Instytut Informatyki Uniwersytetu Wrocławskiego

Cezary Świtała

Numeryczne wyznaczanie wartości dystrybuanty rozkładu Fishera

Rachunek prawdopodobieństwa i statystyka 2020/2021

Wrocław, 14 kwietnia 2021

Spis treści

1. Wstęp

1.1. Cel dokumentu

Tematem niniejszego dokumentu będzie numeryczne wyznaczanie wartości dystrybuanty rozkładu Fishera dla zadanego $t > 0$. Składać się na niego będzie opis użytych metod całkowania numerycznego oraz zastosowania ich w obliczaniu interesującej nas dystrybuanty. Razem z kodem programu stanowi on również rozwiązanie zadania drugiego z pierwszej części egzaminu z przedmiotu Rachunek prawdopodobieństwa i statystyka.

1.2. Rozkład Fishera

Gęstość rozkładu Fishera z parametrami $m, n \in \mathbb{N}$ ma gęstość $f(x)$ zdefiniowaną w następujący sposób:

$$f(x) = \sqrt{\frac{(mx)^m \cdot n^n}{(mx+n)^{m+n}}} / (x \cdot B(m/2, n/2)), \quad x \in (0, \infty),$$

gdzie B to funkcja beta, którą definiuje się

$$B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt.$$

Naszym celem jest policzenie dystrybuanty w zależności od zadanego $t > 0$, czyli wartości całki

$$G(t) = \int_0^t f(x) dx.$$

Zrobimy to wykorzystując złożony wzór trapezów oraz metodę Romberga.

2. Metody całkowania numerycznego

2.1. Złożony wzór trapezów

Jest to przykład kwadratury złożonej, których ideą jest podział przedziału całkowania na części i zastosowanie na nich kwadratur prostych oraz zsumowaniu rezultatów.

$$\int_a^b f(x) = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} f(x),$$

gdzie $t_0, t_1, ..., t_n$ to punkty podziału($t_0 = a, t_n = b$), a całkę na przedziale (t_k, t_{k+1}) przybliżymy wybraną kwadraturą prostą.

W przypadku złożonego wzoru trapezów będzie nią wzór trapezów, czyli całkę będziemy przybliżać przez trapez o wysokości równej długości przedziału i podstawach o długościach $f(k)$ i $f(k+1)$. Czyli

$$\int_{t_k}^{t_{k+1}} f(x) \approx \frac{t_{k+1} - t_k}{2n}(f(t_k) + f(t_{k+1})).$$

W złożonym wzorze trapezów wykorzystuje się punkty równoodległe, dlatego jeśli liczmy całkę na przedziale (a, b) to odległość między początkiem, a końcem jednej części przedziału możemy oznaczyć $h = \frac{a+b}{n}$. Wtedy

$$\int_{t_k}^{t_{k+1}} f(x) \approx \frac{h}{2}(f(t_k) + f(t_{k+1})).$$

Czyli złożony wzór trapezów dla n punktów będzie miał postać

$$T_n(f) = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} f(x) = \sum_{k=0}^{n-1} \frac{h}{2}(f(t_k) + f(t_{k+1})) = h \sum_{k=0}^{n}{}'' f(t_k)$$

Można łatwo pokazać, że ciąg złożonych wzorów trapezów $\{T_n(f)\}$ jest zbieżny do całki przy $n \to \infty$.

Dowód. Pokażemy, że dla dowolnej funkcji f ciągłej na przedziale (a, b) ciąg złożonych wzorów trapezów $\{T_n(f)\}$ zbiega do $\int_a^b f(x)dx$ przy $n \to \infty$.

$$T_n(f) = \sum_{k=0}^{n-1} \frac{h}{2}(f(t_k) + f(t_{k+1})) = \sum_{k=1}^{n} \frac{h}{2}(f(t_k - 1) + f(t_k))$$

Niech m_i będzie argumentem, dla którego $f(x)$ przyjmuje najmniejszą wartość na przedziale (t_{k-1}, t_k), a M_i takim, dla którego przyjmuje w tym przedziale wartość największą. Możemy wtedy zapisać nierówność

$$\sum_{k=1}^{n} \frac{h}{2}(f(m_k) + f(m_k)) \leqslant \sum_{k=1}^{n} \frac{h}{2}(f(t_k - 1) + f(t_k)) \leqslant \sum_{k=1}^{n} \frac{h}{2}(f(M_k) + f(M_k))$$

$$\sum_{k=1}^{n} h \cdot f(m_k) \leqslant T_n(f) \leqslant \sum_{k=1}^{n} h \cdot f(M_k).$$

Zauważmy, że $\sum_{k=1}^{n} h \cdot f(m_k)$ i $\sum_{k=1}^{n} h \cdot f(M_k)$ to sumy Riemanna, czyli przy $n \to \infty$ zbiegają do $\int_a^b f(x)dx$. Czyli z twierdzenia o trzech ciągach $\{T_n(f)\}$ zbiega do $\int_a^b f(x)dx$, przy $n \to \infty$. Co kończy dowód.

