AKADEMIA GÓRNICZO-HUTNICZA

im. Stanisława Staszica w Krakowie

WYDZIAŁ ZARZĄDZANIA

PROJEKT NR 1 - ANALIZA ROZKŁADU A POSTERIORI DLA ROZKŁADU BETA

METODY BAYESOWSKIE

Autorzy: Krok Justyna, Artur Karamon

Kierunek: Informatyka i Ekonometria

Studia: II Stopnia, Stacjonarne

Kraków – 2020

POLECENIE:

Niech cecha X ma rozkład zero-jedynkowy z parametrem p. Założyć pewien rozkład a priori z rodziny rozkładów beta. Następnie z rozkładu zero-jedynkowego z prawd. sukcesu 0,8 wylosować 1000 danych $x_1, ..., x_{1000}$. Rozważając kolejne dane x_i obliczać rozkład a posteriori i na jego podstawie bayesowski estymator parametru p i jego odchylenie standardowe. Dla kolejnych $i = 1, ..., 1000$. porównać te wartości z odpowiadającymi im wartościami uzyskanymi za pomocą MNW. Rozważyć różne rozkłady a priori (powinny one mieć tę samą wartość oczekiwaną (można ja np. wylosować z przedziału [0,1]), ale różne odchylenia standardowe świadczące o różnej pewności co do wartości parametru).

TEORIA:

Rozkład beta:

Rozkład beta $B(\alpha, \beta)$ ma funkcję gęstości postaci:

(1) $f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1} (1-p)^{\beta-1}$ *dla* $p \in [0,1]$, gdzie:

(2) $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$.

Jeżeli zmienna losowa X ma rozkład $B(\alpha, \beta)$, to:

(3) $E(X) = \frac{\alpha}{\alpha+\beta}$,

(4) $D^2(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

A priori/ a posteriori:

Niech cecha X ma rozkład zero-jedynkowy (dwupunktowy) z nieznanym parametrem p (prawdopodobieństwo sukcesu w pojedynczej próbie).

Na podstawie danych $x = \{x_1, ..., x_n\}$, gdzie $n = 1000$ możemy określić liczbę sukcesów k.

Funkcja wiarygodności dla całego zbioru ma postać

(5) $P(x_i|p) = p^k\ (1-p)^{n-k}$, gdzie:

(6) $k = \sum_{i=1}^{n} x_i$.

Przyjmijmy, że rozkład a priori parametru p jest rozkładem beta.

Rozkład a priori $B(\alpha, \beta)$:

(7) $f(p) = \frac{1}{B(\alpha,\beta)}\ p^{\alpha-1}\ (1-p)^{\beta-1}$

Rozkład a posteriori $B(\alpha + k, n - k + \beta)$:

(8) $f(p|x) = \frac{P(x|p)\,f(p)}{\int_o^1 P(x|p)\,f(p)\,dp} = \frac{1}{B(k+\alpha,n-k+\beta)}\ p^{k+\alpha-1}\ (1-p)^{n-k+\beta-1}$

Dla rozkładu a posteriori:

(9) $E(p|k) = \frac{\alpha+k}{\alpha+k+n-k+\beta} = \frac{\alpha+k}{\alpha+n+\beta}$

(10) $D^2(p|k) = \frac{(\alpha+k)(n-k+\beta)}{(\alpha+k+n-k+\beta)^2(\alpha+k+n-k+\beta+1)} = \frac{(\alpha+k)(n-k+\beta)}{(\alpha+n+\beta)^2(\alpha+n+\beta+1)}$

BADANIE:

Badanie zawiera kilka prostych kroków, które pozwolą na otrzymanie oszacowań rozkładów a posteriori. Wymaga ono zdefiniowania 2 funkcji:

