

Matematyczne aspekty analizy danych

semestr zimowy 2025/2026

Dr Anna Muranova
UWM w Olsztynie

Wykład 4

Przykład

Prawdopodobieństwo że będzie padać wynosi 30%, a prawdopodobieństwo, że parasol będzie dostarczony na czas – 40%. Jeżeli to są niezależne zdarzenia, to

- ▶ Chcesz pospacerować w deszcz, ale z parasolem. Jaki jest prawdopodobieństwo, że będzie padać oraz parasol będzie dostarczone na czas?
- ▶ Musisz załatwić swoje sprawy. Jaki jest prawdopodobieństwo, że nie będzie padać lub parasol będzie dostarczone na czas?

Przykład 1

Prawdopodobieństwo że będzie padać wynosi 30%, a prawdopodobieństwo, że parasol będzie dostarczony na czas – 40%. Jeżeli to są niezależne zdarzenia, to

- Chcesz pospacerować w deszcz, ale z parasolem. Jaki jest prawdopodobieństwo, że będzie padać oraz parasol będzie dostarczony na czas?

$$A = \{\text{będzie padać}\}$$

$$B = \{\text{parasol będzie dostarczony na czas}\}$$

$$P(A) = 0.3, P(B) = 0.4$$

$$P(A \cap B) = [\text{niezależność}] = P(A) \cdot P(B) = 0.3 \cdot 0.4 = 0.12$$

Przykład 2

Prawdopodobieństwo że będzie padać wynosi 30%, a prawdopodobieństwo, że parasol będzie dostarczony na czas – 40%. Jeżeli to są niezależne zdarzenia, to

- Musisz załatwić swoje sprawy. Jaki jest prawdopodobieństwo, że nie będzie padać lub parasol będzie dostarczony na czas?

$$A = \{\text{nie będzie padać}\}$$

$$B = \{\text{parasol będzie dostarczony na czas}\}$$

$$P(A) = 1 - P(\{\text{będzie padać}\}) = 1 - 0.3 = 0.7$$

$$P(B) = 0.4$$

Uwaga! Jeżeli A i B są niezależni, to A i nie B też są niezależni.

Dlatego:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.7 + 0.4 - 0.7 \cdot 0.4 = 0.82$$

Prawdopodobieństwo warunkowe 1

Prawdopodobieństwo, że doszło do zdarzenia A , jeśli wystąpiło zdarzenie B .

$$P(A \text{ jeśli } B) = P(A | B)$$

Przykład: rzucamy uczciwy sześcián:

$$A = \{\text{dwójka}\}, B = \{\text{liczba parzysta}\}$$

1, 2, 3, 4, 5, 6

$$P(A | B) = ?.$$

2, 4, 6

$$P(A | B) = 1/3.$$

$$P(B | A) = ?.$$

2

$$P(B | A) = 1.$$

Prawdopodobieństwo warunkowe 2

Wzór:

$$P(A | B) = \frac{\text{Ilość wyników } A \text{ oraz } B}{\text{Ilość wyników } B} = \frac{\frac{\text{Ilość wyników } A \text{ oraz } B}{\text{Ilość wszystkich wyników}}}{\frac{\text{Ilość wyników } B}{\text{Ilość wszystkich wyników}}} = \frac{P(A \cap B)}{P(B)},$$

Takim czynem,

$$\frac{P(A \cap B)}{P(B)} = P(A | B) \text{ oraz } P(B | A) = \frac{P(A \cap B)}{P(A)}$$

oraz

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}.$$

Kierunek warunku ma znaczenie!