2.2. Metoda Romberga

Jest to metoda przyspieszająca zbieżność ciągu złożonych wzorów trapezów. Niech $T_{0,k} = T_{2^k}$, czyli

$$T_{0,k} = h_k \sum_{i=0}^{2^k}{}'' f(x_i^{(k)}),$$

gdzie $h_k = \frac{b-a}{2^k}$ oraz $x_i^{(k)} = a + ih_k$, $(i = 0, 1, ..., 2^k)$. Inaczej mówiąc jest to złożony wzór trapezów dla $2^k + 1$ równoodległych punktów. Definiujemy teraz zależność

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

wartości te możemy obliczać w tablicy, gdzie m oznacza kolumnę, a k oznacza wiersz.

$T_{0,0}$					
$T_{0,1}$	$T_{1,0}$				
$T_{0,2}$	$T_{1,1}$	$T_{2,0}$			
$T_{0,3}$	$T_{1,2}$	$T_{2,1}$	$T_{3,0}$		
$T_{0,4}$	$T_{1,3}$	$T_{2,2}$	$T_{3,1}$	$T_{4,0}$	
$\vdots$					$\ddots$

Tablica 1. Tablica Romberga

Można pokazać, że ciągi tworzone przez kolejne wartości w kolumnach są zbieżne do całki.

Dowód. Pokażemy że dla dowolnej funkcji f ciągłej na przedziale (a, b) ciąg $\{T_{m,k}\}$ jest zbieżny do $\int_a^b f(x)dx$ przy $k \to \infty$ dla każdego $m \in \mathbb{N}$. Dowód będzie przez indukcję.
Podstawa: $m = 0$.

$$\lim_{k\to\infty} T_{0,k} = \lim_{k\to\infty} T_{2^k}$$

Udowodniliśmy już że ciąg złożonych wzorów trapezów jest zbieżny do $\int_a^b f(x)dx$, zatem podstawa indukcji zachodzi.
Krok. Załóżmy prawdziwość tezy dla dowolnego m, pokażemy, że indukuje to prawdziwość dla $m + 1$.

$$\lim_{k\to\infty} T_{m+1,k} = \lim_{k\to\infty} \left(\frac{4^m T_{m-1,k+1} - T_{m-1,k}}{4^m - 1} \right) \overset{z.zał.}{=} \frac{4^m \int_a^b f(x)dx - \int_a^b f(x)dx}{4^m - 1} = \int_a^b f(x)dx$$

Co kończy dowód.

2.2.1. Przyspieszanie obliczania tablicy Romberga

Wyznaczanie wartości w kolejnych kolumnach, znając już wartości pierwszej jest proste, zajmuje stałą liczbę operacji i nie wymaga dodatkowej pamięci, bo możemy je zapisać w miejscu poprzednich.

Skupimy się zatem na optymalizacji wyznaczania pierwszej kolumny, której obliczenie może być kosztowne, gdyż na przykład przy wyznaczaniu drugiego elementu potrzebujemy znać wartość funkcji podcałkowej na początku, na końcu i na środku przedziału całkowania, a dwie pierwsze z tych wartości były już obliczane przy okazji wyznaczania pierwszego elementu, podobnie w każdym następnym elemencie obok wartości w kilku nowych punktach, będziemy używać wszystkich z elementu poprzedniego.

