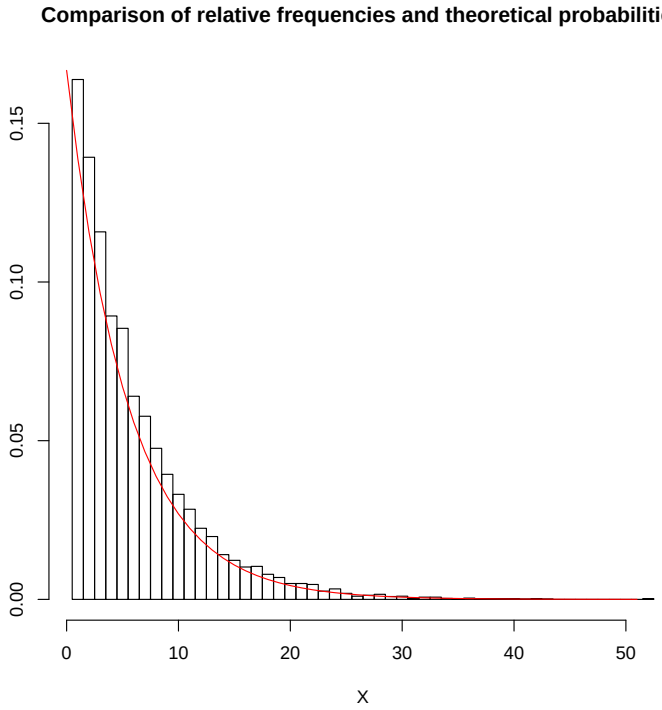
► **Příklad 6.2.** Pro náhodnou veličinu X z předchozího příkladu určete střední hodnotu, rozptyl a směrodatnou odchylku. Porovnejte s výběrovým průměrem, výběrovým rozptylem a výběrovou směrodatnou odchylkou. Výběrový průměr je tzv. „dobrý“ odhad střední hodnoty. Co přesně znamená „dobrý“ se dozvíte později.

► **Příklad 6.3.** Uvažujte následující hru. Vyhrajete 9 $ pokud při hodu 2 kostkami padne součet 2,3,11 nebo 12. Prohrajete 10 $ pokud padne součet 7. Jinak nevyhrajete nic. Určete $P_X((0,\infty))$ a $P_X((-\infty,0))$. [1/6, 1/6]

► **Příklad 6.4.** Za den přijde k jednomu určitému lékaři v průměru 10 lidí. Lékař je schopný ošetřit maximálně 15 pacientů za den. Jaká je pravděpodobnost, že v daný den bude muset lékař pacienty odmítnout? Jaká je pravděpodobnost, že v daný den přijde k lékaři právě 8 pacientů? [0.0487,0.1126]

► **Příklad 6.5.** Házíme kostkou a zajímá nás po kolika hodech padne k-tá šestka. Označme tuto náhodnou veličinu W_k. Pak náhodný jev $(W_k = n)$ značí „v n-tém hodu padla šestka a v předchozích $n-1$ hodech padla šestka právě $k-1$-krát.“ Náhodná veličina W_k má tzv. negativní binomické rozdělení.
Najděte pravděpodobnostní funkci pro náhodnou veličinu W_3, tj. obecný předpis pro $P(W_3 = n)$. V R simulujte tuto náhodnou veličinu N-krát pro $N = 100, N = 1000, N = 10000$. Srovnejte relativní četnosti hodnot náhodné veličiny s teoretickými pravděpodobnostmi.
Vykreslete pravděpodobnostní funkci a distribuční funkci pro negativní binomické rozdělení s různými parametry a funkce porovnejte.

► **Příklad 6.6.** Nicolas Bernoulli navrhl hazardní hru, při které se hází mincí dokud nepadne hlava. Pokud padne hlava v n-tém hodu, dostane hráč od kasína 2^{n-1} Kč. Původně šlo samozřejmě o jinou měnu. Pokud padne hlava v prvním hodu, hráč vyhrává 1 Kč, pokud např. padne hlava ve 4.hodu, hráč vyhraje $2^{4-1} = 8$ Kč.
Jaká cena bude pro hráče za tuto hru férová, tj. jaká je očekávaná hodnota výhry? $[+\infty]$.
Právě kvůli výsledku očekávané hodnoty $+\infty$ se této úloze říká Petrohradský paradox. V reálném životě by ale nikdo za tuto hru nezaplatil „všechno co má,“ neprodal dům či auto. V reálném životě



bychom museli uvažovat, že hra je konečná, omezená např. množstvím peněz, které je kasino schopné vyplatit, nebo omezená dobou života hráče. Uvažujte tedy, že zdroje kasina určené pro vyplácení výher z této hry jsou omezené, ozn. Z. Hráč při vstupu zaplatí částku E. Pro různá Z a E simulujte tuto hru (tj. náhodnou veličinu X-výhra hráče) v R, pro dostatečný počet opakování určete výběrový průměr, srovnejte relativní četnosti hodnot náhodné veličiny X s jejím teoretickým rozdělením.

► **Příklad 6.7.** Předpokládejte, že je náhodně a nezávisle na sobě vybráno 2000 bodů z jednotkového čtverce

$$\{\langle x, y\rangle \,|\, 0 \leq x \leq 1, 0 \leq y \leq 1\}.$$

Nechť W je počet bodů, které padnou do oblasti $A = \{\langle x, y\rangle \,|\, x^2 + y^2 < 1\}$. Určete pravděpodobnostní funkci náhodné veličiny W, její střední hodnotu a rozptyl. $[W \sim Bi(2000, \pi/4), 1570.8, 337.1]$

► **Příklad 6.8.** Nechť náhodná veličina X má Poissonovo rozdělení s parametrem λ. Dokažte, že

$$P_X(\{k+1\}) = \frac{\lambda}{k+1} P_X(\{k\})$$

Reference

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

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

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

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

[5] Matsumoto, Makoto and Nishimura, Takuji: Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.* 8(1):3-30, 1998.

[6] Maria L. Rizzo: Statistical Computing with R
Chapman and Hall/CRC, ISBN: 978-1584885450