

# Rozkład F dostępne funkcje

Krystian Baran 145000

6 kwietnia 2021

## Spis treści

<b>1</b>	<b>Definicja i podstawowe wzory (bez dowodu)</b>	<b>3</b>
<b>2</b>	<b>R</b>	<b>4</b>
<b>3</b>	<b>MS Excel</b>	<b>5</b>
<b>4</b>	<b>GNU Octave</b>	<b>6</b>
<b>5</b>	<b>Matlab</b>	<b>7</b>
5.1	fcd f . . . . .	7
5.2	fpdf . . . . .	7
5.3	finv . . . . .	7
5.4	fstat . . . . .	7
5.5	frnd . . . . .	8
<b>6</b>	<b>TI-82 Stats</b>	<b>9</b>
6.1	Fpdf( . . . . .	9
6.2	Fcdf( . . . . .	9
<b>7</b>	<b>Przykład</b>	<b>10</b>
<b>8</b>	<b>Bibliografia</b>	<b>12</b>

# 1 Definicja i podstawowe wzory (bez dowodu)

Rozkład F jest rozkładem zmiennej losowej typu ciągłego często stosowane w analizie wariancji. Jego funkcja gęstości wygląda następująco:

$$f(x|d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2} x\right)^{-\frac{d_1+d_2}{2}}$$

Gdzie  $B(a, b)$  to funkcja beta z parametrami  $a$  i  $b$ . Często spotykana jest także wersja bez funkcji Beta korzystając z podobieństwa funkcji Beta z funkcją Gamma, to znaczy:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Parametry  $d_1$  i  $d_2$  są to stopnie swobody, często nazywane stopnie swobody licznika i mianownika odpowiednio, i muszą być większe od 0. Natomiast zmienna  $x$  jest wartością większą od 0 dla  $d_1 = 1$  lub większą równą 0 dla innych przypadków  $d_1$ .

Obliczenie dystrybuanty jest trudne bez znania stopni swobody dla tego korzysta się z uregulowanej niepełnej funkcji Beta definiowanej następującym wzorem:

$$I_x(a, b) = \frac{B(x; a, b)}{B(a, b)}$$

Gdzie  $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1} dt$ . Dystrybuantę rozkładu F zapisuje się zatem w następujący sposób:

$$F_X(x) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$$

Wartość oczekiwana wynosi:

$$\mathbb{E}X = \frac{d_2}{d_1 - 2}$$

dla wartości  $d_2 > 2$ .

Wartość modalna wynosi

$$\frac{d_1 - 2}{d_1} \frac{d_2}{d_2 + 2}$$

dla wartości  $d_1 > 2$ .

Wariancja wynosi

$$\mathbb{D}^2(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)}$$

dla wartości  $d_2 > 4$ .

## 2 R

W języku programowania R dostępne są następujące funkcje dla rozkładu F:

- `df(x, df1, df2, ncp, log = FALSE)`
- `pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)`
- `qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)`
- `rf(n, df1, df2, ncp)`

Gdzie:

- **x**, **q** - wektory kwantyli
- **p** - wektor prawdopodobieństw
- **df1**, **df2** - stopnie swobody (można skorzystać z  $\infty$ )
- **ncp** - parametr nie środkowości (gdy nie podany jest to centralny rozkład F)
- **log**, **log.p** - gdy TRUE wartości brane jako  $\log(p)$
- **lower.tail** - jeżeli TRUE liczone  $P(X \leq x)$ , gdy FALSE  $P(X > x)$

Funkcja *df* liczy wartość gęstości dla podanego *x*, funkcja *pf* liczy wartość dystrybucyjną dla podanego *q*, funkcja *qf* liczy kwantyl dla podanego prawdopodobieństwa i funkcja *rf* generuje próbę *n* liczb losowych według rozkładu z danymi parametrami *df1* i *df2*.

### 3 MS Excel

W programie MS Excel dostępna jest funkcja **F.DIST**( $x$ ,  $\text{deg\_freedom1}$ ,  $\text{deg\_freedom2}$ ,  $\text{cumulative}$ ) która oblicza wartość funkcja gęstości lub dystrybuanty w danym  $x$ . Parametry podane są następujące:

- **$x$**  - wartość w którym obliczyć prawdopodobieństwo lub dystrybuantę
- **deg\_freedom1** - stopień swobody nr 1, ten co występuje w liczniku
- **deg\_freedom2** - stopień swobody nr 2, ten co występuje w mianowniku
- **cumulative** - wartość logiczna, FALSE dla gęstości, TRUE dla dystrybuanty

Parametry  $x$  musi być większy od zera a  $\text{deg\_freedom1}$ ,  $\text{deg\_freedom2}$  muszą być większe od 1, inaczej zwracany jest błąd **#NUM!**.

Parametry  $\text{deg\_freedom1}$  i  $\text{deg\_freedom2}$  powinny być także całkowite, w przeciwnym wypadku są obcinane do liczby całkowitej.

Gdy dowolny parametr podany nie jest liczbą funkcja zwraca błąd **#VALUE!**.

