Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

STATYSTYKA MATEMATYCZNA

z pakietem R

III. Rozkłady prawdopodobieństwa i podstawy symulacji

Przemysław Grzegorzewski
Konstancja Bobecka-Wesołowska
Marek Gągolewski

Spis treści

1 Wprowadzenie

1.1 Wybrane rozkłady prawdopodobieństwa

W programie `R` dostępne są funkcje związane z podstawowymi rozkładami prawdopodobieństwa, tj.

Rozkład	Nazwa	Parametry	Identyfikator
$\mathrm{Bin}(n,p)$	dwumianowy (ang. *binomial*)	$n \in \mathbb{N}$, $p \in (0,1)$	`*binom`
$\mathrm{Geom}(p)$	geometryczny	$p \in (0,1)$	`*geom`
$\mathrm{Hyp}(m,n,k)$	hipergeometryczny	$m,n,k \in \mathbb{N}$, $k \leqslant m$	`*hyper`
$\mathrm{NegBin}(n,p)$	ujemny dwumianowy (ang. *negative binomial*)	$n \in \mathbb{N}$, $p \in (0,1)$	`*nbinom`
$\mathrm{Poi}(\lambda)$	Poissona	$\lambda > 0$	`*pois`
$\mathrm{B}(a,b)$	beta	$a > 0$, $b > 0$	`*beta`
$\mathrm{C}(l=0,s=1)$	Cauchy'ego	$l \in \mathbb{R}$, $s > 0$	`*cauchy`
${\chi^2}_d$	chi-kwadrat (ang. *chi-square*)	$d \in \mathbb{N}$	`*chisq`
$\mathrm{Exp}(\lambda=1)$	wykładniczy (ang. *exponential*)	$\lambda > 0$	`*exp`
$\mathrm{F}^{[d_1,d_2]}$	F-Snedecora	$d_1, d_2 \in \mathbb{N}$	`*f`
$\Gamma(a,s)$	gamma	$a > 0$, $s > 0$	`*gamma`
$\mathrm{Logis}(\mu=0,s=1)$	logistyczny	$\mu \in \mathbb{R}$, $s > 0$	`*logis`
$\mathrm{LogN}(\mu=0,\sigma=1)$	logarytmiczno-normalny	$\mu \in \mathbb{R}$, $\sigma > 0$	`*lnorm`
$\mathrm{N}(\mu=0,\sigma=1)$	normalny	$\mu \in \mathbb{R}$, $\sigma > 0$	`*norm`
$\mathrm{U}([a=0,b=1])$	jednostajny (ang. *uniform*)	$a < b$	`*unif`
$\mathrm{t}^{[d]}$	t-Studenta	$d \in \mathbb{N}$	`*t`
$\mathrm{Wei}(a,s=1)$	Weibulla	$a > 0$, $s > 0$	`*weibull`

Nazewnictwo tych funkcji jest ujednolicone. By uzyskać odpowiednią wartość, należy zastąpić znak `*` przed identyfikatorem rozkładu odpowiednią literą, wg wzoru:

Przedrostek	Znaczenie
`d`	gęstość (ang. *density*) $f(x)$ lub rozkład prawdopodobieństwa $P(X = x)$
`p`	dystrybuanta (ang. *distribution function*) $F(x) = P(X \leqslant x)$
`q`	funkcja kwantylowa (ang. *quantile function*) $\simeq F^{-1}(p)$
`r`	generowanie liczb pseudolosowych (ang. *random deviates generation*)

Uwaga

Zwróćmy uwagę, że w przypadku kilku rozkładów niektóre parametry posiadają, dla wygody, wartości domyślne. Zostały one przedstawione w kolumnie *Rozkład*.

1.1.1 Dystrybuanta

Jeśli dla wybranego rozkładu prawdopodobieństwa chcemy obliczyć wartość dystrybuanty w danym punkcie x, to wystarczy przed identyfikatorem rozkładu postawić przedrostek `p`, np.

```
> pnorm(0) # wartość dystrybuanty rozkładu standardowego normalnego w punkcie 0
```

```
[1] 0.5
```

Zgodnie z R-ową zasadą działania na wektorach, pierwszym argumentem tej funkcji może być też wektor liczb, dla których chcemy obliczyć wartość dystrybuanty:

```
> pnorm(c(1,2,3)) # a teraz wartość dystrybuanty w punktach 1,2 oraz 3

[1] 0.8413447 0.9772499 0.9986501
```

Pozostałe argumenty tej funkcji określają parametry rozkładu, dla którego chcemy znaleźć wartość dystrybuanty, np.

```
> pnorm(0, 2, 1) # wartość dystrybuanty rozkładu N(2,1) w punkcie 0

[1] 0.02275013
```

Jeśli, zamiast wartości dystrybuanty w danym punkcie x, chcemy wyznaczyć wartość funkcji przeżycia w punkcie x (czyli wartość $G(x) = 1 - F(x) = P(X > x)$), to podajemy dodatkowy argument: `lower.tail=F`, np.

```
> pnorm(0, lower.tail=F) # wartość funkcji przeżycia rozkładu N(0,1) w punkcie 0

[1] 0.5
```

Oczywiście to samo mogliśmy obliczyć, pisząc:

```
> 1-pnorm(0)

[1] 0.5
```

1.1.2 Gęstość i prawdopodobieństwo

Przedrostek `d` przed identyfikatorem rozkładu prawdopodobieństwa określa funkcję liczącą wartość gęstości (w przypadku rozkładów absolutnie ciągłych) lub prawdopodobieństwa (w przypadku rozkładów dyskretnych) w danym punkcie x (lub w punktach zadanych przez elementy wektora wejściowego), np.

```
> dexp(0) # wartość f(0), gdzie f jest gęstością rozkładu Exp(1)

[1] 1


> dexp(c(0,0.5,1), 0.5) # wartości f(0), f(0.5), f(1) dla rozkładu Exp(0.5)

[1] 0.5000000 0.3894004 0.3032653


> pr <- dbinom(0:8, 8, 0.25); # wartości Pr(X=i) dla X~Bin(8, 1/4), i=0,1,...,8
> round(pr, 3); # wyświetl zaokrąglone do 3 miejsc po przecinku

[1] 0.100 0.267 0.311 0.208 0.087 0.023 0.004 0.000 0.000
```