„85% chorych na raka pilo kawę”

Stany Zjednoczone:

$$P(Kawa) = 0.65$$

(statista.com, 65%)

$$P(Rak) = 0.005$$

(cancer.gov. 0.5%)

„85% chorych na raka pilo kawę”: $P(Kawa | Rak) = 0.85$

Co nas naprawdę interesuje jest

$$P(Rak | Kawa) = \frac{P(Kawa | Rak)P(Rak)}{P(Kawa)} = \frac{0.85 \cdot 0.005}{0.65} = 0,0065(0,65\%)$$

$$P(Rak | nie Kawa) = ?$$

Z czym możemy porównać?

$$P(Rak \mid nie\ Kawa) = \frac{P(Rak \cap nie\ Kawa)}{P(nie\ Kawa)} = \frac{P(Rak \cap nie\ Kawa)}{1 - P(Kawa)}?$$

$$P(Rak \cap nie\ Kawa) = 1 - P(nie\ Rak) - P(Rak\ i\ Kawa)$$

Ponieważ, $P(Kawa\ i\ Rak) = P(Kawa \mid Rak)P(Rak) = 0.85 \cdot 0.005 = 0.00425$,
mamy

$$P(Rak \cap nie\ Kawa) = 1 - (1 - 0.005) - 0.00425 = 0.00075$$

$$P(Rak \mid nie\ Kawa) = \frac{P(Rak \cap nie\ Kawa)}{1 - P(Kawa)} = \frac{0.00075}{0.35} = 0.002$$

Wniosek:

$$P(Rak) = 0.005(0,5\%)$$

$$P(Rak \mid Kawa) = 0.0065(0,65\%)$$

$$P(Rak \mid nie\ Kawa) = 0.002(0,2\%)$$

Co definiuje „pijącego kawę”?

Moglibyśmy tu uwzględnić inne zmienne, w szczególności to, co kwalifikuje osobę jako „pijącego kawę”. Jeśli ktoś pije kawę raz na miesiąc, a ktoś inny codziennie, czy powinniśmy określić obie te osoby jako „pijące kawę”? Co z osobą, która zaczęła pić kawę miesiąc temu, w porównaniu z osobą, która pije ją od 120 lat? Jak często i przez jaki długi czas musi pić kawę, aby przekroczyć próg bycia „pijącym kawę” w tych badaniach nad rakiem?

Są to ważne pytania, które pokazują, dlaczego dane rzadko opowiadają całą historię. Jeśli ktoś pokazuje Ci tabelę pacjentów z prostymi odpowiedziami „TAK/NIE” na pytanie, czy piją kawę, ten próg musi być zdefiniowany! Albo potrzebny jest bardziej miarodajny parametr, taki jak „liczba kaw wypitych w ciągu ostatnich trzech lat”. Nie komplikujemy tego teoretycznego obliczenia, i nie umawiamy, co kwalifikuje kogoś jako „pijącego kawę”, ale w rzeczywistych badaniach zawsze warto dobrze zastanowić się nad danymi.

Przykład

Prawdopodobieństwo że będzie padać wynosi 30%, a prawdopodobieństwo, że parasol będzie dostarczony na czas wynosi 40% jeżeli nie będzie padać oraz 20%, jeżeli będzie.

- ▶ Chcesz pospacerować w deszcz, ale z parasolem. Jaki jest prawdopodobieństwo, że będzie padać oraz parasol będzie dostarczony na czas?
- ▶ Musisz załatwić swoje sprawy. Jaki jest prawdopodobieństwo, że nie będzie padać lub parasol będzie dostarczony na czas?

Przykład 1

Prawdopodobieństwo że będzie padać wynosi 30%, a prawdopodobieństwo, że parasol będzie dostarczony na czas wynosi 40% jeżeli nie będzie padać oraz 20%, jeżeli będzie.

- ▶ Chcesz pospacerować w deszcz, ale z parasolem. Jaki jest prawdopodobieństwo, że będzie padać oraz parasol będzie dostarczony na czas?

$$A = \{\text{będzie padać}\}$$

$$B = \{\text{parasol będzie dostarczony na czas}\}$$

$$P(A) = 0.3, P(B) = 0.4, P(B | A) = 0.2$$

$$P(A \cap B) = P(B | A)P(A) = 0.2 \cdot 0.3 = 0.06.$$

Przykład 2

Prawdopodobieństwo że będzie padać wynosi 30%, a prawdopodobieństwo, że parasol będzie dostarczony na czas wynosi 40% jeżeli nie będzie padać oraz 20%, jeżeli będzie.

