Zadanie 1.1: Obliczanie dystrybuanty rozkładu F metodą trapezów

Dawid Pawliczek
347081

6 kwietnia 2025

1 Wstęp

Rozkład F odgrywa kluczową rolę w statystyce, szczególnie w analizie wariancji (ANOVA) oraz w innych F-testach. F-testy to procedury statystyczne służące do porównywania wariancji między dwoma lub więcej grupami, a także do testowania dopasowania modeli. Nazwa rozkładu F pochodzi od nazwiska Ronalda Fishera, wybitnego statystyka, który znacząco przyczynił się do rozwoju metod analizy wariancji.

2 Teoria rozkładu F

2.1 Definicja i właściwości rozkładu F

- Rozkład jest asymetryczny i jego kształt zależy od dwóch parametrów: liczby stopni swobody dla licznika ($df1$) oraz dla mianownika ($df2$).
- Wartość oczekiwana istnieje dla $df2 > 2$ i wynosi $\frac{df2}{df2-2}$.
- Stosowany jest do testowania hipotez, na przykład w analizie wariancji, gdzie sprawdza się, czy różnice między średnimi grup wynikają ze zmienności wewnątrz grup, czy też są statystycznie istotne.

2.2 Gęstość rozkładu F

Funkcja gęstości rozkładu F opisuje, jak prawdopodobieństwo jest rozłożone na różnych wartościach zmiennej losowej. Wzór na funkcję gęstości rozkładu F dla $t \geq 0$ ma postać:

$$f(t) = c\, t^{\frac{df1}{2}-1} \left(1 + \frac{df1}{df2} t\right)^{-\frac{df1+df2}{2}},$$

gdzie c jest stałą normalizacyjną. Stała ta jest wyznaczana przy użyciu funkcji beta, która jest powiązana z funkcjami gamma, i zapewnia, że całkowite prawdopodobieństwo wynosi 1 (czyli całka z gęstości po całej dziedzinie jest równa 1). Funkcje gamma i beta są powszechnie stosowane w statystyce do opisu różnych rozkładów ciągłych i pomagają prawidłowo skalibrować funkcję gęstości rozkładu F.

3 Metoda numeryczna: Złożona metoda trapezów

3.1 Podstawy metody trapezów

Metoda trapezów to technika numeryczna służąca do przybliżonego obliczania wartości całek oznaczonych. Polega ona na podziale przedziału całkowania na mniejsze podprzedziały, a następnie przybliżeniu obszaru pod krzywą funkcji przez trapezy.

Dla funkcji $f(t)$ zdefiniowanej na przedziale $[a, b]$, całka

$$\int_a^b f(t)\, dt$$

jest przybliżana przez wzór:

$$\int_a^b f(t)\, dt \approx \frac{h}{2} \left[f(a) + 2 \sum_{i=1}^{n-1} f(a + i\, h) + f(b) \right],$$

gdzie $h = \frac{b-a}{n}$ jest szerokością każdego podprzedziału, a n to liczba podziałów.

4 Implementacja w OCTAVE

4.1 Opis struktury kodu

- **Definicja funkcji:** Na początku definiujemy funkcję `z1` z trzema argumentami:
  - `x` – punkt, w którym obliczana jest dystrybuanta,
  - `df1` i `df2` – liczby stopni swobody dla licznika i mianownika rozkładu F.
- **Liczba przedziałów (N):**

  ```
  N = 1 000 000;
  ```

  Ustawiamy dużą liczbę podziałów przedziału $[0, x]$, aby zwiększyć dokładność przybliżenia całki metodą trapezów.
- **Obliczanie stałej normalizacyjnej:**

  ```
  B = gamma(df1/2) * gamma(df2/2) / gamma((df1+df2) / 2);
  c = (df1/df2)^(df1/2) / B;
  ```

  Korzystamy z funkcji `gamma` do obliczenia funkcji beta, która służy do wyznaczenia stałej c.
- **Definicja funkcji gęstości:**

  ```
  f = @(t) c * t.^(df1/2 - 1) .* (1 + (df1/df2)*t).^(-(df1+df2)/2);
  ```

  Definiujemy funkcję `f`, która odpowiada gęstości rozkładu F.
- **Tworzenie siatki punktów:**

  ```
  t = linspace(0, x, N+1);
  h = x / N;
  ```

Funkcja `linspace` tworzy równomiernie rozmieszczone punkty w przedziale od 0 do x. Zmienna `h` określa szerokość pojedynczego przedziału.

- **Obliczenie przybliżonej wartości całki:**

```
integral_val = h * ( sum(f(t)) - 0.5*(f(t(1))) + f(t(end)) );
```

W tej linii używamy złożonej metody trapezów do przybliżenia wartości całki z funkcji gęstości `f` na przedziale $[0, x]$. Sumujemy wartości funkcji w punktach siatki, odejmując połowę wartości funkcji w pierwszym i ostatnim punkcie.

5 Wyniki i dyskusja

5.1 Porównanie wyników

W celu weryfikacji poprawności implementacji metody trapezów, porównano wyniki uzyskane z własnej implementacji z wartościami dystrybuanty rozkładu F obliczonymi za pomocą funkcji `fcdf` z pakietu `statistics`.

Dla przykładowych parametrów:

- **Przykład 1:**
  - Parametry: $x = 0.5$, $df1 = 5$, $df2 = 3$
  - Wynik funkcji `fcdf`: $y = 0.232623918000079$
  - Wynik własnej implementacji (dla $N = 10^6$): $y \approx 0.232624346689244$
- **Przykład 2:**
  - Parametry: $x = 10$, $df1 = 4$, $df2 = 17$
  - Wynik funkcji `fcdf`: $y = 0.999761880364224$
  - Wynik własnej implementacji: $y \approx 0.999761882366273$
- **Przykład 3:**
  - Parametry: $x = 1$, $df1 = 2$, $df2 = 3$
  - Wynik funkcji `fcdf`: $y = 0.535241998455110$

– Wynik własnej implementacji: $y \approx 0.535242416737439$

Różnice między wynikami dla wszystkich przypadków są bardzo niewielkie, rzędu 10^{-7}, co potwierdza poprawność zastosowanej metody numerycznej. Zwiększenie liczby podziałów N jeszcze bardziej zbliży wyniki własnej implementacji do wartości otrzymanych z funkcji `fcdf`.

6 Bibliografia

- `https://homepage.divms.uiowa.edu/~mbognar/applets/f.html`
- `https://en.wikipedia.org/wiki/F-distribution`
- `https://adamdjellouli.com/articles/statistics_notes/probability_distributions/continuous_distributions/f_distribution`
- `https://medium.com/data-science/f-distribution-simply-explained-45d0e6768a4`