1.1.3 Funkcja kwantylowa

Wartości teoretycznych kwantyli wyznaczamy, pisząc przed nazwą rozkładu przedrostek q, przy czym pierwszym argumentem tej funkcji jest rząd kwantyla, np.

```
> qt(0.95, 5) # kwantyl rzędu 0.95 rozkładu t o 5 stopniach swobody

[1] 2.015048


> qt(0.95, c(1,5,10,15)) # różne stopnie swobody

[1] 6.313752 2.015048 1.812461 1.753050


> qt(0.95, Inf) # to jest rozkład normalny standardowy

[1] 1.644854


> qnorm(0.95)

[1] 1.644854


> qt(0.95, 1) # rozkład Cauchy'ego standardowy

[1] 6.313752


> qcauchy(0.95)

[1] 6.313752


> qt(c(0.95, 0.975, 0.99, 0.995), 5) # różne rzędy kwantyla

[1] 2.015048 2.570582 3.364930 4.032143


> qt(c(0.95, 0.975, 0.99, 0.995), c(1,5,10,15)) # a to co?

[1] 6.313752 2.570582 2.763769 2.946713
```

W przypadku rozkładów dyskretnych funkcja kwantylowa zmiennej X w punkcie q zwraca najmniejszą wartość $x \in \mathrm{supp}(X)$, dla której $P(X \leqslant x) \geqslant q$, gdzie $\mathrm{supp}(X)$ oznacza nośnik rozkładu zmiennej X.

```
> pbinom(0:5, 5, 0.5) # dla porównania

[1] 0.03125 0.18750 0.50000 0.81250 0.96875 1.00000


> qbinom(c(0.4, 0.5, 0.6), 5, 0.5)

[1] 2 2 3
```

1.1.4 Generowanie liczb pseudolosowych

Generowanie liczb pseudolosowych[1] z danego rozkładu prawdopodobieństwa uruchamiamy, pisząc przed nazwą rozkładu przedrostek `r`, przy czym pierwszym argumentem tej funkcji jest liczba wartości, które chcemy wygenerować, np.

```
> runif(5) # wygenerowanie 5 obserwacji z rozkładu jednostajnego na [0,1]

[1] 0.3364676 0.2858941 0.8558148 0.3210381 0.7831911


> runif(10,0,5) # wygenerowanie realizacji 5-elementowej próby z rozkładu U([0,5])

 [1] 1.3949899 1.6287845 4.7443366 0.3689878 3.5892525 2.8202595 2.8934075
 [8] 1.8794554 4.0163759 1.9004927


> rpois(20, 4) # wygenerowanie 20 obserwacji z rozkładu Poi(4)

 [1] 6 2 4 1 6 3 1 4 5 4 4 3 5 6 3 5 8 4 5 6
```

1.2 Losowanie bez i ze zwracaniem

Do generowania próbek będących wynikiem n-krotnego losowania elementów danego zbioru S można użyć funkcji `sample()` (S jest pierwszym argumentem tej funkcji, n — drugim). Domyślnie otrzymujemy wynik losowania bez zwracania. Gdy chcemy losować ze zwracaniem, to jako kolejny argument funkcji `sample()` piszemy `replace=TRUE`.

Na przykład wynik $n = 15$ rzutów monetą można otrzymać następująco:

```
> sample (c("O","R"), 15, replace=TRUE);

 [1] "R" "R" "R" "R" "R" "R" "R" "O" "O" "O" "R" "O" "R" "O" "R"
```

Parametr n można także pominąć — poniższy kod wygeneruje nam losową permutację zbioru $\{1, 2, \ldots, 10\}$.

```
> sample(1:10); # losowanie bez zwracania 10 elementów z 10-elementowego zbioru

 [1]  9  5  8  4  7  1  6 10  2  3
```

[1] Czyli będących rezultatem wykonania pewnego ciągu ściśle deterministycznych operacji arytmetycznych. Operacje te dają w wyniku liczby *przypominające* liczby losowe z zadanego rozkładu, tzn. spełniające pewne statystyczne kryteria, których spełniania spodziewalibyśmy się po liczbach „prawdziwie" losowych (np. powstałych w wyniku pomiaru fizycznego, rzutu monetą itp.). Dla wygody dalej często pomijać będziemy przedrostek „pseudo" i mówić po prostu o liczbach losowych.

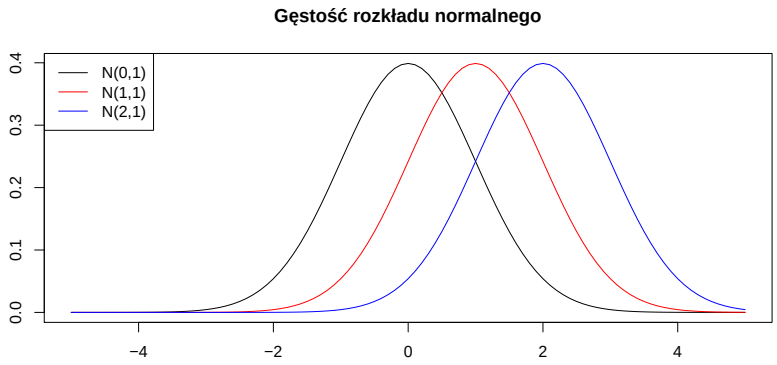
2 Zadania rozwiązane

Zadanie 3.1. Utwórz wykresy gęstości, dystrybuanty i funkcji przeżycia dla zmiennych losowych z rozkładów normalnych o parametrach $\mathrm{N}(0,1)$, $\mathrm{N}(1,1)$, $\mathrm{N}(2,1)$.

Rozwiązanie.
Do rysowania wykresów użyjemy funkcji `curve(f(x), from=x1, to=x2)`. Próbkuje ona w wielu punktach `x` z przedziału $[\texttt{x1}, \texttt{x2}]$ wartości funkcji[2] `f` i przedstawia je na rysunku (podobnie jak funkcja `plot(..., type="l")`).

