Generowanie ciągu liczb pseudolosowych o rozkładzie normalnym metodą eliminacji.

Tomasz Chwiej

13 stycznia 2015

1 Wstęp

Funkcję gęstości prawdopodobieństwa dla rozkładu normalnego definiujemy następująco:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{1}$$

gdzie: μ to wartość oczekiwana, a σ jest odchyleniem standardowym.

Gęstość prawdopodobieństwa używana jest w definicji dystrybuanty:

$$F(x) = \int_{-\infty}^{x} f(y)dy = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy \tag{2}$$

która posłuży nam do wyznaczania prawdopodobieństwa. W celu łatwiejszego numerycznego wyznaczania dystrybuanty przekształcamy powyższy wzór:

$$\begin{aligned}
F(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy & (3)\\
&= 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_{x}^{\infty} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy & (4)\\
&= \left| t = \frac{y-\mu}{\sqrt{2}\sigma}, \quad dt = \frac{dy}{\sqrt{2}\sigma}, \quad x \to x' = \frac{x-\mu}{\sqrt{2}\sigma} \right| & (5)\\
&= 1 - \frac{1}{2}\frac{2}{\sqrt{\pi}} \int_{x'}^{\infty} exp\left(-t^2\right) dt & (6)\\
&= 1 - \frac{1}{2} erfc(x') = \frac{1 + erf(x')}{2} & (7)
\end{aligned}$$

gdzie: $erf(x)$ jest funkcją błędu, a $erfc(x) = 1 - erf(x)$ jest jej dopełnieniem.

Funkcja $erf(x)$ jest funkcją specjalną, której wartość można obliczyć przy użyciu procedury z Numerical Recipes: **erff(float x)**. Dla rozkładu normalnego łatwo teraz znaleźć prawdopodobieństwo wylosowania liczby z przedziału $[x_a, x_b]$, gdyż jest ono równe:

$$P(x_a < x \leqslant x_b) = F(x_b) - F(x_a) \tag{8}$$

2 Zadania do wykonania

2.1 Rozkład jednorodny

Startując od $x_0 = 10$ należy wygenerować $n = 10^4$ liczb pseudolosowych przy użyciu generatora mieszanego

$$x_{n+1} = (ax_n + \ c) \ mod \ m \tag{9}$$

o parametrach (**typu long**):

a) $a = 123$, $c = 1$, $m = 2^{15}$

b) $a = 69069$, $c = 1$, $m = 2^{32}$

Proszę w obu przypadkach sporządzić rysunek $X_{i+1} = f(X_i)$ ($X_i = x_i/(m + 1.0)$ **z warunku normalizacji do rozkładu U(0,1)**). Czy porównując oba rysunki można stwierdzić, który generator ma lepsze własności statystyczne? W sprawozdaniu proszę uzasadanić odpowiedź. W sprawozdaniu proszę także zamieścić histogram (dla $k = 12$ podprzedziałów) rozkładu gętości prawdopodobieństwa dla $n = 10^4$ liczb pseudolosowych o rozkładzie równomiernym (oba przypadki). Proszę także podać obliczone wartości μ i σ i porównać je z wartościami teoretycznymi.

2.2 Rozkład normalny

Wykorzystując generator mieszany z podpunktu (b) należy wygenerować ciąg $n = 10^4$ liczb pseudolosowych o rozkładzie normalnym z parametrami $\mu = 0.2$ i $\sigma = 0.5$ metodą eliminacji. Liczby pseudolosowe mają zawierać się w przedziale $x \in [\mu - 3\sigma, \mu + 3\sigma]$.

2.3 Testowanie generatora o rozkładzie $N(\mu, \sigma)$ - test χ^2

Zadania do wykonania:

1. Obliczyć średnią arytmetyczną uzyskanego rozkładu normalnego: $\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Obliczyć wariancję
$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{10}$$
i odchylenie standardowe.
Obliczone wartości μ_n i σ_n zapisać do pliku.
3. Podzielić przedział $[\mu-3\sigma, \mu+3\sigma]$ na $k = 12$ rozłącznych podprzedziałów o identycznej długości.
4. W każdym z podprzedziałów określić ilość liczb pseudolosowych (n_i), która do niego trafia. Wartości n_i zapisać do pliku.
5. Wyznaczyć wartość statystyki testowej
$$\chi^2 = \sum_{i=1}^{k}\frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} \tag{11}$$
gdzie: n jest całkowitą ilością liczb pseudolosowych, n_i ilość liczb w i-tym podprzedziale, p_i teoretyczne prawdopodobieństwo wylosowania liczby z i-tego podprzedziału. Aby wyznaczyć wartości p_i w każdym z podprzedziałów należy skorzystać z wzoru (8). Wartości: p_i oraz $n \cdot p_i$ dla każdego z podprzedziałów zapisać do pliku. Do obliczenia p_i proszę użyć założonych na początku wartości μ i σ.
6. Testujemy hipotezę H_0: wygenerowany rozkład jest rozkładem $N(\mu, \sigma)$ wobec H_1 że nie jest to prawdą. Korzystając z odpowiednich tabel statystycznych proszę sprawdzić czy nasza hipoteza jest prawdziwa na poziomie istotności $\alpha = 0.05$ (α jest prawdopodobieństwem pierwszego rodzaju czyli prawdopodobieństwem odrzucenia hipotezy H_0 gdy ta jest prawdziwa). W tym celu definiujemy obszar krytyczny testu:
$$K = \{\boldsymbol{X} : \chi^2(\boldsymbol{X}) > \varepsilon\} \tag{12}$$
gdzie: $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ jest ciągiem liczb pseudolosowych, $\chi^2(\boldsymbol{X})$ wartością statystyki dla danego ciągu $\boldsymbol{X}$, ε jest poziomem krytycznym danego rozkładu dla określonej liczby stopni swobody i założonego poziomu istotności. Liczbę stopni swobody określamy jako $\nu = k - r - 1$,

gdzie: k jest liczbą podprzedziałów, a $r = 2$ jest liczbą parametrów testowanego rozkładu (μ i σ). Jeśli $\chi^2 < \varepsilon$ to stwierdzamy że dla zadanego poziomu istotności hipoteza H_0 jest prawdziwa - nasz rozkład jest typu $N(\mu, \sigma)$.

7. Określić poziom ufności dla obliczonej statystyki χ^2:

$$P(\chi^2|\nu) = 1 - \tilde{\alpha} \tag{13}$$

gdzie: $\nu = k - r - 1$ jest liczbą stopni swobody, natomiast $\tilde{\alpha}$ jest poziomem istotności którego nie znamy (a chcemy go poznać), korzystając z procedury bibliotecznej:

$$P(\chi^2|\nu) = gammp\left(\frac{\nu}{2}, \frac{\chi^2}{2}\right) \tag{14}$$

Uwaga: można tu odwrócić zagdanienie tj. zadać sobie pytanie - jaka powinna być wartość χ^2 dla określonej wartości α? - i w ten sposób poszukiwać lewych granic obszarów krytycznych testu. Do poszukiwania wartości χ^2 można użyć np. metody bisekcji.

8. W sprawozdaniu proszę zamieścić histogram pokazujący wartości n_i/n dla każdego z podprzedziałów, na tym samym rysunku proszę także zamieścić przebieg funkcji gęstości prawdopodobieństwa dla rozkładu normalnego.