Potrzebujemy zatem sposobu na wyliczanie wartości tylko w nowych punktach i tutaj pomocny okazuje się fakt, że ciąg złożonych wzorów trapezów spełnia poniższy związek

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

gdzie

$$M_n(f) = h_n \sum_{k=1}^{n} f\left(a + \frac{1}{2}(2k-1)h_n\right),\quad h_n = \frac{b-a}{n}.$$

Dowód. Rozpiszmy T_n i M_n, gdzie T_n zdefiniujemy

$$T_n(f) = h_n \sum_{k=0}^{n}{}'' f(x_k),\quad x_k = a + kh_n$$

Rozwijamy sumy:

$$T_n(f) = h_n \left(\frac{1}{2}f(a) + f(a+h_n), + f(a+2h_n) + ... + \frac{1}{2}f(a+nh_n)\right)$$

$$M_n(f) = h_n \left(f\left(a+\frac{1}{2}h_n\right) + f\left(a+\frac{3}{2}h_n\right) + ... + f\left(a+\frac{2n-1}{2}h_n\right)\right).$$

Widać teraz, że M_n dokłada wartości funkcji pomiędzy wcześniejsze zatem możemy napisać

$$T_n(f) + M_n(f) = h_n \sum_{k=0}^{2n}{}'' f\left(a+\frac{k}{2}h_n\right) = h_n \sum_{k=0}^{2n}{}'' f\left(a+kh_{2n}\right).$$

Mając to, możemy teraz łatwo pokazać związek, którego dowodzimy

$$\frac{1}{2}(T_n(f) + M_n(f)) = \frac{1}{2}h_n \sum_{k=0}^{2n}{}'' f\left(a+kh_{2n}\right) = h_{2n} \sum_{k=0}^{2n}{}'' f\left(a+kh_{2n}\right) = T_{2n}.$$

Co kończy dowód.

Mamy teraz sposób, żeby wyznaczyć kolejny element pierwszej kolumny, na podstawie wartości poprzedniego oraz wartości funkcji w nowych węzłach, co znacznie przyspiesza nasze obliczenia.

3. Wyznaczanie dystrybuanty rozkładu Fishera

Przypomninam, że interesuje nas wartość

$$G(t) = \int_0^t f(x)dx = \int_0^t \sqrt{\frac{(mx)^m \cdot n^n}{(mx+n)^{m+n}}}/(x \cdot B(m/2, n/2))dx.$$

Wyznaczając jej wartość metodą Romberga, będziemy bardzo często wyliczać wartość funkcji podcałkowej w jakimś punkcie, więc warto uprościć jej obliczanie. Możemy to zrobić, zauważając że wartośc $B(m/2, n/2)$ oraz $m^m \cdot n^m$ jest niezależna od zmiennej x, czyli wyrażenia te mogą zostać wyciągnięte przed całkę. Możemy również bez problemu wciągnąć $\frac{1}{x}$ pod pierwiastek, bo $x > 0$.

$$\int_0^t \sqrt{\frac{(mx)^m \cdot n^n}{(mx+n)^{m+n}}}/(x \cdot B(m/2, n/2))dx = \frac{\sqrt{m^m \cdot n^n}}{B(m/2, n/2)} \int_0^t \sqrt{\frac{x^{m-2}}{(mx+n)^{m+n}}}dx$$

Problemem może być fakt, że dla $m = 1$ nasza funkcja nie ma wartości w zerze, ale możemy to obejść ustalając ją na 0, co nie powinno mieć dużego znaczenia przy gęstym podziale na trapezy. Następnym krokiem będzie przyjrzenie się funkcji B, gdyż parametry m, n są naturalne, zatem potencjalnie można wyznaczyć jej wartość dokładniej niż za pomocą metody Romberga.

3.1. Wyznaczanie wartości funkcji beta

Uwaga: W poniższych rozważaniach będziemy traktować twierdzenia udowodnione na ćwiczeniach z Rachunku prawdopodobieństwa i statystyki jako fakty z podaniem numeru listy X oraz zadania Y w postaci ***L.X.Y****.*

Wiemy (**L.3.9**), że dla dowolnych $p, q \in \mathbb{R}_+$ możemy zapisać równość

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{1}$$

Problemem jest teraz wyznaczenie wartości funkcji Γ. Wiemy, że argumentami funkcji B będą parametry rozkładu podzielone przez dwa, które z kolei są dodatnimi liczbami naturalnymi. Zatem w przypadku kiedy parametry były parzyste, to zadanie mamy ułatwione gdyż (**L.1.3**)

$$\Gamma(n) = (n-1)!, \quad n \in \mathbb{N}$$

Jeśli jeden z parametrów jest nieparzysty, sprawa się komplikuje, ale dalej jesteśmy w stanie wyznaczyć wygodny wzór. Wykorzystamy w tym celu związek

$$\Gamma(z)\Gamma\left(z+\frac{1}{2}\right) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$$