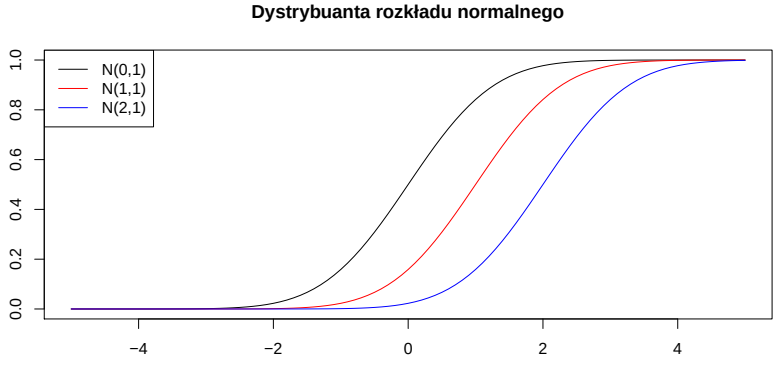
Gęstości rozkładów normalnych o różnych parametrach położenia:

```
> curve(dnorm(x), from=-5, to=5, col=1, main="Gęstość rozkładu normalnego")
> curve(dnorm(x, 1, 1), col=2, add=T)  # dodanie kolejnej krzywej
> curve(dnorm(x, 2, 1), col=4, add=T)  # i jeszcze jednej
> legend("topleft", c("N(0,1)", "N(1,1)", "N(2,1)"), col=c(1,2,4), lty=1);
```



W podobny sposób możemy sporządzić wykresy dystrybuant rozkładów normalnych o różnych parametrach położenia:

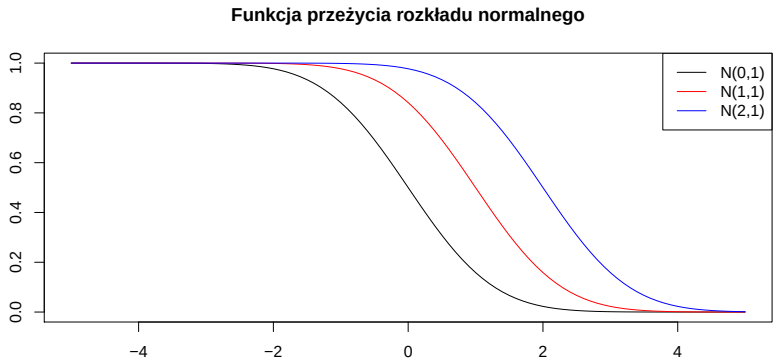
```
> curve(pnorm(x), from=-5, to=5, col=1, main="Dystrybuanta rozkładu normalnego")
> curve(pnorm(x, 1, 1), col=2, add=T)
> curve(pnorm(x, 2, 1), col=4, add=T)
> legend("topleft", c("N(0,1)", "N(1,1)", "N(2,1)"), col=c(1,2,4), lty=1);
```



Wykresy funkcji przeżycia rozkładów normalnych o różnych parametrach położenia:

[2] Bądź wyrażenia, w którym pojawia się `x`.

```
> curve(pnorm(x, lower.tail=F), from=-5, to=5, col=1,
     main="Funkcja przeżycia rozkładu normalnego")
> curve(pnorm(x, 1, 1, lower.tail=F), col=2, add=T)
> curve(pnorm(x, 2, 1, lower.tail=F), col=4, add=T)
> legend("topright", c("N(0,1)", "N(1,1)", "N(2,1)"), col=c(1,2,4), lty=1);
```



⊡

Zadanie 3.2. Sprawdź tzw. regułę 3-sigmową dla rozkładu normalnego. Utwórz graficzną ilustrację tej reguły.

Rozwiązanie.
Reguła trzech sigm dla $X \sim \mathrm{N}(\mu, \sigma)$:

$$P\left(\mu - 3\sigma \leqslant X \leqslant \mu + 3\sigma\right) \simeq 0{,}99730. \tag{1}$$

Intuicyjnie, głosi ona, że prawie cała masa prawdopodobieństwa zmiennej losowej X zawiera się w przedziale $[\mu - 3\sigma, \mu + 3\sigma]$.

Obliczmy to prawdopodobieństwo (czyli np. $P\left(X \in [-3, 3]\right)$, gdzie $X \sim \mathrm{N}(0, 1)$) za pomocą R-a:

```
> pnorm(3)-pnorm(-3)

[1] 0.9973002
```

Reguła 3-sigmowa dla rozkładu $\mathrm{N}(\mu, \sigma)$ może być przykładowo zilustrowana (np. w podręczniku do rachunku prawdopodobieństwa) w sposób następujący.

1. Narysujemy funkcję gęstości.
2. Pokolorujemy odpowiedni fragment pola pod krzywą (za pomocą funkcji `polygon()`, służącej do rysowania wielokątów).
3. Umieścimy odpowiednie etykietki tekstowe (funkcje `arrows()`, `text()`).

Dokonywać będziemy obliczeń dla rozkładu standaryzowanego. Zachęcamy Czytelnika do samodzielnego przestudiowania stron systemu pomocy dotyczących używanych funkcji graficznych.