- **BetaParameters** - oblicza wartości parametrów rozkładu Beta (alpha, beta) na podstawie wartości oczekiwanej oraz wariancji rozkładu
- **aposteriori** - losuje wektor liczb z rozkładu zero-jedynkowego z prawdopodobieństwem 0,8; generuje parametry rozkładu o zadanych: wartości oczekiwanej oraz wariancji; wykonuje 1000 iteracji, w których obliczany jest rozkład a posteriori oraz bayesowski estymator p i jego odchylenie standardowe, w każdej kolejnym przejściu pętli rozważany zakres informacji zwiększany jest o 1 daną

W funkcji aposteriori do obliczenia wartości oczekiwanej parametru p (zmienna theta_ex) korzystano ze wzoru nr 9, do obliczenia wariancji p (zmienna theta_var)

korzystano ze wzoru nr 10. Wartości parametrów alfa oraz beta rozkładu a posteriori obliczono na podstawie wartości oczekiwanej i wariancji p (wykorzystano funkcję BetaParameters).

```
BetaParameters <- function(EX, VAR) {
  alpha <- EX*(EX*(1-EX)-VAR)/VAR
  beta <- (1-EX)*(EX*(1-EX)-VAR)/VAR
  return(params = list(alpha = alpha, beta = beta))
}
```

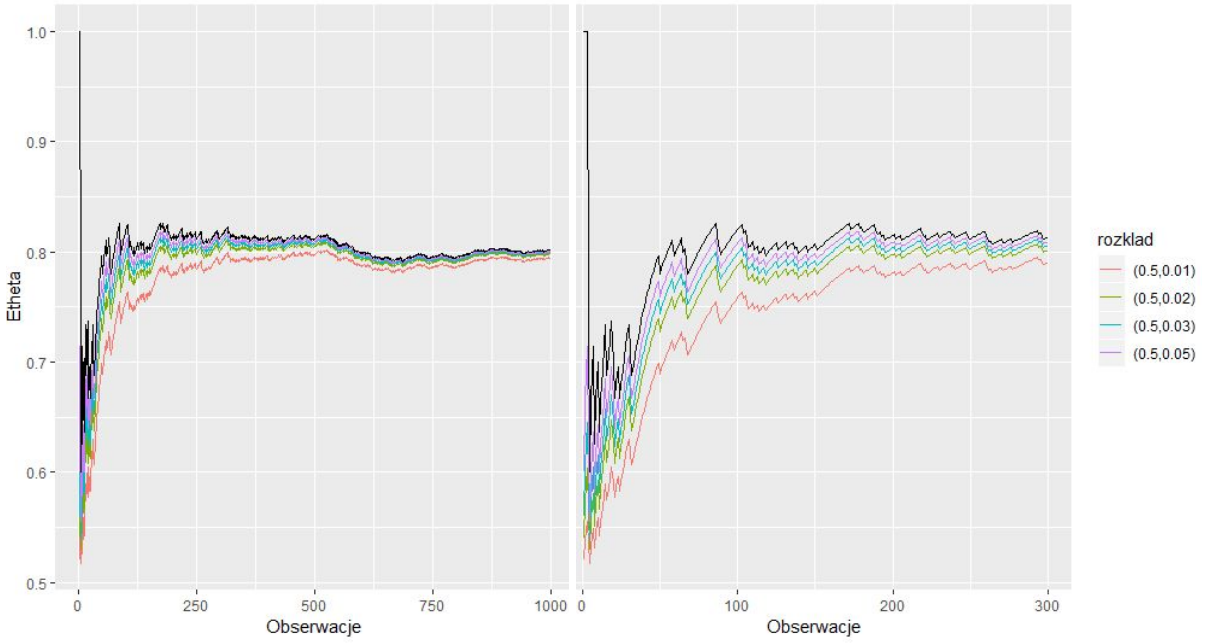
Rysunek 1: Definicja funkcji "BetaParameters".

```
aposteriori<-function(mi,tau){
  n<-1000
  set.seed(123)
  vector<-rbinom(n,1,0.8)

  alpha<-BetaParameters(mi,tau)[[1]]
  beta<-BetaParameters(mi,tau)[[2]]
  p = seq(0,1, length=n)
  rapriori<-dbeta(p, alpha, beta)

  x<-c()
  k<-c()
  theta_ex<-c()
  theta_var<-c()
  raposteriori_list<-list()

  for (i in 1:n) {
    x[i]<-mean(vector[1:i])
    k[i]<-sum(vector[1:i])
    theta_ex[i]<-(k[i]+alpha)/(alpha+i+beta)
    theta_var[i]<-((k[i]+alpha)*(i-k[i]+beta))/((alpha+i+beta)^2*(alpha+i+beta+1))
    alpha_post<-BetaParameters(theta_ex[i],theta_var[i])[[1]]
    beta_post<-BetaParameters(theta_ex[i],theta_var[i])[[2]]
    raposteriori<-dbeta(p, alpha_post, beta_post)
    raposteriori_list[[i]]<-raposteriori
  }
  return(list(x=x,theta_ex=theta_ex,theta_var=theta_var,rapriori=rapriori,raposteriori=raposteriori_list))
}
```