- Musisz załatwić swoje sprawy. Jaki jest prawdopodobieństwo, że nie będzie padać lub parasol będzie dostarczony na czas?

$$\begin{aligned} &P(\{\text{nie będzie padać lub parasol będzie dostarczony na czas}\}) \\ &= 1 - P(\{\text{będzie padać i parasol nie będzie dostarczony na czas}\}) \end{aligned}$$

$$A = \{\text{będzie padać}\}$$

$$B = \{\text{parasol nie będzie dostarczony na czas}\}$$

$$P(A) = 0.3, P(B | A) = 1 - 0.2 = 0.8$$

$$P(A \cap B) = P(B | A)P(A) = 0.8 \cdot 0.3 = 0.24.$$

$$P(\{\text{nie będzie padać lub parasol będzie dostarczony na czas}\}) = 1 - 0.24 = 0.76$$

Przykład 2

Prawdopodobieństwo że będzie padać wynosi 30%, a prawdopodobieństwo, że parasol będzie dostarczony na czas wynosi 40% jeżeli nie będzie padać oraz 20%, jeżeli będzie.

- Musisz załatwić swoje sprawy. Jaki jest prawdopodobieństwo, że nie będzie padać lub parasol będzie dostarczony na czas?

Inne rozwiązanie:

$$\begin{aligned} &P(\{\text{nie będzie padać lub parasol będzie dostarczony na czas}\}) \\ &= P(\{\text{nie będzie padać i parasol będzie dostarczony na czas}\}) \\ &+ P(\{\text{nie będzie padać i parasol nie będzie dostarczony na czas}\}) \\ &\quad + P(\{\text{będzie padać i parasol będzie dostarczony na czas}\}) \\ &\quad = P(\{\text{nie będzie padać}\}) \\ &\quad + P(\{\text{będzie padać i parasol będzie dostarczony na czas}\}) \\ &\quad = (1 - 0.3) + 0.06 = 0.76 \end{aligned}$$

Ponieważ:

$$\begin{aligned} A &= \{\text{będzie padać}\} \\ B &= \{\text{parasol będzie dostarczony na czas}\} \\ P(A \cap B) &= P(B | A)P(A) = 0.2 \cdot 0.3 = 0.06 \end{aligned}$$

Rozkład dwumianowy: prosty przykład

- Jaki jest prawdopodobieństwo, że w trzech rzutach uczciwych:

0. nie wypadnie żadnej reszki
1. wypadnie dokładnie jedna reszka
2. wypadnie dwie reszki
3. wypadnie trzy reszki

OOO

OOR

ORO

ORR

ROO

ROR

RRO

RRR

0. nie wypadnie żadnej reszki – $1/8$
1. wypadnie dokładnie jedna reszka – $3/8$
2. wypadnie dwie reszki – $3/8$
3. wypadnie trzy reszki – $1/8$

Rozkład dwumianowy: przykład

- Jaki jest prawdopodobieństwo, że w trzech rzutach:

0. nie wypadnie żadnej reszki
1. wypadnie dokładnie jedna reszka
2. wypadnie dwie reszki
3. wypadnie trzy reszki

jeżeli prawdopodobieństwo reszki jest p ?

$$OOO - p^3$$

$$OOR - p(1 - p)^2$$

$$ORO - p(1 - p)^2$$

$$ORR - p^2(1 - p)$$

$$ROO - p(1 - p)^2$$

$$ROR - p^2(1 - p)$$

$$RRO - p^2(1 - p)$$

$$RRR - p^3$$

0. nie wypadnie żadnej reszki – $(1 - p)^3$
1. wypadnie dokładnie jedna reszka – $3p(1 - p)^2$
2. wypadnie dwie reszki – $3p^2(1 - p)$
3. wypadnie trzy reszki – p^3

Rozkład dwumianowy

- ▶ Jaki jest prawdopodobieństwo, że w n rzutach wypadnie k reszek?