Dowód. Z definicji funkcji B oraz (1), wiemy że

$$\frac{\Gamma(z)\Gamma(z)}{\Gamma(2z)} = \int_0^1 t^{z-1}(1-u)^{z-1}dt$$

Stosujemy podstawienie $t = \frac{1+x}{2}$, $dt = \frac{1}{2}dx$

$$\int_0^1 t^{z-1}(1-u)^{z-1}dt = \frac{1}{2}\int_{-1}^1 \left(\frac{1+x}{2}\right)^{z-1}\left(\frac{1-x}{2}\right)^{z-1}dx = \frac{1}{2^{2z-1}}\int_{-1}^1 (1-x^2)^{z-1}dx.$$

Z symetryczności funkcji podcałkowej

$$\frac{1}{2^{2z-1}}\int_{-1}^1 (1-x^2)^{z-1}dx = \frac{1}{2^{2z-1}}2\int_0^1 (1-x^2)^{z-1}dx.$$

Możemy teraz zapisać

$$\frac{\Gamma(z)\Gamma(z)}{\Gamma(2z)} = \frac{1}{2^{2z-1}}2\int_0^1 (1-x^2)^{z-1}dx,$$

czyli

$$2^{2z-1}\Gamma(z)\Gamma(z) = \Gamma(2z)2\int_0^1 (1-x^2)^{z-1}dx. \tag{2}$$

Pozostaje wyznaczyć występującą we wzorze podwojoną całkę. Zastosujmy podstawienie $t = x^2$, $dt = 2x\,dx$ do definicji funkcji B.

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt = 2\int_0^1 x^{2p-1}(1-x^2)^{q-1}dx.$$

Jeśli za parametry przyjmiemy $\frac{1}{2}$ i z, to dostaniemy

$$B\left(\frac{1}{2}, z\right) = 2\int_0^1 (1-x^2)^{z-1}dx. \tag{3}$$

Korzystając z wzorów (2) i (3), możemy napisać równość

$$2^{2z-1}\Gamma(z)\Gamma(z) = \Gamma(2z)B\left(\frac{1}{2}, z\right).$$

Teraz ponownie korzystając z wzoru (1) oraz z faktu, że $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ (**L.5.4**)

$$2^{2z-1}\Gamma(z)\Gamma(z) = \Gamma(2z)\frac{\Gamma(\frac{1}{2})\Gamma(z)}{\Gamma\left(z+\frac{1}{2}\right)} = \Gamma(2z)\sqrt{\pi}\frac{\Gamma(z)}{\Gamma\left(z+\frac{1}{2}\right)}$$

Zatem

$$\Gamma(z)\Gamma\left(z+\frac{1}{2}\right) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$$

Co kończy dowód.

Wyliczanie wartości $\Gamma\left(n/2\right)$ będzie zatem polegało na rozwiązaniu równania

$$\Gamma(n/2) = \begin{cases} \sqrt{\pi}, & n = 1 \\ (n/2-1)!, & 2|n \\ 2^{1-2k}\sqrt{\pi}\frac{\Gamma(2k)}{\Gamma(k)} & dla\ n = 2k+1,\ k \in \mathbb{Z}_+ \end{cases}$$

4. Wyniki obliczeń

Użyta autorska implementacja metody numerycznego wyznaczania całki za pomocą złożonego wzoru trapezów i metody Romberga, zgodna z tą opisaną wyżej w ninejszym dokumencie zostanie załączona w pliku *romberg.py*.