Rysunek 2: Definicja funkcji "aposteriori".

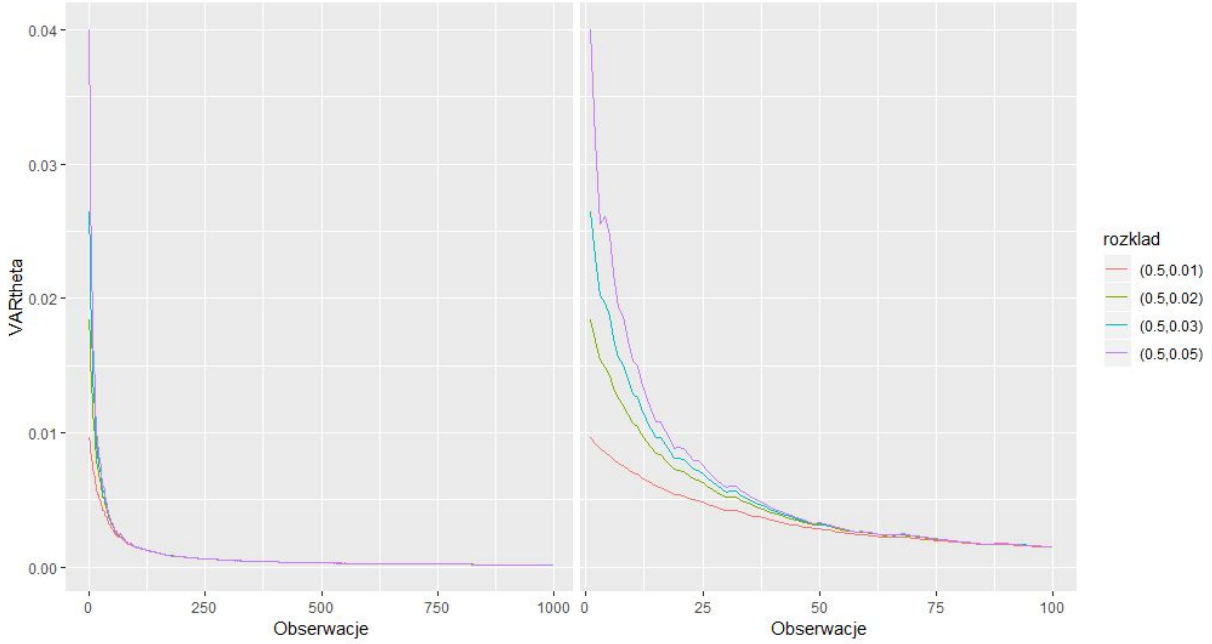
WYNIKI:

Wyniki programu zawierają dużą liczbę danych, a więc znacznie lepszą metodą przedstawienia rezultatów jest wyświetlanie wykresów, a nie otrzymanych wartości. Poniżej znajdują się wykresy przedstawiające zmianę wartości oczekiwanej i wariancji parametru p wraz ze wzrostem liczby obserwacji n.