Funkcja ta dostępna jest dla następujących wersji MS Excel:

*Excel for Microsoft 365, Excel for Microsoft 365 for Mac, Excel for the web, Excel 2019, Excel 2016, Excel 2019 for Mac, Excel 2013, Excel 2010, Excel 2016 for Mac, Excel for Mac 2011, Excel Web App, Excel Starter 2010*

## 4 GNU Octave

W GNU Octave dostępne są następujące funkcje dla rozkładu F:

- **fpdf**(**x**, **m**, **n**) - gęstość rozkładu F
- **fcdf**(**x**, **m**, **n**) - dystrybuanta rozkładu F
- **finv**(**x**, **m**, **n**) - odwrotność dystrybuanty rozkładu F, czyli funkcja kwantylowa

Gdzie:

- **x** - wartość dla której jest obliczana funkcja
- **m** - stopnie swobody w liczniku
- **n** - stopnie swobody w mianowniku

Funkcje te dostępne są w GNU Octave pod warunkiem że zainstalowany i załadowany został pakiet *statistics* który można łatwo zainstalować.

## 5 Matlab

W Matlab dostępne są następujące funkcje dla rozkładu F:

- **fpdf** - gęstość rozkładu F
- **fcdf** - dystrybucja rozkładu F
- **finv** - funkcja kwantylowa rozkładu F, czyli odwrotność dystrybucji
- **fstat** - funkcja obliczająca wartość oczekiwaną i wariancję rozkładu F
- **frnd** - funkcja generująca losowe liczby według rozkładu F

Do każdej funkcji należą co najmniej parametry **V1** i **V2** które są stopnie swobody odpowiednio te w liczniku i te w mianowniku.

### 5.1 fcdf

$fcdf(x, v1, v2, 'upper')$

Dla każdego elementu tablicy  $x$  zwracana jest wartość gęstości w tym punkcie. Parametr *'upper'* jest parametrem nie obowiązkowym, służy do obliczenia wartości gęstości algorytmem bardziej dokładnym dla skrajnych wartości gęstości.

### 5.2 fpdf

$fpdf(X, V1, V2)$

Dla każdego elementu tablicy  $x$  zwracana jest wartość dystrybucji.  $V1$  i  $V2$  mogą także być tablicami pod warunkiem że mają ten sam wymiar.

### 5.3 finv

$finv(P, V1, V2)$

Dla każdego elementu tablicy prawdopodobieństw  $P$  zwracany jest odpowiadający mu kwantyl. Podobnie jak wcześniej  $V1$  i  $V2$  mogą być tablicami pod warunkiem że zgadzają się wymiary. Wartości  $P$  powinny należeć do przedziału  $[0, 1]$ , w przeciwnym wypadku zwracany jest błąd.

### 5.4 fstat

$[M, V] = fstat(V1, V2)$

Dla każdego elementu tablicy  $V1$  i  $V2$ , pod warunkiem że mają one ten sam wymiar (jeżeli liczymy tylko dla jednych wartości stopni swobody będą to tablice jednoelementowe) obliczana jest wartość oczekiwana ( $M$  - mean) i wariancja ( $V$  - variance) rozkładu F z podanymi stopniami swobody.

## 5.5 frnd

$R = frnd(V1, V2)$

$R = frnd(V1, V2, m, n, \dots)$

$R = frnd(V1, V2, [m, n, \dots])$

Jak poprzednio  $V1$  i  $V2$  mogą być tablicami. Funkcja zwraca jedną wartość losową według rozkładu  $F$  lub tablice  $m$  na  $n$  wartości losowych według rozkładu  $F$ . Wymiary tablicy mogą być podawane osobno lub w wektorze wymiarów, zatem ostatnie dwie funkcje są równoważne.



## 6 TI-82 Stats

Na kalkulatorze graficznym *Texas Instruments TI-82 Stats* dostępne są dwie funkcje dla rozkładu F.

- **Fpdf**( - dla gęstości rozkładu F
- **Fcdf**( - dla dystrybuanty rozkładu F

### 6.1 Fpdf(

*Fpdf(X, numerator df, denominator df)*

Funkcja ta zwraca wartość gęstości dla podanego  $X$  i dla podanych stopni swobody *numerator df* w liczniku i *denominator df* w mianowniku. Liczby te muszą być większe od zera inaczej zwracany jest błąd **DOMAIN**. Możliwe jest narysowanie gęstości poprzez wpisane **Fpdf**( w polu "Y =".

### 6.2 Fcdf(

*Fcdf(lowerbound, upperbound, numerator df, denominator df)*

Funkcja ta zwraca wartość dystrybuanty pomiędzy punktami *lowerbound* i *upperbound* (tj  $P(a < X < b)$ ). Podane stopnie swobody są jak poprzednio. Aby obliczyć wartość dystrybuant w punkcie zamiast w przedziale wystarczy przyjąć jako *lowerbound* 0 ponieważ rozkład F nie jest zdefiniowany dla wartości mniejszych od 0.

