Dystrybuanta rozkładu chi kwadrat

Zadanie nr 3

Antoni Pokusiński

Wprowadzenie

Rozkład chi kwadrat z k stopniami swobody definiuje się jako sumę kwadratów k niezależnych zmiennych losowych, które podlegają rozkładowi $N(0,1)$.
Niech $X_1,\ldots,X_k$ będą niezależnymi zmiennymi losowymi oraz $X_i \sim N(0,1)$. Jeśli Z to zmienna losowa, taka że

$$Z = \sum_{i=1}^{k}(X_i)^2$$

to wówczas:

$$Z \sim \chi^2(k)$$

Funkcja gęstości tego rozkładu to

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} \cdot x^{k/2-1}e^{-x/2},\ x \in (0,\infty)$$

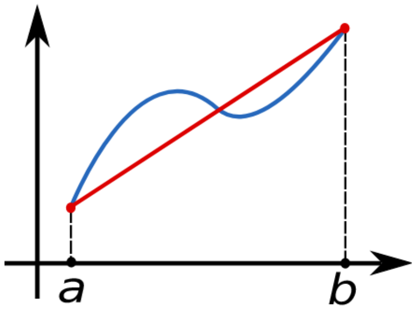
Dystrybuanta $G(t)$ rozkładu chi kwadrat jest całką z powyższej funkcji, tj. $G(t) = \int_0^t f(x)\,dx$ dla $t \geq 0$. Naszym zadaniem jest znalezienie jak najlepszego przybliżenia tej całki. W tym celu użyjemy metod całkowania numerycznego - złożonego wzoru trapezów i metody Romberga.

Spis treści

1 Złożony wzór trapezów

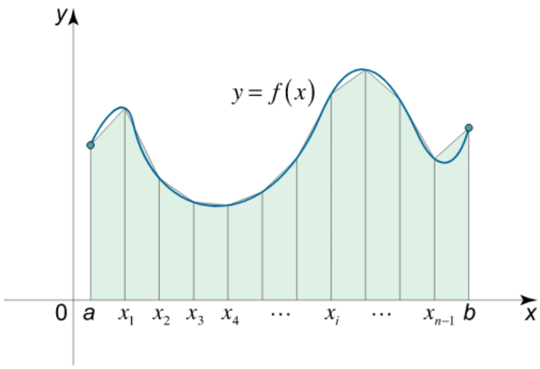
Mając daną funkcję $f(x)$ na przedziale $[a, b]$ możemy oszacować jej całkę używając **wzoru trapezów**:

$$\int_a^b f(x)\,dx \;\approx\; \big(f(a) + f(b)\big) \cdot \frac{b-a}{2}$$



Źródło: `https://en.wikipedia.org/wiki/Trapezoidal_rule`
Rys. 1: Interpretacja geometryczna wzoru trapezów

Metoda polegająca na oszacowaniu całki przez pole jednego trapezu da zawsze dobre przybliżenie jedynie gdy f jest funkcją liniową - w ogólnym przypadku jest niedokładna. Można jednak zauważyć, że gdyby podzielić przedział $[a, b]$ na wiele równych, małych podprzedziałów, na każdym z nich policzyć przybliżenie całki wzorem trapezów, a następnie zsumować wyniki z tych podprzedziałów, to wówczas dostaniemy zdecydowanie lepsze przybliżenie całki $\int_a^b f(x)\,dx$. Właśnie na tej idei opiera się **złożony wzór trapezów**.



