Metody nieparametryczne w statystyce

Tomasz Wójtowicz

Wydział Zarządzania
AGH Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie

Generatory liczb pseudolosowych

Częsty problem w statystyce:

jakie własności ma pewna funkcja $T(X_1, \dots, X_n)$, gdzie $X_1, \dots, X_n$ jest próbą losową prostą z pewnego rozkładu (znanego lub nie).

Często jest to trudne lub niemożliwe do zrobienia:

- gdy postać funkcji T jest skomplikowana,
- gdy rozkład badanej cechy odbiega od normalnego lub jest nieznany.

Praktyczne rozwiązanie:

- symulacje.

Generatory liczb pseudolosowych

Jak generować próby losowe proste o zadanym rozkładzie?

Wystarczy generować próbę losową prostą z rozkładu jednostajnego U na przedziale $[0,1]$.

W praktyce generuje się ciągi deterministyczne, które dobrze imitują próbę losową prostą o zadanym rozkładzie.

Metoda (dla rozkładu $\boldsymbol{U[0,1]}$):

- dla pewnej ustalonej funkcji $G: [0,1] \to [0,1]$
- dla pewnej ustalonej wartości początkowej $u_0 \in [0,1]$
- definiujemy:
  $u_1 = G(u_0), u_2 = G(u_1), \dots, u_i = G(u_{i-1}), \dots$ dla $i = 1,2, \dots$

Generatory liczb pseudolosowych

Generatorem liczba pseudolosowych z rozkładu jednostajnego $U[0,1]$ musi mieć następującą własność:

wartości $u_1, \dots, u_n$ imitują zachowanie realizacji próby losowej prostej $U_1, \dots, U_n$ z rozkładu jednostajnego $U[0,1]$, tzn. standardowy zestaw testów nie daje podstaw do odrzucenia hipotezy głównej, że $u_1, .. u_n$ jest realizacją próby $U_1, \dots, U_n$.

W szczególności:

- na podstawie $u_1, .., u_n$ nie powinno dać się obliczyć u_{n+1} z prawdopodobieństwem istotnie większym niż $1/2$,
- wszystkie własności prób losowych prostych powinny być (w przybliżeniu) spełnione dla prób pseudolosowych, czyli np. dla dużych n częstośc generowania liczb z przedziału $[a, b] \subset [0,1]$ powinna być w przybliżeniu równa $|b - a|$.

Generatory liczb pseudolosowych

Liniowy kongruencyjny generator liczb pseudolosowych:

$$u_n = (a \cdot u_{n-1} + b)\ mod\ m$$

Przykład:

dla $A = 2147483647$

$$u_n = \big((16807 \cdot u_{n-1})\ mod\ A\big)/A$$

$$u_0 = C/A$$

gdzie C jest dowolna liczbą naturalna z przedziału $(1, \dots, A-1)$

$(16807 \cdot u_{n-1})\ mod\ A$ - liczba naturalna z przedziału $1, \dots, A-1$

$\big((16807 \cdot u_{n-1})\ mod\ A\big)/A$ - liczba z przedziału $(0,1)$

Generatory liczb pseudolosowych

Przykład:

- dysponując próbą pseudolosową $u_1, \dots, u_n$ z rozkładu jednostajnego wygenerować podzbiór k-elementowy spośród liczb $\{1, \dots, n\}$

- wybrać indeksy $i_1, \dots, i_k$ odpowiadające k największym wartościom spośród $u_1, \dots, u_n$.

Metoda przekształcenia kwantylowego

Mając próbę losową $U_1, \ldots U_n$ z rozkładu $U[0,1]$ można łatwo uzyskać próbę losową dowolnego rozkładu F.