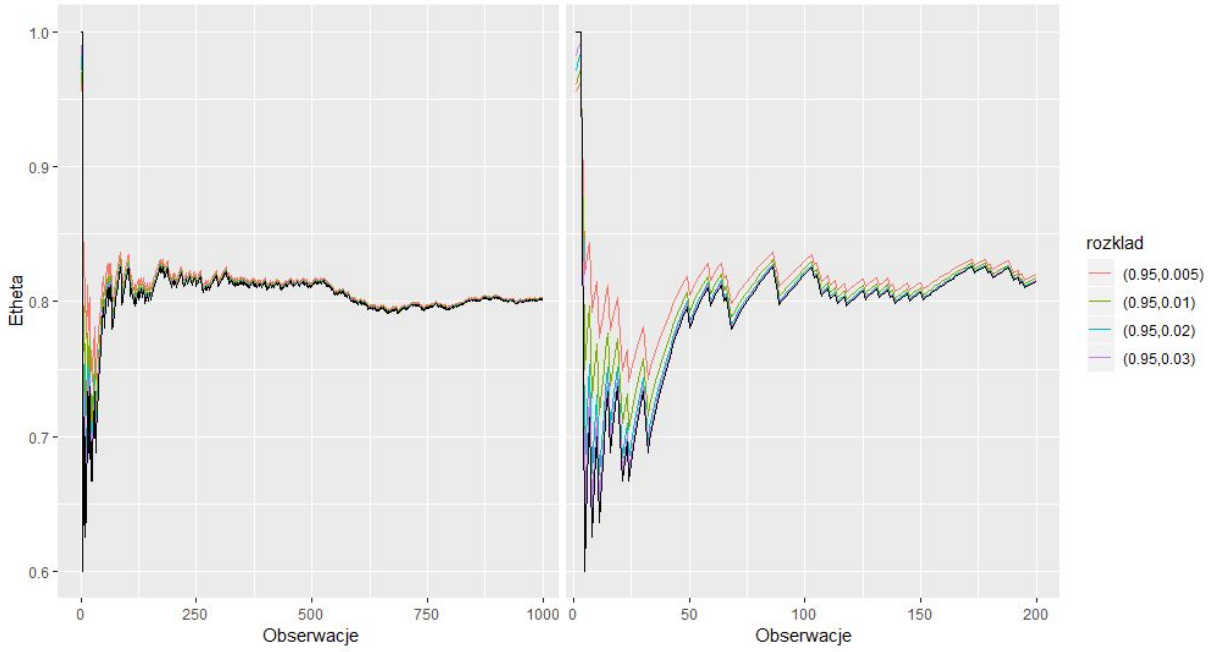
Wykres 1: Wartość oczekiwana a posteriori.

Na podstawie wykresów wartości oczekiwanej rozkładu a posteriori dla pierwszej grupy rozkładów, o stałej wartości oczekiwanej $EX = 0,5$ oraz zmiennej wariancji, można stwierdzić, że wzrost zakładanej wariancji w rozkładzie a priori wpływa na większe podobieństwo między oszacowaniami wartości oczekiwanej metodą bayesowską oraz metodą MNW. Różnice w oszacowaniach są tym mniejsze, im wyższa jest wariancji początkowa a priori.