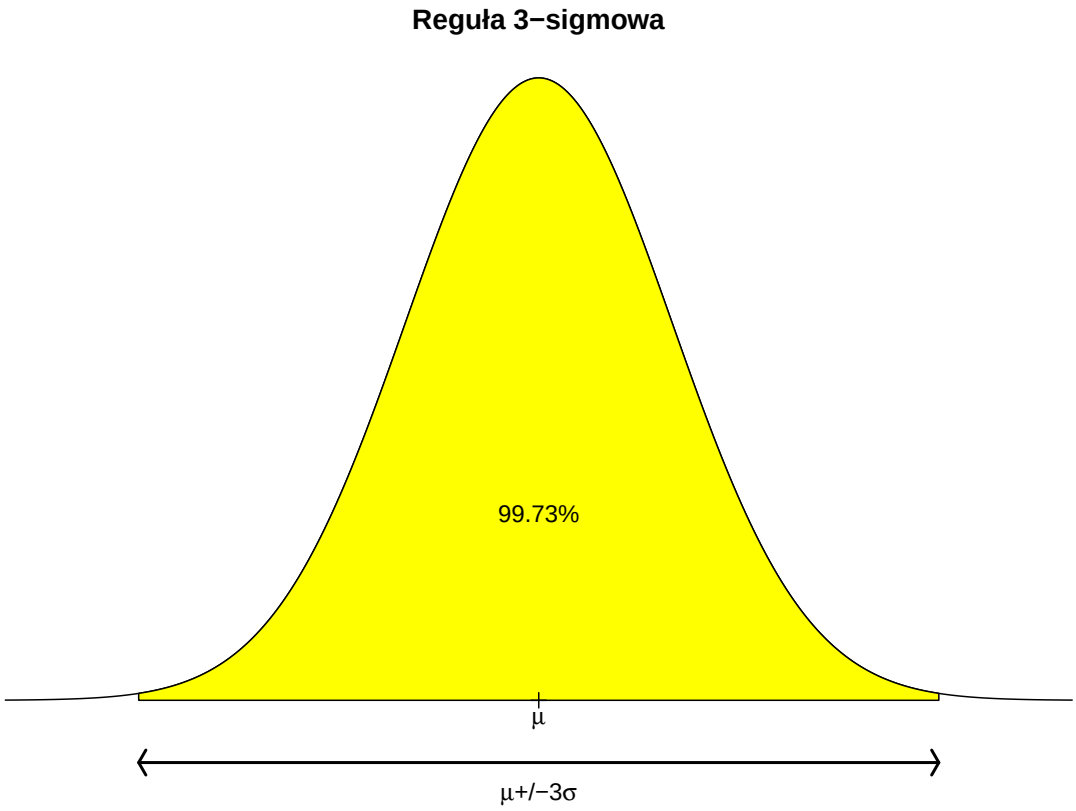
```
x <- seq(-5, 5, by=0.01)      # wektor argumentów, dla których obliczymy wartości
                              # gęstości rozkładu N(0,1)
y <- dnorm(x)
plot(x, y, type="l", main="Reguła 3-sigmowa", xlab=NA, ylab=NA, axes=F,
   ylim=c(-max(y)*0.2, max(y)))          # wykres gęstości (p. 1.)
```

Definiujemy współrzędne punktów, które będą wierzchołkami wielokąta, który następnie wypełnimy kolorem żółtym (p. 2.):

```
wx <- c(-3, x[x>=-3 & x<=3], 3);
wy <- c( 0, y[x>=-3 & x<=3], 0);
polygon(wx, wy, col="yellow");
```

Dodajemy oznaczenie punktu $x = \mu$, rysujemy podpisaną strzałkę wskazującą interesujący nas przedział na osi poziomej i dodajemy informację o wartości pola pod fragmentem krzywej (p. 3.):

```
points(0, 0, pch=3) # Oznaczenie punktu x=mi
text(0, 0, expression(paste(mu)), pos=1)
# Generujemy strzałkę dla przedziału:
arrows(-3,-max(y)*0.1,3,-max(y)*0.1,angle=60, length=0.1,code=3,lwd=2)
text(0, -max(y)*0.15, # Etykietka pod strzałką
 expression(paste(mu,"+/- 3",sigma))) # expression() generuje litery greckie
text(0, max(y)*0.3, paste(round(100*(pnorm(3)-pnorm(-3)), 2), "%", sep=""))
```



□

Zadanie 3.3. Wzrost pewnej grupy osób opisany jest rozkładem normalnym o wartości oczekiwanej 173 cm i odchyleniu standardowym 6 cm.

a) Jakie jest prawdopodobieństwo, że losowo wybrana osoba ma nie więcej niż 179 cm wzrostu?

b) Jaka jest frakcja osób mających wzrost pomiędzy 167 i 180 cm?

c) Jakie jest prawdopodobieństwo, że losowo wybrana osoba ma więcej niż 181 cm wzrostu?

d) Wyznacz wartość wzrostu, której nie przekracza 60% badanej populacji osób.

Rozwiązanie.
Załóżmy, że mamy do czynienia ze zmienną losową $X \sim \mathrm{N}(173, 6)$, opisującą wzrost losowo napotkanej osoby z danej grupy. W pierwszym przypadku obliczamy $P(X \leqslant 179)$:

```
> pnorm(179, 173, 6);

[1] 0.8413447
```

Dalej obliczamy $P(167 < X \leqslant 180)$:

```
> pnorm(180, 173, 6)-pnorm(167, 173, 6);

[1] 0.7196722
```

Trzecie pytanie dotyczy $P(X > 181)$:

```
> 1-pnorm(181, 173, 6); # lub równoważnie:

[1] 0.09121122


> pnorm(181, 173, 6, lower.tail=F);

[1] 0.09121122
```

W ostatnim zagadnieniu szukamy kwantyla $q_{0.6}$ rozkładu $\mathrm{N}(173, 6)$:

```
> qnorm(0.6,173,6)

[1] 174.5201
```

□

Zadanie 3.4. Utwórz tablicę wartości dystrybuanty rozkładu standardowego normalnego.

Rozwiązanie.
Stworzymy tablicę z wartościami dystrybuanty $\Phi(x)$ rozkładu $\mathrm{N}(0, 1)$ dla $x \in \{0, 0{,}2, \ldots, 3{,}8\}$.

```
> x <- seq(0, 3.8, 0.2);
> y <- pnorm(x);
> length(x)
```

```
[1] 20
```

Następnie przekonwertujemy wektor y na macierz. Można to zrobić w następujący sposób:

```
> dim(y)=c(5,4)
> y
```

```
          [,1]      [,2]      [,3]      [,4]
[1,] 0.5000000 0.8413447 0.9772499 0.9986501
[2,] 0.5792597 0.8849303 0.9860966 0.9993129
[3,] 0.6554217 0.9192433 0.9918025 0.9996631
[4,] 0.7257469 0.9452007 0.9953388 0.9998409
[5,] 0.7881446 0.9640697 0.9974449 0.9999277
```

Zauważmy, że w naszej tablicy w pierwszym wierszu mamy wartości $\Phi(x)$ dla $x \in \{0,0{,}2, 0{,}4, 0{,}6, 0{,}8\}$, w drugim — dla $x \in \{1, 1{,}2, 1{,}4, 1{,}8\}$ itd. Jeśli chcemy mieć tablicę do czytania „wierszami" możemy zrobić np. tak:

```
> matrix(y, 4, 5, byrow=T)
```

```
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.5000000 0.5792597 0.6554217 0.7257469 0.7881446
[2,] 0.8413447 0.8849303 0.9192433 0.9452007 0.9640697
[3,] 0.9772499 0.9860966 0.9918025 0.9953388 0.9974449
[4,] 0.9986501 0.9993129 0.9996631 0.9998409 0.9999277
```