Źródło: `https://www.math24.net/trapezoidal-rule`
Rys. 2: Interpretacja geometryczna zł. wzoru trapezów

Podzielmy przedział $[a, b]$ na n równych podprzedziałów. Niech $h_n = \frac{b-a}{n}$ oraz $t_i = a + i\,h_n$ dla $0 \le i \le n$. Zgodnie z podaną wyżej intuicją szacujemy całkę:

$$\int_a^b f(x)\,dx \approx \big(f(t_0)+f(t_1)\big)\frac{h_n}{2} + \big(f(t_1)+f(t_2)\big)\frac{h_n}{2} + \ldots + \big(f(t_{n-1})+f(n)\big)\frac{h_n}{2} =$$

$$= h_n\big(\frac{1}{2}f(t_0) + f(t_1) + \ldots + f(t_{n-1}) + \frac{1}{2}f(t_n)\big) = h_n \sum_{i=0}^{n}{}'' f(t_i)$$

gdzie $\sum''$ - suma, w której pierwszy i ostatni element mnożymy przez $\frac{1}{2}$.
Dla wygody wprowadźmy oznaczenie $T_n(f) := h_n \sum_{i=0}^{\prime\prime n} f(t_i)$.

Powyższa metoda daje już dość dobre oszacowanie całki. Pozostaje jedynie zauważyć, że większa liczba podprzedziałów, na które podzielimy $[a, b]$, oznacza lepsze przybliżenie, w szczególności $\lim\limits_{n\to\infty} T_n(f) = \int_a^b f(x)\,dx$.

2 Metoda Romberga

2.1 Opis metody

Metoda Romberga polega na konstrukcji trójkątnej tablicy (jak poniżej) z wykorzystaniem złożonego wzoru trapezu, a następnie pewnej zależności rekurencyjnej.

$$\begin{array}{cccccc}
T_{0,0} & & & & & \\
T_{0,1} & T_{1,0} & & & & \\
T_{0,2} & T_{1,1} & T_{2,0} & & & \\
T_{0,3} & T_{1,2} & T_{2,1} & T_{3,0} & & \\
\vdots & \vdots & \vdots & \vdots & \ddots & \\
T_{0,m} & T_{1,m-1} & T_{2,m-2} & T_{3,m-3} & \ldots & T_{m,0}
\end{array}$$

Pierwszą kolumnę powyższej tablicy wypełniamy korzystając ze złożonego wzoru trapezów (oznaczenia jak w poprzednim rozdziale):

$$T_{0k} = T_{2^k}(f) = h_{2^k} \sum_{i=0}^{2^k}{}'' f(t_i)$$

Kolejne elementy tablicy obliczamy korzystając z następującej zależności rekurencyjnej:

$$T_{m,k} = \frac{4^m T_{m-1,k+1} - T_{m-1,k}}{4^m - 1}$$

Po wyznaczeniu wszystkich wartości w tablicy Romberga należy zwrócić jako wynik $T_{m,0}$ - jest to najlepsze przybliżenie całki. Pozostaje tylko dodać, że im większą liczbę m ustalimy na początku, tym dokładniejszy wynik uzyskamy.

Korzystając z podanych do tego momentu narzędzi można już zaimplementować algorytm, który będzie obliczał przybliżenie całki zadanej funkcji na określonym przedziale z bardzo dobrą dokładnością. Można jednak wprowadzić pewne ulepszenia, które znacząco poprawią działanie programu.

2.2 Uwagi implementacyjne

- **Złożoność pamięciowa**. Naiwna implementacja metody Romberga wymaga użycia trójkątnej tablicy 2-wymiarowej - złożoność jest więc kwadratowa $O(m^2)$. Można jednak zauważyć, że do obliczenia kolumny $i+1$ wystarczy znać jedynie kolumnę i. Interesuje nas tylko $T_{m,0}$, więc nie musimy pamiętać wcześniejszych wartości tablicy. Jeśli będziemy obliczać elementy zaczynając "od dołu" kolumny, to okazuje się, że potrzebna jest nam jedynie 1-wymiarowa tablica rozmiaru m, czyli dostaniemy w ten sposób złożoność pamięciową $O(m)$.