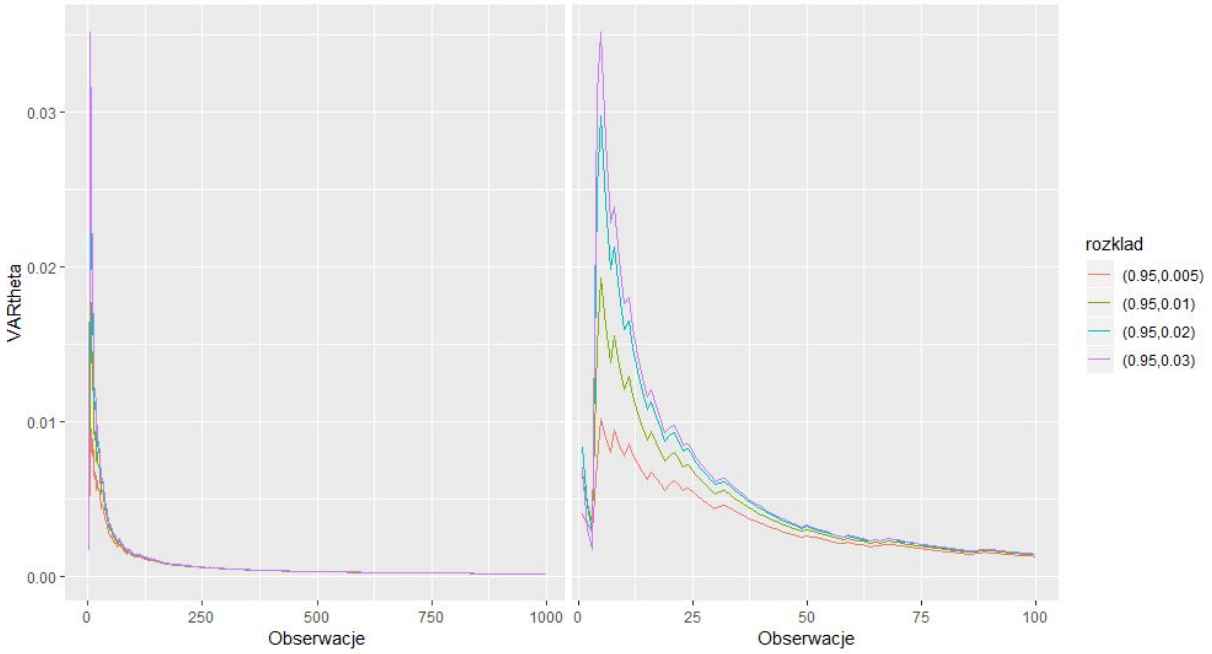
Wykres 2: Wariancja a posteriori.

Na podstawie wykresów wariancji rozkładu a posteriori dla pierwszej grupy rozkładów można stwierdzić, że następuje jej dynamiczny spadek już przy znajomości pierwszych kilkudziesięciu obserwacjach. Po tej liczbie wariancja dla wszystkich rozkładów ma już niemalże identyczną wartość. Wykres wariancji ma postać hiperboliczną. W dalszej perspektywie zmiany są już niewielkie, jednak widać, że jej wartość dąży do 0.



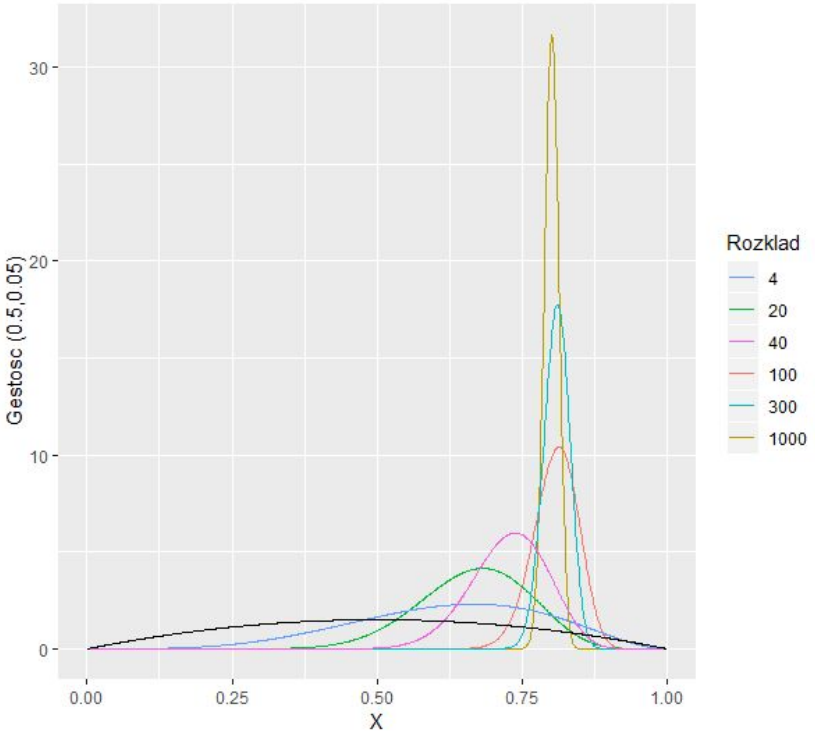
Wykres 3: Wartość oczekiwana a posteriori.