Dodatkowo, dobrze będzie nadać nazwy poszczególnym wierszom i kolumnom naszej tablicy:

```
> matrix(y, 4, 5, byrow=T, dimnames=list(c(0,1,2,3),c(0.0,0.2,0.4,0.6,0.8)));
```

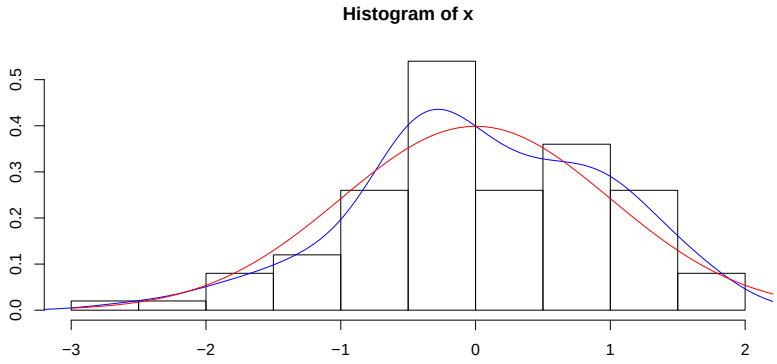
```
          0       0.2       0.4       0.6       0.8
0 0.5000000 0.5792597 0.6554217 0.7257469 0.7881446
1 0.8413447 0.8849303 0.9192433 0.9452007 0.9640697
2 0.9772499 0.9860966 0.9918025 0.9953388 0.9974449
3 0.9986501 0.9993129 0.9996631 0.9998409 0.9999277
```

□

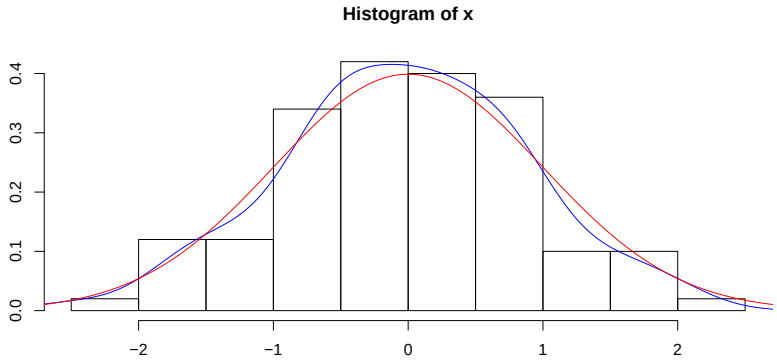
Zadanie 3.5. Wygeneruj n-elementową ($n = 100$) próbę losową z rozkładu normalnego standardowego. Utwórz histogram oraz estymator jądrowy dla tej próby. Nałóż na uzyskany obraz wykres gęstości teoretycznej rozkładu normalnego.

Rozwiązanie.
Rozwiązanie tego zadania jest bardzo proste:

```
> n <- 100;
> x <- rnorm(n); # próba losowa z rozkładu N(0,1)
> hist(x, prob=T)
> lines(density(x), col="blue")
> curve(dnorm(x), from=-3, to=3, col="red", add=T)
```



Oczywiście, kolejna wygenerowana próba będzie (z prawdopodobieństwem 1) składała się z innych obserwacji. Warto więc przyjrzeć się wykresom dla kilku realizacji tego eksperymentu, wywołując powyższy kod kilkakrotnie.



□

Zadanie 3.6. Sporządź wykres funkcji masy prawdopodobieństwa rozkładów dwumianowych: $\mathrm{Bin}(10, 0{,}5), \mathrm{Bin}(10, 0{,}25), \mathrm{Bin}(50, 0{,}25)$.

Rozwiązanie.
Obliczamy najpierw $P(X = k)$ dla $k = 0, 1, \ldots, 10$ dla $X \sim \mathrm{Bin}(10, 0{,}5)$:

```
> x <- dbinom(0:10, 10, 0.5)
```

Podobnie czynimy dla pozostałych rozkładów:

```
> y <- dbinom(0:10, 10, 0.5)
> z <- dbinom(0:50, 50, 0.25)
```

Rysujemy funkcje masy prawdopodobieństwa tych rozkładów jako wykresy słupkowe.

```
> barplot(x, names.arg=0:10)
> barplot(y, names.arg=0:10)
> barplot(z, names.arg=0:50)
```

Sporządzenie wynikowych wykresów i wyciągnięcie wniosków pozostawiamy Czytelnikowi.
□

Zadanie 3.7. Korzystając z generatora liczb losowych o rozkładzie jednostajnym na przedziale $[0,1]$, wygeneruj próbkę losową z rozkładu Pareto z parametrem $a = 2$.

Rozwiązanie.

Metoda odwracania dystrybuanty

Korzystamy tu z następującego faktu. Niech F oznacza dystrybuantę zmiennej losowej X. Definiujemy funkcję kwantylową F^{-1} (lub uogólnioną funkcję odwrotną):

$$F^{-1}(x) = \inf\{t : F(t) \geqslant x\}. \tag{2}$$

Wtedy zmienna losowa X ma taki sam rozkład jak $F^{-1}(U)$, gdzie U jest zmienną losową o rozkładzie jednostajnym $\mathrm{U}([0,1])$. Jeśli F jest dystrybuantą ciągłą, to $F(X)$ ma taki sam rozkład jak U. Zauważmy przy tym, że gdy F jest ciągła i rosnąca, to F^{-1} jest funkcją odwrotną do F w zwykłym sensie.

Powyższy sposób konstruowania generatorów liczb losowych z zadanego rozkładu na podstawie generatora dla rozkładu $\mathrm{U}([0,1])$ nazywamy *metodą odwracania dystrybuanty*.