## 7 Przykład

Przypuszcmy że chcemy zbadać wpływ napoju na czas pracy księgowych, to znaczy czy typ napoju wpływa na czas pracy księgowego. Księgowi podzielono na trzy grupy pod innym napojem, Soda, Napój witaminy-B i kawa. Dla każdego księgowego zbadano czas pracy gdy każdy księgowy dostał to samo zadanie.

Soda		Napój witaminy-B		Kawa	
Księgowy	Godz. pracy	Księgowy	Godz. pracy	Księgowy	Godz. pracy
1	8	6	5	11	7
2	8	7	6	12	6
3	10	8	6	13	7
4	7	9	4	14	7
5	10	10	8	15	9

Następnie potrzebujemy obliczyć Sumę Kwadratów w grupach (SKG), Sumę kwadratów między grupami (SKMG) i całkowitą sumę kwadratów (CSK). Sumę kwadratów w każdej grupie jest to suma kwadratów dla poszczególnych grup która wyraża się wzorem:  $\sum_{i=1}^n (X - \bar{X})^2$ . Zatem musimy dla każdej grupy najpierw wyliczyć wartość średnia. Uzyskane wartości w tabeli poniżej.

Soda		Napój witaminy-B		Kawa	
Śr	SK	Śr	SK	Śr	SK
8.6	7.2	5.8	8.8	7.2	4.8

SKG będzie sumą otrzymanych SK, czyli  $7.2 + 8.8 + 4.8 = 20.8$ .

Następnie obliczymy CSK, czyli wyznaczyliśmy wartość średnia dla całej populacji bez grupowania i skorzystaliśmy z poprzedniego wzoru. Średnia z całej populacji wynosi **7.2** natomiast CSK wynosi **40.4**.

Aby wyznaczyć SKMG potrzebujemy od średniej każdej grupy odjąć średnią całej populacji, następnie uzyskane wartości podnieść do kwadratu i sumować. Otrzymany wynik pomnożony razy liczebność grupy jest szukane SKMG.

- Grupa 1 =  $8.6 - 7.2 = 1.4$
- Grupa 2 =  $5.8 - 7.2 = -1.4$
- Grupa 3 =  $7.2 - 7.2 = 0$

$$1.4^2 + (-1.4)^2 + 0^2 = 1.96 + 1.96 = 3.92$$

Zatem szukane SKMG wynosi  $3.92 \cdot 5 = 19.6$

Teraz potrzebujemy obliczyć stopnie swobody dla każdej grupy i między grupowe. Stopnie swobody między grupowe dane są wzorem **liczba grup - 1** czyli w naszym przypadku wynosi to 2. Stopnie swobody dla każdej grupy wyraża się

wzorem **Liczebność całkowita - liczba grup** czyli w naszym przypadku 12. Następnie SKG i SKMG dzieli przez odpowiadające im stopnie swobody czyli

$$SKG) \frac{20.8}{12} = 1.73$$

$$SKMG) \frac{19.6}{2} = 9.8$$

Wyznamy współczynnik F jako iloczyn otrzymanych wartości tak aby największa znajdowała się w liczniku. Tę wartość porównamy z wartością krytyczną którą poniżej obliczymy.

Zakładając że nasz procent pewności jest 95% nasz level  $\alpha$  wynosi 0.05. Aby znaleźć wartość krytyczną potrzebujemy obliczyć kwantyl odpowiadając za level  $\alpha$  ale z prawej strony wykresu dla rozkładu F ze stopniami swobody SKMG i SKG. Zatem wartość krytyczna K wynosi:

$$K \stackrel{R}{=} qf(0.05, 2, 12, lower.tail = FALSE) \approx 3.885294$$

Wartość F jest większa od wartości K, oznacza to że nasza teza, mówiąca że typ napoju nie wpływa na czas pracy księgowego jest nie prawdziwa. Zatem typ napoju wpływa na czas pracy księgowego.

## 8 Bibliografia

- <https://en.wikipedia.org/wiki/F-distribution>
- [https://en.wikipedia.org/wiki/Beta\\_function#Incomplete\\_beta\\_function](https://en.wikipedia.org/wiki/Beta_function#Incomplete_beta_function)
- <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/FDist>
- <https://support.microsoft.com/en-us/office/f-dist-function-a887efdc-7c8e-46cb-a74a-f884cd29b25d>
- <https://octave.org/doc/v4.2.0/Distributions.html>
- [https://www.mathworks.com/help/stats/referencelist.html?type=function&category=f-distribution-1&s\\_tid=CRUX\\_topnav](https://www.mathworks.com/help/stats/referencelist.html?type=function&category=f-distribution-1&s_tid=CRUX_topnav)
- <https://www.mathworks.com/help/stats/f-distribution.html>
- <https://www.mathworks.com/help/stats/fcdf.html>
- <https://www.mathworks.com/help/stats/fpdf.html>
- <https://www.mathworks.com/help/stats/finv.html>
- <https://www.mathworks.com/help/stats/fstat.html>
- TI-82 STATS GRAPHING CALCULATOR GUIDEBOOK - Texas Instruments Incorporated
- <https://www.whatissixsigma.net/anova-f-test/>
- <https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/p-value/>
- <https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-an-alpha-level/>