Na podstawie wykresów wartości oczekiwanej rozkładu a posteriori dla drugiej grupy rozkładów, o stałej wartości oczekiwanej $EX = 0,95$ oraz zmiennej wariancji, może wysnuć podobne wnioski, wzrost wariancji przybliża oszacowania metodą bayesowską do oszacować metodą MNK. Wykres wartości oczekiwanej dla pierwszej i drugiej grupy znacznie się różni, spowodowane jest to zmianą parametrów alfa i beta wraz ze zmianą wartości oczekiwanej czy wariancji.



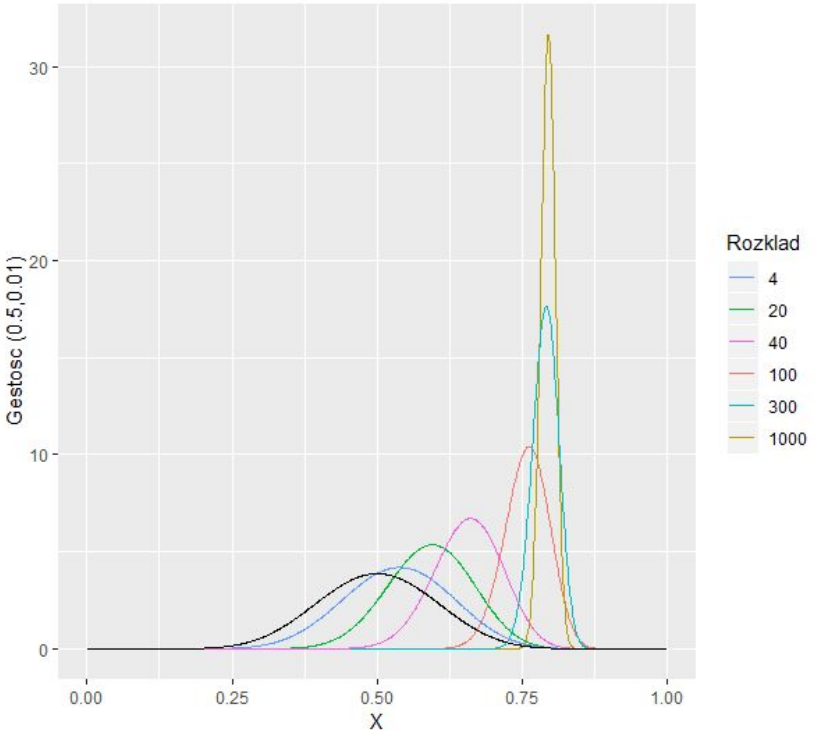
Wykres 4: Wariancja a posteriori.

Wykres wariancji dla drugiej grupy rozkładów zachowuje się podobnie jak w pierwszym przypadku. Funkcja przypomina hiperbole i drastycznie spada. Analizując wykres dla pierwszych stu wartości możemy zauważyć, że większa wariancji rozkładu a priori skutkuje większą wariancją rozkładu a posteriori, różnica wariancji ma wpływ tylko dla początkowych wartości n. Wraz ze wzrostem n, różnice te zanikają.