Zmienna losowa X o rozkładzie Pareto z parametrem a ma rozkład o gęstości:

$$f(x) = \frac{a}{x^{a+1}}, \tag{3}$$

dla $x > 1$. Dystrybuanta ma postać:

$$F(x) = (1 - 1/x^a), \tag{4}$$

stąd

$$F^{-1}(x) = (1-x)^{-1/a}, \tag{5}$$

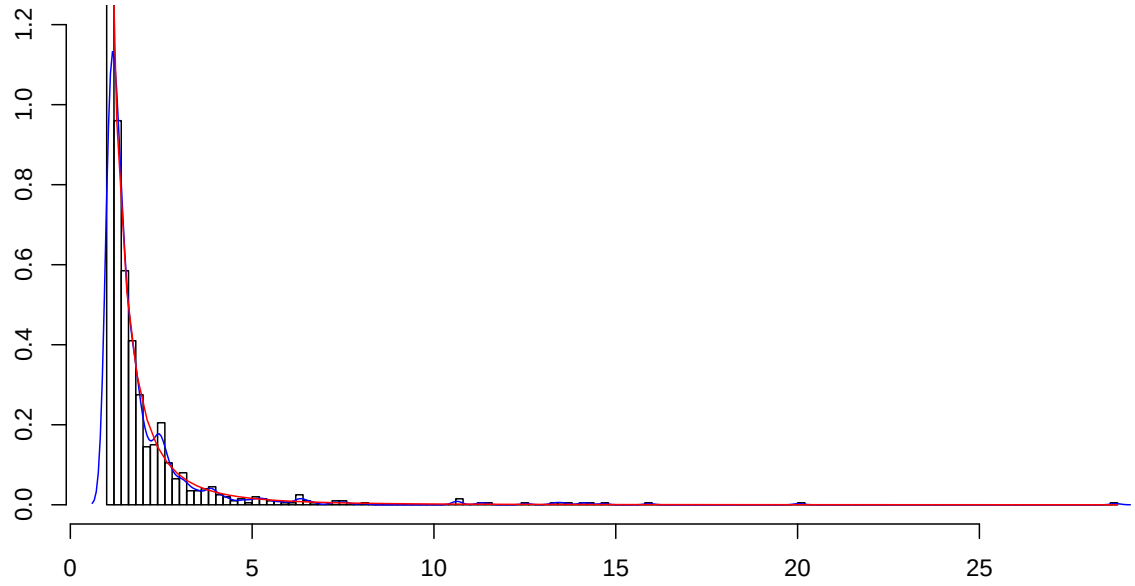
czyli zmienna losowa $F^{-1}(U) = (1-U)^{-1/a}$ dla $a = 2$, gdzie $U \sim \mathrm{U}([0,1])$, będzie miała interesujący nas rozkład Pareto o dystrybuancie F.

Próbkę losową generujemy zatem w sposób następujący:

```
> n <- 1000;
> u <- runif(n);
> x <- u^(-0.5); # rozkład 1-u jest taki sam jak rozkład u
```

Narysujmy teraz histogram dla naszej próbki, wykres estymatora jądrowego gęstości oraz gęstości teoretycznej:

```
> hist(x, prob=T, main=NA, ylim=c(0.0, 1.2), breaks=100);
> lines(density(x), col="blue");
> curve(2/x^3, add=T, col="red", from=1);
```



□

Zadanie 3.8. Posługując się metodą Monte Carlo, oblicz pole powierzchni obszaru $A = \{(x, y) \in \mathbb{R}^2 : 0 < x < 1; 0 < y < x^2\}$.

Porównaj uzyskane w ten sposób wyniki z dokładnymi rezultatami otrzymanymi na drodze analitycznej.

Rozwiązanie.

Metoda całkowania Monte Carlo

Skorzystamy tu z następującego faktu. Niech $X_1, Y_1, X_2, Y_2, \ldots$ będą niezależnymi zmiennymi losowymi o rozkładzie jednostajnym $U([0, 1])$. Dla funkcji borelowskiej $f : [0, 1] \to [0, 1]$ definiujemy

$$Z_i = \mathbf{1}\left(Y_i \leqslant f(X_i)\right), \tag{6}$$

gdzie $\mathbf{1}(\cdot)$ jest funkcją indykatorową. Wówczas, z mocnego prawa wielkich liczb (MPWL), mamy z prawdopodobieństwem równym 1:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = \int_0^1 f(x)\, dx. \tag{7}$$

Uwaga

Uogólnienie tej metody na funkcje zdefiniowane dla innych dziedzin i przeciwdziedzin (przedziałowych) pozostawiamy jako ćwiczenie dla Czytelnika.

Pole obszaru A to

$$\int_A dx\, dy = \int_0^1 \left(\int_0^{x^2} dy \right) dx = \int_0^1 x^2\, dx = \frac{1}{3}.$$

Wyznaczymy zatem pole obszaru A, obliczając w sposób przybliżony całkę $\int_0^1 x^2\, dx$ powyższą metodą, zwaną *całkowaniem Monte Carlo*[3].

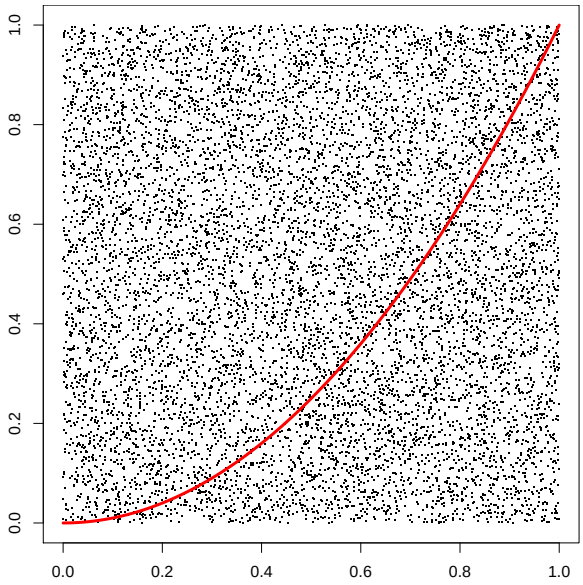
Generujemy najpierw próbkę losową $(U_1, V_1, \ldots, U_n, V_n)$ z rozkładu jednostajnego:

[3]Metoda całkowania Monte Carlo została zaproponowana przez polskiego matematyka Stanisława Ulama, biorącego udział w tzw. projekcie Manhattan.

```
> n <- 10000;
> u <- runif(n);
> v <- runif(n);
```

Zaznaczmy te punkty na obrazku i nałóżmy na niego wykres funkcji $y = x^2$ dla $x \in [0, 1]$.

```
> plot(u, v, xlim=c(0,1), ylim=c(0,1), pch='.') # punkty oznaczamy "kropką"
> curve(x*x, col="red", type="l", lwd=3, add=T)
```