Dla uproszczenia załóżmy, że dystrybuanta F jest rosnąca, tzn.

$$\text{jeżeli } x < y, \text{ to } F(x) < F(y).$$

Twierdzenie

Niech $U_1, \ldots, U_n$ będzie próbą losową prostą z rozkładu $U[0,1]$. Wtedy ciąg kwantyli rzędu $U_1, \ldots, U_n$ dla rozkładu F, tzn. ciąg $q_F(U_1), \ldots, q_F(U_n)$ jest próbą losową prostą z rozkładu F.

Generatory liczb pseudolosowych

Na podstawie powyższego twierdzenia można uzyskać próbę losową prostą dowolnego rozkładu F.

W ten sam sposób, mając próbę pseudolosową $u_1, \dots, u_n$ z rozkładu $U[0,1]$ można uzyskać próbę pseudolosową z dowolnego rozkładu F.

Problemem jest tylko znajomość postaci dystrybuanty F. Nie zawsze dystrybuanta F dana jest w postaci jawnej. Czasami obliczenie transformacji kwantylowej jest niepraktyczne.

W przypadku niektórych rozkładów można dosyć łatwo obliczyć transformację kwantylową, np. w przypadku rozkładu wykładniczego:

$$F(x) = 1 - e^{-\lambda x}$$

$$X_i = -\ln(U_i)\,/\lambda$$

Generatory liczb pseudolosowych

Rozkład **wykładniczy** z parametrem $\lambda > 0$ ma funkcję gęstości postaci:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Rozkład χ^2 o n stopniach swobody ma gęstość postaci:

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{n/2-1} e^{-\frac{x}{2}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

gdzie $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt$ jest funkcją Eulera .

Metoda oparta na reprezentacji zmiennych losowych

Gdy postać dystrybuanty F uniemożliwia (lub utrudnia) skorzystanie z metody przekształcenia kwantylowego, można spróbować wyrazić generowany rozkład jako funkcję zmiennych losowych o rozkładach, które łatwo można estymować.

Przykłady:

- rozkład χ^2 o 2 stopniach swobody:
$$X_i = -2\ln(U_i)$$
- rozkład χ^2 o $2k$ stopniach swobody:
$$X_{2k} = -2\sum_{i=1}^{k}\ln(U_i)$$
- rozkład χ^2 o $2k+1$ stopniach swobody:
$$X_{2k+1} = X_{2k} + Z^2$$

Generowanie próby losowej prostej z rozkładu normalnego

Niech (X_1, X_2) będzie parą niezależnych zmiennych losowych o rozkładach $N(0,1)$. Niech $(R,\ \theta)$ będzie ich przedstawieniem we współrzędnych biegunowych, tzn.

$$R = \sqrt{X_1^2 + X_2^2} \qquad\qquad \theta = \arccos \frac{X_1}{\sqrt{X_1^2 + X_2^2}}$$

Wtedy:

- zmienna R^2 ma rozkład χ^2 o 2 stopniach swobody,
- zmienna θ ma rozkład $U[0,2\pi]$,
- zmienne R i θ są niezależne.

Generowanie próby losowej prostej z rozkładu normalnego

Procedura:

1. wygenerować dwie niezależne zmienne losowe U_1 i U_2 o rozkładzie jednostajnym $U[0,1]$,
2. zdefiniować:

$$X_1 = \sqrt{-2\ln U_1}\cos(2\pi U_2)$$

$$X_2 = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

3. tak zdefiniowane X_1 i X_2 są niezależnymi zmiennymi losowymi o rozkładzie normalnym $N(0,1)$.

Generowanie próby losowej prostej z rozkładu Poissona

Rozkład wykładniczy – czasy pomiędzy wystąpieniami pewnego zdarzenia

Rozkład Poissona – liczba wystąpień tego zdarzenia

Jeżeli $W_1, W_2, \ldots$ są niezależnymi zmiennymi losowymi o rozkładzie wykładniczym z pearametrem 1, to definiujemy

$$X = k,$$

gdzie k jest taka największą liczbą, że

$$W_1 + W_2 + \cdots + W_k \leq \lambda.$$

Gdy $W_1 > \lambda$, to przyjmujemy $X = 0$.