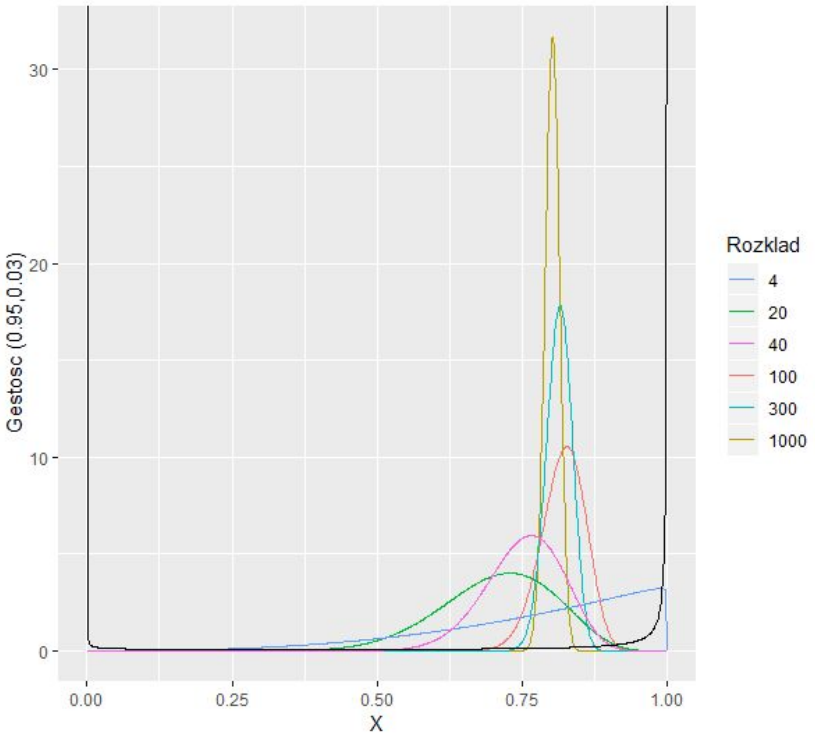
Wykres 5: Rozkłady a priori i a posteriori dla kolejnych danych.

Powyżej znajduje się wykres rozkładu a posteriori dla $EX = 0,5$ oraz $D^2X = 0,05$. Czarna linia przedstawia średnią (oszacowanie metodą MNW). Wraz ze wzrostem n rozkłady są coraz wyższe i węższe, oznacza to, że wzrasta pewność rozkładu (zmniejsza się wariancja). W momencie kiedy n wynosi 100, dalsze zwiększanie n zwiększa pewność rozkłady, ale nie przesuwa go w bok (wartość oczekiwana wydaje się ustabilizowana).



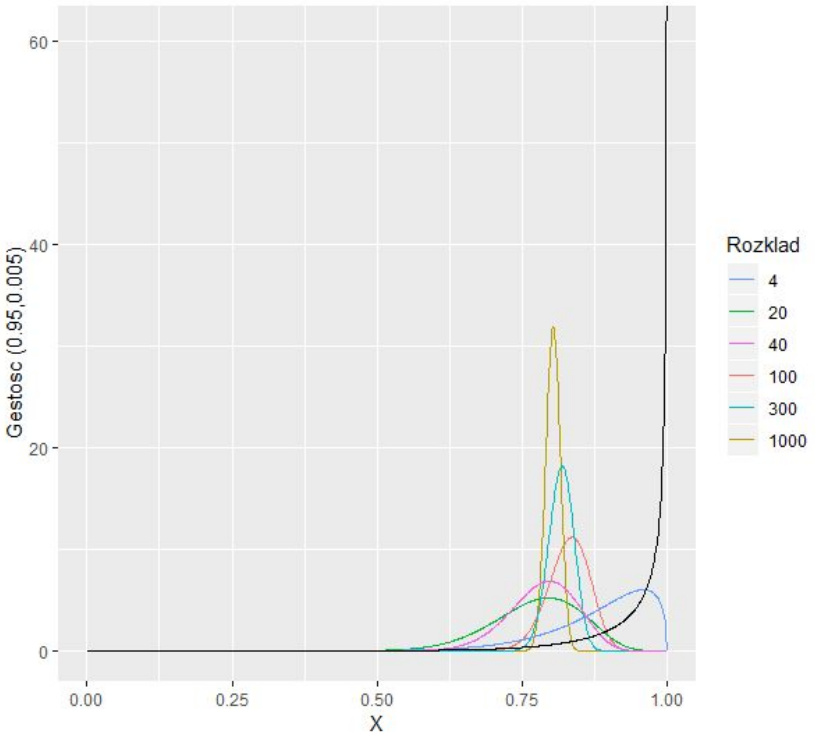
Wykres 6: Rozkłady a priori i a posteriori dla kolejnych danych.