Zliczmy teraz punkty z naszej próbki, które znalazły się pod wykresem funkcji $y = x^2$:

```
> z <- (v <= u*u);
> sum(z); # przypominamy, TRUE ma wartość 1, FALSE - 0

[1] 3397
```

Pole interesującego nas obszaru wynosi więc w przybliżeniu:

```
> mean(z);

[1] 0.3397
```

Uwaga

W programie R dostępna jest także funkcja do całkowania numerycznego o nazwie `integrate()`.

```
> integrate(dnorm, -1.96, 1.96)

0.9500042 with absolute error < 1.0e-11


> pnorm(1.96)-pnorm(-1.96)

[1] 0.9500042


> integrate(dnorm, -Inf, Inf)
```

```
1 with absolute error < 9.4e-05
```

`integrate()` jako pierwszy argument przyjmuje funkcję do scałkowania. Chcąc policzyć całki występujące w naszym zadaniu, należy stworzyć własną funkcję. Czyni się to w R wg następującej składni:

Własne funkcje

```
> nazwaFunkcji <- function(argument1, argument2, (...))
{
   (...) różne operacje (...)
   return(wynik);
}
```

Spróbujmy więc:

```
> funkcjaA <- function(x) { return(x^2); } # stworzenie własnej funkcji
> integrate(funkcjaA, 0, 1);

0.3333333 with absolute error < 3.7e-15
```

□

3 Zadania do rozwiązania

Zadanie 3.9. Utwórz wykresy gęstości, dystrybuanty i funkcji przeżycia dla zmiennych losowych z rozkładów normalnych o parametrach $N(0,1)$, $N(0,0{,}5)$, $N(0,2)$.

Zadanie 3.10. Utwórz tablicę podstawowych kwantyli (tzn. rzędu $0{,}9, 0{,}95, 0{,}975, 0{,}99, 0{,}995$) rozkładu standardowego normalnego.

Zadanie 3.11. Utwórz wykresy gęstości zmiennych losowych o rozkładzie chi-kwadrat o 5, 10 oraz 40 stopniach swobody. Przeanalizuj, jak zmienia się gęstość rozkładu χ^2 wraz ze wzrostem liczby stopni swobody.

Zadanie 3.12. Utwórz tablicę podstawowych kwantyli rozkładu chi-kwadrat o różnych stopniach swobody (tzn. kwantyli rzędu $0{,}005, 0{,}01, 0{,}025, 0{,}05, 0{,}1, 0{,}9, 0{,}95, 0{,}975, 0{,}99, 0{,}995$).

⋆ **Zadanie 3.13.** Przeprowadź eksperyment symulacyjny pokazujący, że rozkład chi-kwadrat, wraz ze wzrostem liczby stopni swobody, zbiega do rozkładu normalnego.

Zadanie 3.14. Utwórz tablicę podstawowych kwantyli (tzn. rzędu $0{,}9, 0{,}95, 0{,}975, 0{,}99, 0{,}995$) rozkładu t-Studenta o różnych stopniach swobody.

Zadanie 3.15. Utwórz wykresy gęstości zmiennych losowych o rozkładzie t-Studenta z 1, 5 i 30 stopniami swobody. Porównaj otrzymane wykresy z wykresem gęstości zmiennej losowej o rozkładzie normalnym.

⋆ **Zadanie 3.16.** W wielu tablicach statystycznych sugeruje się, że rozkład t-Studenta już od 30 stopni swobody można dobrze przybliżać rozkładem normalnym standardowym.

Niech Φ oznacza dystrybuantę rozkładu $\mathrm{N}(0,1)$, a F_d — dystrybuantę rozkładu $\mathrm{t}^{[d]}$. Dla różnych liczb stopni swobody d zbadaj wartości funkcji błędu:

$$e(d) = \sup_{x\in\mathbb{R}} |F_d(x) - \Phi(x)|,$$

którą można aproksymować za pomocą wyrażenia

$$e(d) \simeq \max_{x=-\lambda,-\lambda+\delta,\ldots,\lambda} |F_d(x) - \Phi(x)|,$$

gdzie np. $\lambda = 5$ oraz $\delta = 0{,}001$.

Zadanie 3.17. Utwórz wykresy gęstości zmiennych losowych o rozkładzie gamma z parametrami:

a) $\Gamma(1,1), \Gamma(0{,}5,1), \Gamma(2,1), \Gamma(3,1)$,

b) $\Gamma(2,1), \Gamma(2,2), \Gamma(2,3)$.

Sformułuj wnioski dotyczące wpływu obu parametrów rozkładu na kształt wykresu gęstości.

Zadanie 3.18. Utwórz wykresy gęstości zmiennych losowych o rozkładzie beta: $B(1,1)$, $B(2,2), B(5,2)$ i $B(2,5)$. Sformułuj wnioski dotyczące wpływu obu parametrów rozkładu na kształt wykresu gęstości.

Zadanie 3.19. Utwórz wykresy gęstości zmiennych losowych o rozkładzie F-Snedecora: $\mathrm{F}^{[10,5]}, \mathrm{F}^{[10,10]}, \mathrm{F}^{[10,20]}$.

Zadanie 3.20. Średnio jedna na dziesięć osób mijających pewien sklep wchodzi do tego sklepu. Niech X oznacza numer pierwszej osoby, która weszła do sklepu, podczas gdy $X-1$ osób, które wcześniej mijały ów sklep, nie weszło do środka. Oblicz prawdopodobieństwa $P(X=1), P(X=2), P(X=3), P(X=4)$ oraz $P(X>11)$.

Zadanie 3.21. W partii towaru liczącej 200 sztuk znajduje się 5 sztuk niespełniających wymagań jakościowych. Jakie jest prawdopodobieństwo, że w losowej próbie 10 sztuk pobranych z tej partii nie znajdzie się ani jedna sztuka wadliwa?

Zadanie 3.22. Czas poprawnej pracy aparatu telefonicznego ma rozkład wykładniczy o intensywności awarii 0,0001 [1/h].

a) Oblicz prawdopodobieństwo, że aparat ten nie uszkodzi się w ciągu: 1000, 10000, 30000 godzin pracy.

b) Ile co najmniej godzin powinien przepracować bezawaryjnie ten aparat z prawdopodobieństwem 0,9?