Trzeba wybrać k miejsc w ciągu n prób. Wtedy prawdopodobieństwa wystąpienia ciągu, gdzie na wybranych miejscach jest reszki, a na innych orły jest $p^k(1-p)^{n-k}$.

Liczba możliwości wybrania k miejsc w ciągu długości n oznacza się $\binom{n}{k}$ i wynosi:

$$\binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-(k-1))}{k(k-1)\dots 1} = \frac{n!}{(n-k)!k!}$$

Takim czynnem, prawdopodobieństwo, że w n rzutach wypadnie k reszek jest $\binom{n}{k} p^k (1-p)^{n-k}$

Rozkład dwumianowy: definicja

Skończony ciąg niezależnych powtórzeń tego samego doświadczenia o dwóch możliwych wynikach nazywamy *schematem Bernoulliego*.

Możliwe wyniki pojedynczego zdarzenia oznaczmy jako 1 (sukces) oraz 0 (porażka). Niech prawdopodobieństwo sukcesu wynosi p (wówczas prawdopodobieństwo porażki wynosi $1 - p$).

Prawdopodobieństwo pojawienia się dokładnie k sukcesów w schemacie n prób Bernoulliego nazywa się *rozkładem dwumianowym* i wynosi

$$\binom{n}{k} p^k (1 - p)^{n-k},$$

gdzie p oznacza prawdopodobieństwo sukcesu w pojedynczej próbie.

Rozkład dwumianowy: jeszcze przykład

- ▶ Jaki jest prawdopodobieństwo, że w 10 rzutach uczciwych kostki wypadnie 6 razy liczba, podzielna przez 3?

$$p = 1/3, n = 10, k = 6$$

$$\binom{n}{k} p^k (1-p)^{n-k} = \binom{10}{6} (1/3)^6 (1 - 1/3)^{10-6} = 0.056$$

```
import math
print(math.factorial(10)/(math.factorial(6)*
math.factorial(4))*(1/3)**6*(2/3)**(4))
```

Rozkład dwumianowy: Python

$$p = 1/3, n = 10, k = 6$$

```
import sympy
from sympy.stats import Binomial, density

print(sympy.binomial(10,6)*(1/3)**6*(2/3)**(4))
X = Binomial('X', 10, 1/3, succ=1, fail=0)
print(density(X).dict)
print(density(X).dict[6])
```

0.0569018950363257

```
{0: 0.0173415299158326, 1: 0.0867076495791631, 2: 0.195092211553117, 3:
0.260122948737489, 4: 0.227607580145303, 5: 0.136564548087182, 6:
0.0569018950363258, 7: 0.0162576842960931, 8: 0.00304831580551745, 9:
0.000338701756168606, 10: 1.69350878084303e-5}
```

0.0569018950363258

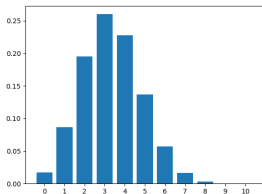
Rozkład dwumianowy: wykres

```
import matplotlib.pyplot as plt
from sympy.stats import Binomial, density

X = Binomial('X', 10, 1/3, succ=1, fail=0)
data = density(X).dict

names = list(data.keys())
values = list(data.values())

plt.bar(range(len(data)), values, tick_label=names)
plt.show()
```



Rozkład dwumianowy: definicja

Skończony ciąg niezależnych powtórzeń tego samego doświadczenia o dwóch możliwych wynikach nazywamy *schematem Bernoulliego*.

Możliwe wyniki pojedynczego zdarzenia oznaczmy jako 1 (sukces) oraz 0 (porażka). Niech prawdopodobieństwo sukcesu wynosi p (wówczas prawdopodobieństwo porażki wynosi $1 - p$).