Kolejny rozkład a posteriori ma parametry $EX = 0,5$ oraz $D^2X = 0,01$. Wariancja w stosunku do poprzednio rozpatrywanego rozkładu jest niższa. Można zauważyć, że tutaj wartość oczekiwana stabilizuje się dopiero przy $n = 300$. Pomimo, że ustalona niepewność rozkładu a priori była mniejsza, wartość oczekiwana odbiega od tej rzeczywistej, dlatego duża liczba danych konieczna była aby to skorygować.



Wykres 7: Rozkłady a priori i a posteriori dla kolejnych danych.

Powyższy wykres rozkładu a posteriori ma parametry $EX = 0,95$ oraz $D^2X = 0,03$. Wzrost n powoduje zmianę kształtu rozkładu, wystarczy kilka danych, aby uformować ostateczny kształt, jednak dopiero około sto danych przybliża wartość oczekiwaną rozkładu do wartości 0,8. Mając 1000 danych, rozkład a posteriori formowany jest z dużą pewnością.



Wykres 8: Rozkłady a priori i a posteriori dla kolejnych danych.

Ostatni wykres ma parametry parametry $EX = 0,95$ oraz $D^2X = 0,005$. Wariancja jest mniejsza niż na poprzednim wykresie, a więc również niepewność rozkładu a priori jest mniejsza. Na tym wykresie można zauważyć, że już 20 danych pozwoliło na przybliżenie wartości oczekiwanej rozkładu do 0,8. Spowodowane jest to faktem, że oszacowana wartość oczekiwana rozkładu 0,95 jest bliska wartości rzeczywistej 0,8. Tylko wykresy dla kilku początkowych danych odbiegają od rzeczywistych wartości.

WNIOSKI:

Przeprowadzone badanie jest źródłem cennych informacji na temat wpływu zastosowanych rozkładów a priori oraz liczby danych na ostateczny rozkład a posteriori w rozważaniach bayesowskich. W tym przypadku wzięto pod uwagę

rozkłady beta z parametrami α i β o stałej wartości oczekiwanej oraz zmiennej wariancji. Pozwoliło to na otrzymanie następujących wniosków. Wzrost wariancji w rozkładzie a priori wpływa na większe podobieństwo między oszacowaniami wartości oczekiwanej metodą bayesowską oraz MNW. Większa wariancja rozkładu a priori skutkuje większą wariancją rozkładu a posteriori. Wraz ze wzrostem liczby danych różnice te zanikają, a wariancja dąży do 0. Sytuacja, w której wartość oczekiwana a priori jest daleka od średniej MNW, a wariancja jest niska, wymaga dużej ilości danych, aby rozkład a posteriori był zbliżony do klasycznych oszacowań danych. Natomiast większa wariancja pozwoli w tym przypadku na szybsze przybliżenie się rozkładu a posteriori do otrzymanych danych. Można na tej podstawie stwierdzić, że niska wariancja w rozkładzie a priori może być używana przy dużej pewności na temat rozkładu badanej cechy. Najlepiej, gdy wynika to z wcześniej potwierdzonych informacji.