Zadanie 3.23. Z dotychczasowych obserwacji wynika, że liczba klientów przybywających w ciągu godziny do oddziału banku ma rozkład Poissona o średniej 4 [klientów/h].

a) Jaki jest rozkład prawdopodobieństwa czasu między przyjściem kolejnych klientów?

b) Jaki jest średni czas oraz odchylenie standardowe czasu pomiędzy chwilami przybycia kolejnych klientów?

c) Jeżeli w danej chwili do oddziału wszedł klient, to jakie jest prawdopodobieństwo, że kolejny klient przybędzie do oddziału w ciągu najbliższych 30 minut?

d) Jakie jest prawdopodobieństwo, że w ciągu godziny do oddziału banku nie przyjdzie ani jeden klient?

Zadanie 3.24. Wygeneruj $n = 100$ liczb z rozkładu $U([0, 10])$. Znajdź maksimum i minimum otrzymanej próbki.

Zadanie 3.25. Wygeneruj $n = 100$ liczb z rozkładu $\mathrm{N}(3, 3)$. Ile z nich jest ujemnych?

Zadanie 3.26. Wygeneruj $n = 1000$ liczb z rozkładu $\mathrm{N}(1, 2)$. Ile z nich różni się od średniej o więcej niż 2 odchylenia standardowe?

Zadanie 3.27. Za pomocą `R`-a wykonaj $n = 20$ rzutów symetryczną monetą. Ile razy wypadła reszka?

Zadanie 3.28. W urnie jest $n = 60$ kul, ponumerowanych od 1 do n. Wylosuj (bez zwracania) $m = 30$ z nich. Jaki jest największy i najmniejszy numer wylosowanej kuli? Powtórz eksperyment losując ze zwracaniem.

Zadanie 3.29. Za pomocą `R`-a wykonaj $n = 100$ rzutów symetryczną kostką do gry. Ile razy wypadła „szóstka" lub „piątka"?

Zadanie 3.30. Za pomocą `R`-a wylosuj (ze zwracaniem) $n = 1000$ kart do gry. Ile otrzymaliśmy asów?

Zadanie 3.31. Rzucamy $n = 1000$ razy dwiema symetrycznymi monetami. Wygeneruj odpowiednią próbkę za pomocą `R`-a. Ile razy otrzymaliśmy dwa orły?

Zadanie 3.32. Korzystając z generatora liczb losowych o rozkładzie jednostajnym na przedziale $[0, 1]$, wygeneruj próbkę losową z rozkładu wykładniczego z parametrem $\lambda = 5$. Narysuj histogram dla uzyskanych danych.

Zadanie 3.33. Korzystając z generatora liczb losowych o rozkładzie jednostajnym na przedziale $[0, 1]$, wygeneruj próbkę losową z rozkładu logistycznego. Narysuj histogram dla uzyskanych danych.

⋆ **Zadanie 3.34.** Posługując się metodą Monte Carlo, oblicz pole powierzchni obszaru $B = \{(x, y) \in \mathbb{R}^2 : x^2 < y < 1 - x^2\}$. Porównaj uzyskane w ten sposób wyniki z dokładnymi rezultatami otrzymanymi na drodze analitycznej.

⋆ **Zadanie 3.35.** Posługując się metodą Monte Carlo, wyznacz aproksymację liczby π.

Empiryczna weryfikacja reguły 3-sigmowej

Zadanie 3.36. Niech X oznacza zmienną losową o rozkładzie normalnym standardowym. Oblicz wartości następujących prawdopodobieństw:

a) $P(-1 < X < 1)$,

b) $P(-2 < X < 2)$,

c) $P(-3 < X < 3)$.

Wygeneruj $n = 10000$ elementową próbę $(X_1, \ldots, X_n)$ z rozkładu normalnego standardowego. Porównaj częstości wystąpienia zdarzeń: $A : X_i \in (-1, 1)$, $B : X_i \in (-2, 2)$, $C : X_i \in (-3, 3)$ z wartościami odpowiednich prawdopodobieństw wyznaczonych powyżej.

Empiryczna weryfikacja CTG

Zadanie 3.37. Wygeneruj $m = 100$ próbek $n = 200$ elementowych $(U_1, \ldots, U_n)$ z rozkładu jednostajnego na przedziale $[0, 1]$. Utwórz histogramy dla zmiennych $(Z_1, \ldots, Z_m)$, gdzie

$$Z_k = \frac{\sum_{i=1}^{k} U_i - k/2}{\sqrt{k/12}},$$

dla $k = 1, \ldots, m$. Nałóż na histogram wykres gęstości rozkładu normalnego standardowego. Sformułuj wnioski odnośnie zmiany kształtu histogramu zmiennej Z_k wraz ze wzrostem k.

Empiryczne badanie jakości aproksymacji

⋆ **Zadanie 3.38.** Niech X oznacza zmienną losową o rozkładzie dwumianowym $\mathrm{Bin}(n, p)$. Wyznacz tablice prawdopodobieństw $P(X \leqslant k)$ dla kilku wybranych wartości k. Porównaj te prawdopodobieństwa z wartościami prawdopodobieństw otrzymanymi za pomocą aproksymacji

a) rozkładem Poissona,

b) rozkładem normalnym (tw. Moivre'a-Laplace'a),

c) rozkładem normalnym z korektą ciągłości.

Porównaj również wykres dystrybuanty zmiennej losowej X z wykresami dystrybuant rozkładów użytych do aproksymacji X. Sformułuj wnioski dotyczące jakości aproksymacji, biorąc pod uwagę różne wartości parametrów n oraz p np. $n = 20, 30, 50$ oraz $p = 0{,}1, 0{,}2, 0{,}3, 0{,}5$.

4 Wskazówki i odpowiedzi

Wskazówka do zadania 3.20. X ma rozkład geometryczny.

Wskazówka do zadania 3.21. X ma rozkład hipergeometryczny.

Wskazówka do zadania 3.23. $X \sim \mathrm{Exp}(\lambda)$, gdzie $\mathbb{E}X = \frac{1}{4}$.

Wskazówka do zadania 3.33. Dystrybuanta rozkładu logistycznego ma postać: $F(x) = 1/(1 + \exp(-x))$.

Wskazówka do zadania 3.35. Pole koła o promieniu r wynosi πr^2.