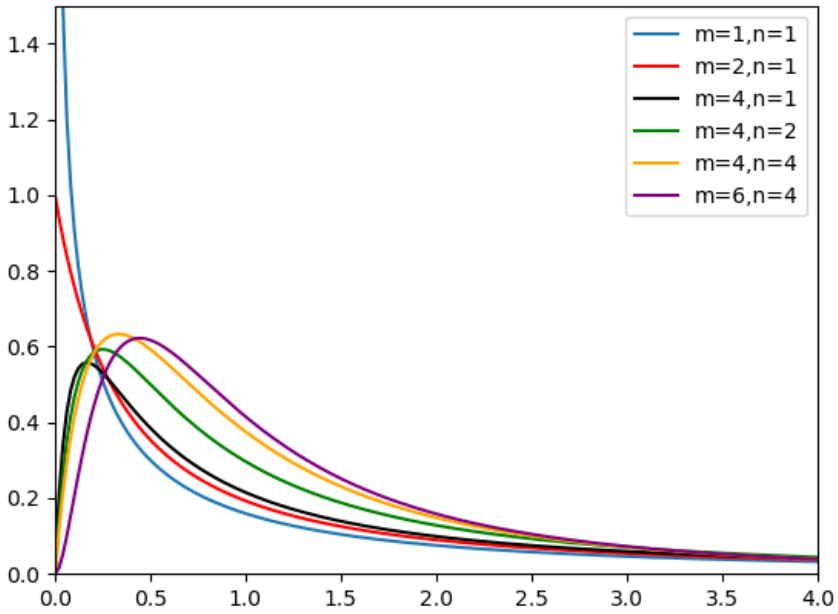
Podobnie funkcja wyznaczająca wartości funkcji gęstości rozkładu Fishera oraz funkcje pomocnicze z nią związane zostaną załączone w pliku *fisher.py*.

Cały kod za pomocą którego generowane były poniższe wykresy załączony zostanie w pliku *charts.py*, jednak do uruchomienia wymagane są zainstalowane biblioteki *numpy* i *matplotlib*.

Kod w powyższych plikach napisany został w języku Python w wersji trzeciej i **nie jest kompatybilny ze wcześniejszymi wersjami**.

4.1. Wykresy funkcji gęstości z wybranymi parametrami

Możemy teraz dla przykładu wygenerować wykresy funkcji gęstości dla wybranych parametrów m, n i argumentów z przedziału $x \in (0, 4]$.

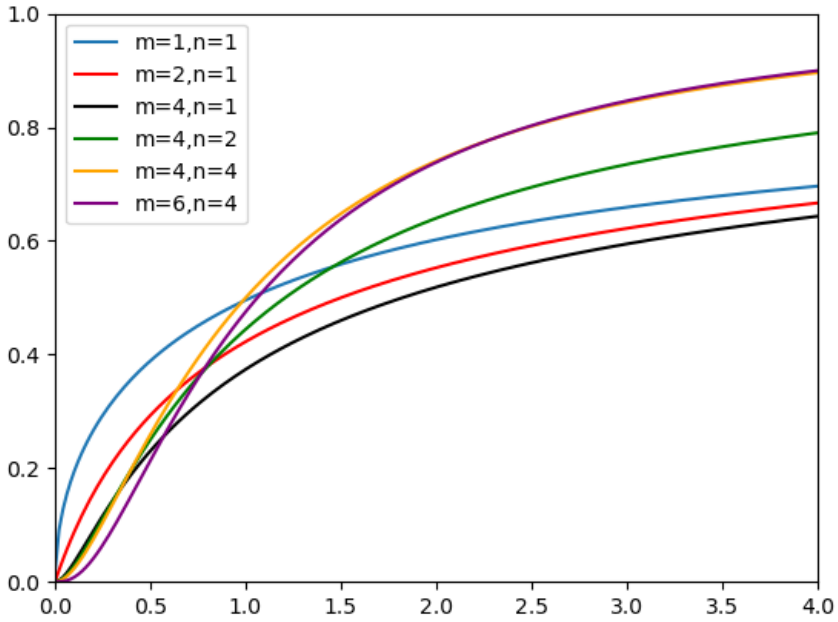


Rysunek 1. Wykresy funkcji gęstości rozkładu Fishera dla wybranych parametrów

Jesteśmy w stanie wyznaczać już wartości funkcji gęstości, więc korzystając z metody Romberga jesteśmy teraz również w stanie wyznaczać wartości dystrybuanty.

4.2. Wykresy dystrybuanty dla wybranych parametrów

Dla funkcji przedstawionych na 1 wyliczymy i przedstawimy na wykresie wartości dystrybuanty, czyli interesującej nas funkcji $G(t)$, dla wartości $t \in (0, 4]$.



Rysunek 2. Wykresy funkcji dystrybuanty dla wybranych parametrów

Moduł z pliku *fisher.py* udostępnia funkcję *get_cumulative_fisher _distr(m,n,romberg_size)*, która zwraca funkcję $G(t)$ dla zadanych parametrów $m, n \in \mathbb{N}$ oraz parametru opcjonalnego *romberg_size* , który definiuje ile wierszy romberga będziemy używać przy liczeniu całki i której można następnie użyć do obliczenia wybranego $t > 0$.