- **Obliczanie wartości funkcji**. Przyjrzyjmy się pierwszej kolumnie tablicy Romberga. Zauważmy, że licząc każdy element bezpośrednio ze zł. wzoru trapezów, wielokrotnie obliczamy wartości funkcji f dla tych samych argumentów. W szczególności $f(a)$ i $f(b)$ zostają policzone aż m razy, a np. $f(\frac{b-a}{2})$ – $m-1$ razy. Można jednak zastosować następujący lemat na obliczanie kolejnego elementu pierwszej kolumny:

  Lemat. *Niech $T_n(f)$ oznacza złożony wzór trapezów dla n podprzedziałów oraz $M_n(f) = h_n \sum_{i=1}^{n} f\big(a + \frac{1}{2}(2i-1)h_n\big)$. Wówczas dla n = 1, 2, 3 ...*

  $$T_{2n}(f) = \frac{1}{2}\big(T_n(f) + M_n(f)\big)$$

  Dowód.

  $$T_{2n}(f) = h_{2n} \sum_{i=0}^{2n}{}'' f(t_i) = h_{2n} \sum_{i=0}^{n}{}'' f(t_{2i}) + h_{2n} \sum_{i=1}^{n} f(t_{2i-1}) =$$

  $$= \frac{1}{2}T_n(f) + h_{2n} \sum_{i=1}^{n} f(a+(2i-1)h_{2n}) = \frac{1}{2}T_n(f) + \frac{1}{2}h_n \sum_{i=1}^{n} f(a+\frac{1}{2}(2i-1)h_n) =$$

  $$= \frac{1}{2}T_n(f) + \frac{1}{2}M_n(f)$$

  □

3 Wyniki obliczeń

Dystrybuantę rozkładu χ^2, czyli całkę na zadanym przedziale, obliczamy więc za pomocą złożonego wzoru trapezów i metody Romberga. Należy zauważyć, że metoda Romberga daje dokładniejsze wyniki i ma lepsze zastosowanie praktyczne. (Doświadczenie zostało przeprowadzone w języku Python w arytmetyce 64-bitowej.)

Przykład dla $\chi^2(4)$ i $t = 2$:

Dokładnym wynikiem jest $\frac{e-2}{e} = 0.2642411176571153568...$

n	**Wynik zł. wzoru trapezów**	**Rząd błędu**
2^1	0.24360252522101894	10^{-2}
2^2	0.2590450401914125	10^{-2}
2^3	0.26293980164730035	10^{-3}
2^4	0.2639156448023526	10^{-3}
2^5	0.26415974044776863	10^{-4}
2^6	0.2642207727924738	10^{-5}
2^7	0.26423603140580976	10^{-5}
2^8	0.2642398460920923	10^{-5}
2^9	0.26424079976572223	10^{-6}
2^{10}	0.2642410381842585	10^{-7}
2^{11}	0.26424109778890065	10^{-7}
	0.2642411176571153568	0

Powyższa tabela przedstawia wyniki obliczeń dla złożonego wzoru trapezów. Jak widać, błąd zbiega do 0, jednak jest to dość wolna zbieżność. Zdecydowanie lepiej sprawdza się metoda Romberga:

n	**Wynik T_{n0} metody Romberga**	**Rząd błędu**
2^1	0.2634901267661182	10^{-3}
2^2	0.2642393730759054	10^{-5}
2^3	0.26424111672947576	10^{-9}
2^4	0.26424111765699954	10^{-11}
2^5	0.2642411176571154	10^{-16}
2^6	0.26424111765711533	10^{-16}
2^7	0.2642411176571154	10^{-16}
2^8	0.2642411176571152	10^{-16}
	0.2642411176571153568	0

W tym przypadku błąd zbiega do 0 zdecydowanie szybciej - jedyne co zaczyna nas ograniczać to dokładność reprezentacji liczb w arytmetyce zmiennopozycyjnek. Widzimy więc, że metoda Romberga jest efektywnym narzędziem do aprosymowania całek oznaczonych. W szczególności może mieć zastosowanie w teorii prawdopodobieństwa, gdzie często należy z dużą dokładnością przybliżać dystrybuanty zmiennych losowych.