Prawdopodobieństwo pojawienia się dokładnie k sukcesów w schemacie n prób Bernoulliego nazywa się *rozkładem dwumianowym* i wynosi

$$\binom{n}{k} p^k (1 - p)^{n-k},$$

gdzie p oznacza prawdopodobieństwo sukcesu w pojedynczej próbie.

Przykład: rozważanie o rozkładzie dwumianowym

Przypuśćmy, że pracujesz nad nowym silnikiem turbodoładowym i przeprowadziłeś 10 testów. Wynikiem są osiem sukcesów i dwa niepowodzenia:

x v v v v v v x v v

Liczyłeś na 90-procentowy wskaźnik sukcesu, ale na podstawie tych danych dochodzisz do wniosku, że testy zawiodły, ponieważ sukcesem zakończyło się tylko 80% prób.

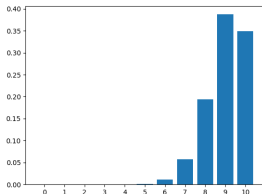
Ile naprawdę wynosi prawdopodobieństwo uzyskać 8 sukcesów z 10 (przy prawdopodobieństwie sukcesu 0.9)?

0.193710244500000

Przykład: rozważanie o rozkładzie dwumianowym

```
import matplotlib.pyplot as plt
from sympy.stats import Binomial, density

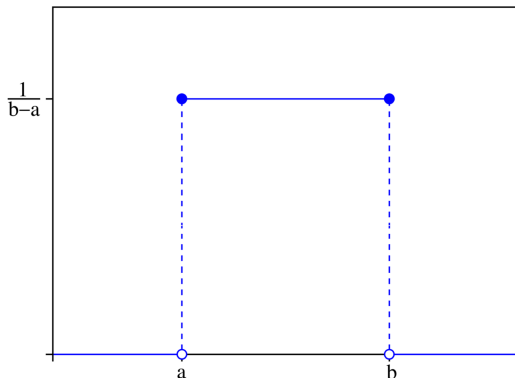
X = Binomial('X', 10, 0.9, succ=1, fail=0)
data = density(X).dict
names = list(data.keys())
values = list(data.values())
print(density(X).dict[8])#0.193710244500000
plt.bar(range(len(data)), values, tick_label=names)
plt.show()
```



Rozkład jednostajny i funkcja gęstości

Nich x przyjmuje wartości rzeczywiste na $[0, 1]$ (lub $[a, b]$), wszystkie z równym prawdopodobieństwem. Z jakim?

Funkcja gęstości:



Domena publiczna, [https:](https://commons.wikimedia.org/w/index.php?curid=117047)

[//commons.wikimedia.org/w/index.php?curid=117047](https://commons.wikimedia.org/w/index.php?curid=117047)

Rozkład beta

Musimy jakoś odwrócić ten model. Jak oszacować wiarygodność sukcesu po liczbie sukcesów w próbie?

Rozkład beta pozwala nam sprawdzić, jaki jest prawdopodobieństwo różnych bazowych prawdopodobieństw zajścia zdarzenia przy α sukcesów i β niepowodzeniach.

Rozkład beta jest zdefiniowany przy pomocy funkcji **gęstości**

$$f(\alpha, \beta, x) = c_{\alpha, \beta} \cdot x^{\alpha-1} (1-x)^{\beta-1}$$

$$c_{\alpha, \beta} = \frac{1}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)},$$

– stała, gdzie B jest funkcją beta, Γ – funkcją gamma.