Generowanie próby losowej prostej z rozkładu jednostajnego w k punktach

Jeżeli U jest zmienną losową o rozkładzie $U[0,1]$, to zmienna $W = i$, gdy $U \in \left(\frac{i-1}{k}, \frac{i}{k}\right]$ dla $i = 1, \dots, k$ ma rozkład jednostajny na zbiorze $\{1, \dots, k\}$.

Dzieląc odcinek $[0,1]$ na odcinki o długościach $p_1, \dots, p_n$ takich, że:

$$0 < p_i < 1 \qquad\qquad \sum_{i=1}^{n} p_i = 1$$

można wygenerować dowolny rozkład dyskretny X, taki, że $P(X = i) = p_i$.

Generowanie mieszaniny rozkładów

Niech X będzie zmienna losową o rozkładzie F, który jest mieszanina rozkładów $F_1, \dots, F_n$, tzn.:

$$F = \sum_{i=1}^{n} w_i F_i.$$

gdzie $w_1 + \cdots + w_n = 1.$

Jeżeli potrafimy wygenerować rozkłady $F_1, \dots, F_n$, to możemy wygenerować F w następujący sposób:

1. generujemy indeks I taki, że $P(I = i) = w_i$,
2. jeżeli $I = j$, to generujemy liczbę z rozkładu F_j.

Generowanie dwuwymiarowego rozkładu normalnego

Zmienna losowa $(X_1, \dots, X_n)$ ma n-wymiarowy rozkład normalny o wektorze wartości oczekiwanych $\boldsymbol{\mu} = [\mu_1, \dots, \mu_n]^T$ i macierzy kowariancji Σ jeżeli jej gęstość ma postać:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

gdzie $\boldsymbol{x} = (x_1, \dots, x_n)$.

W przypadku $n = 2$ macierz kowariancji Σ ma postać:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & cov(X_1, X_2) \\ cov(X_1, X_2) & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2) \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

gdzie ρ jest współczynnikiem korelacji pomiędzy X_1 i X_2.

Generowanie dwuwymiarowego rozkładu normalnego

Jeżeli Z_1 i Z_2 są niezależnymi zmiennymi losowymi o rozkładach $N(0,1)$,

to zmienne losowe:

$$X_1 = \mu_1 + \sigma_1 Z_1$$

$$X_2 = \mu_2 + \sigma_2 (\rho Z_1 + \sqrt{1-\rho^2} Z_2$$

mają dwuwymiarowy rozkład normalny $N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

Generatory liczb pseudolosowych

Metoda eliminacji

Niech rozkład F będzie zadany za pomocą funkcji gęstości f.

Znamy (i potrafimy na jej podstawie generować liczby) gęstość g i stałą M

takie, że

$$\forall x \in \mathbb{R}\ f(x) \leq Mg(x)$$

Wtedy możemy generować zmienną Y z rozkładu o gęstości f w następujący sposób:

1. Generujemy niezależne zmienne $X \sim g$ i $U \sim U[0,1]$
2. Jeżeli $U \leq f(X)/Mg(X)$, to $Y = X$.
3. W przeciwnym wypadku powtarzamy krok 1.

Metoda eliminacji

Szybkość działania powyższej procedury zależy od tego jak dobre jest oszacowanie gęstości f przez Mg.

W ten sposób można generować np. zmienna losową o rozkładzie $N(0,1)$

wiemy, że $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ jest gęstością zm. losowej Z o rozkładzie $N(0,1)$

wtedy $\phi(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ dla $x \geq 0$ jest gęstością zmiennej $W = |Z|$.

Przyjmując:

$g(x) = e^{-x}$ - gęstość rozkładu wykładniczego z $\lambda = 1$, $M = \sqrt{\frac{2e}{\pi}} \approx 1{,}32$

możemy generować W, a po pomnożeniu przez losowy znak – zmienną Z.