Funkcje Gamma i beta

Funkcja B (*funkcja beta*) zwana też całką Eulera pierwszego rodzaju – funkcja specjalna określona dla liczb $x > 0, y > 0$ takich że ich części rzeczywiste są dodatnie, dana wzorem

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

Funkcja Γ (zwana też funkcją gamma Eulera) – funkcja specjalna, która rozszerza pojęcie silni na zbiór liczb rzeczywistych i zespolonych. Jeżeli $x > 0$, to

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt.$$

Funkcje Gamma i Beta: wzory

- ▶ Dla liczb naturalnych: $\Gamma(n+1) = n!$,
- ▶ Dla dowolnych liczb mamy:

$$\Gamma(x) = \lim_{n \rightarrow +\infty} \frac{n! n^x}{x(x+1)(x+2) \dots (x+n)} = \frac{1}{x} \prod_{n=1}^{\infty} \frac{(1 + \frac{1}{n})^x}{1 + \frac{x}{n}}.$$

- ▶ Dla każdego $x, y > 0$ zachodzi:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

- ▶ Dla liczb naturalnych m, n zachodzi:

$$B(n, m) = \frac{(n-1)!(m-1)!}{(n+m-1)!}$$

- ▶ Dla każdego $x, y > 0$ zachodzi:

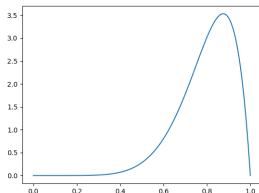
$$B(x, y) = \int_0^{\infty} \frac{t^{x-1}}{(1+t)^{x+y}} dt$$

Funkcja gęstości

Rozkład beta jest zdefiniowany przy pomocy funkcji **gęstości**

$$f(\alpha, \beta, x) = c_{\alpha, \beta} \cdot x^{\alpha-1} (1-x)^{\beta-1}$$

$$\alpha = 8, \beta = 2$$



Ponieważ parametr θ liczby sukcesów może być każdą liczbą rzeczywistą od 0 do 1, a liczb tych jest nieskończone wiele, to prawdopodobieństwo definiuje się dla zakresów:

$$P(a < \theta < b) = \int_a^b f(\alpha, \beta, x) dx.$$

Wykres w Pythonie

Wykres w Pythonie:

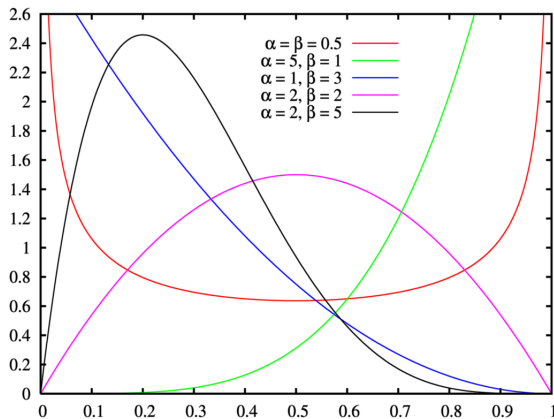
```
from sympy.stats import Beta, density
import matplotlib.pyplot as plt
from sympy import*

alpha = 8
beta = 2
X = Beta("X", alpha, beta)
x = symbols('x')
print(density(X)(x))#x**7*(1 - x)/beta(8, 2)
x0 = [i/100 for i in range (0,101)]
y0 = [density(X)(i).evalf() for i in x0]
plt.plot(x0,y0)
plt.show()
```

Obliczenia, ze $0.85 < \theta < 0.95$:

```
print(integrate(density(X)(x),(x, 0.85,0.95)).evalf())#0.32930944871875
#Uwaga:
print(integrate(density(X)(x),(x, 0,1)).evalf())#1.0000000000000000
```

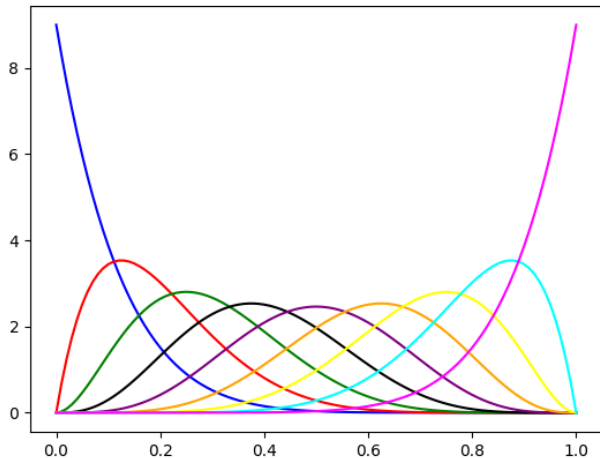
Rozkład beta dla różnych α i β



Rysunek: Rozkłady beta. Źródło:

https://pl.wikipedia.org/wiki/Rozk%C5%82ad_beta

Rozkład beta dla różnych α i β w Pythonie



Rozkład beta dla różnych α i β w Pythonie

```
from sympy.stats import Beta, density
import matplotlib.pyplot as plt

color = ["blue", "red", "green",
         "black", "purple", "orange",
         "yellow", "cyan", "magenta"]
for alpha in range(1,10):
    beta = 10 - alpha
    X = Beta("X", alpha, beta)
    x0 = [i/100 for i in range (0,101)]
    y0 = [density(X)(i).evalf() for i in x0]
    plt.plot(x0,y0, color= color[alpha-1])

plt.show()
```

Przykład

Założmy, że mamy sekwencję nukleotydów (część DNA, nukleotydy: A, C, G, T) i jesteśmy zainteresowani konkretnym nukleotydem, powiedzmy A. Założmy, że prawdopodobieństwo wystąpienia A, $P(A) = \theta$, jest nieznane, ale pozostaje stałe. Niech w obserwowanych danych o rozmiarze $n = 100$ mamy 43 nukleotydy A.

Jaka jest funkcja gęstości? Narysować wykres funkcji gęstości. Jaki jest prawdopodobieństwo ze $0.4 < \theta < 0.5$ oraz $0.1 < \theta < 0.2$?

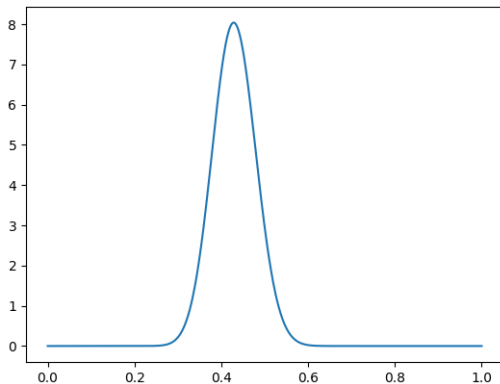
Przykład: kod

$$\alpha = 43, \beta = 100 - 43$$

```
from sympy.stats import Beta, density
import matplotlib.pyplot as plt
from sympy import *

alpha = 43
beta = 100-43
X = Beta("X", alpha, beta)
x = symbols('x')
print(density(X)(x))
x0 = [i/1000 for i in range (0,1001)]
y0 = [density(X)(i).evalf() for i in x0]
plt.plot(x0,y0)
plt.show()
```

Przykład: wykres



Funkcja gęstości:

$$f(43, 57, x) = \frac{x^{42}(1-x)^{56}}{B(43, 57)}$$

Obliczenia $0.4 < \theta < 0.5$ oraz $0.1 < \theta < 0.2$

```
from sympy.stats import Beta, density
from sympy import *

alpha = 43
beta = 57
X = Beta("X", alpha, beta)

x = symbols('x')
print(integrate(density(X)(x), (x, 0.4, 0.5)).evalf())
#86552614.7769390 Uwaga!!! Błąd obliczeniowy!
print(integrate(density(X)(x), (x, Rational(4,10),
    Rational(5,10))).evalf())
#0.645898492575897
print(integrate(density(X)(x), (x, 0.1, 0.2)).evalf())
#1.04028876231591e-7
print(integrate(density(X)(x), (x, Rational(1,10),
    Rational(2,10))).evalf())#1.04028884394972e-7
print(integrate(density(X)(x), (x, Rational(35,100),
    Rational(55,100))).evalf())#0.941365862146943
#przedzial ufnosci
```