Centralne Twierdzenie Graniczne od podstaw

Michał Danaj

31 stycznia 2016

Centralne twierdzenie graniczne

Jeśli X_i, dla $i = 1, 2, ..., n$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie, o wartości oczekiwanej μ i skończonej wariancji σ^2, to zmienna losowa o postaci

$$Z = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

zbiega według rozkładu do standardowego rozkładu normalnego gdy n rośnie do nieskończoności.

Lub alternatywnie i bardziej intuicyjnie. Średnia ciągu niezależnych zmiennych losowych o jednakowym rozkładzie, o wartości oczekiwanej μ i skończonej wariancji σ^2, zbiega wg rozkładu do rozkładu normalnego o wartości oczekiwanej μ i odchyleniu standardowym $\frac{\sigma}{\sqrt{n}}$.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} N(\mu, \frac{\sigma}{\sqrt{n}})$$

W dalszej części będę się odnosić do tej postaci twierdzenia.

Wyjaśnienia

W dalszej części zostaną wytłumaczone pojęcia które zostały użyte w twierdzeniu oraz, jak mam nadzieję, uda się je instazobrazować w trochę bardziej namacalny i intuicyjny sposób.

Zmienna losowa

Zmienna losowa to funkcja która z prawdopodobieństwem zadanym przez jej rozkład przyjmuje określone wartości liczbowe. Np. zmienna losowa X_i z rozkładu dwumianowego, przyjmuje wartości {0, 1}:

- 1 z prawdopodobieństwem p,
- 0 z prawdopodobieństwem $1 - p$.

Losując wartość zmiennej losowej, generujemy realizację zmiennej losowej z zadanego rozkładu.

```
p <- 0.5 #będziemy generować rozkład dwumianowy z prawdopodibieństwem p wylosowania jedynki
X1 <- rbinom(1, size=1 ,prob=p)
```

Możemy teraz obejrzeć realizację:

```
X1
```

```
## [1] 0
```

Kolejne losowanie wygeneruje nam kolejną realizację:

```
X1<-rbinom(1, size=1 ,prob=p)
X1
```

```
## [1] 0
```

I jeszcze kilka realizacji dla zmiennej losowej z rozkładu jednostajnego na odcinku (0,1).

```
X1<-runif(1)
X1
```

```
## [1] 0.5728534
```

```
X1<-runif(1)
X1
```

```
## [1] 0.9082078
```

```
X1<-runif(1)
X1
```

```
## [1] 0.2016819
```

Ciąg zmiennych losowych

Mając n zmiennych losowych, możemy mówić o ciągu zmiennych losowych.

Przykładowo, ciąg pięciu zmiennych losowych z rokładu dwumianowego:

```
p <- 0.5 #będziemy określać rozkład dwumianowy z prawdopodibieństwem p wylosowania jedynki
X <- rbinom(5, size=1, prob=p)
names(X) <- c('X1','X2','X3','X4','X5') #nadaję nazwy poszczególnym zmiennym losowym
```

Możemy obejrzeć jedną realizację ciągu zmiennych losowych:

```
## X1 X2 X3 X4 X5
##  1  1  1  1  0
```

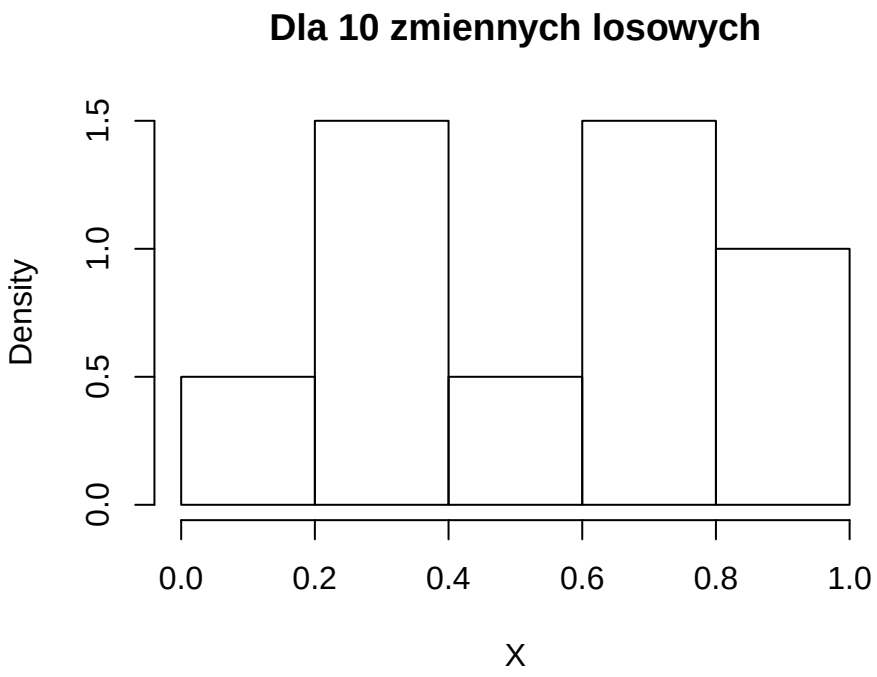
I analogicznie dla rozkładu jednostajnego:

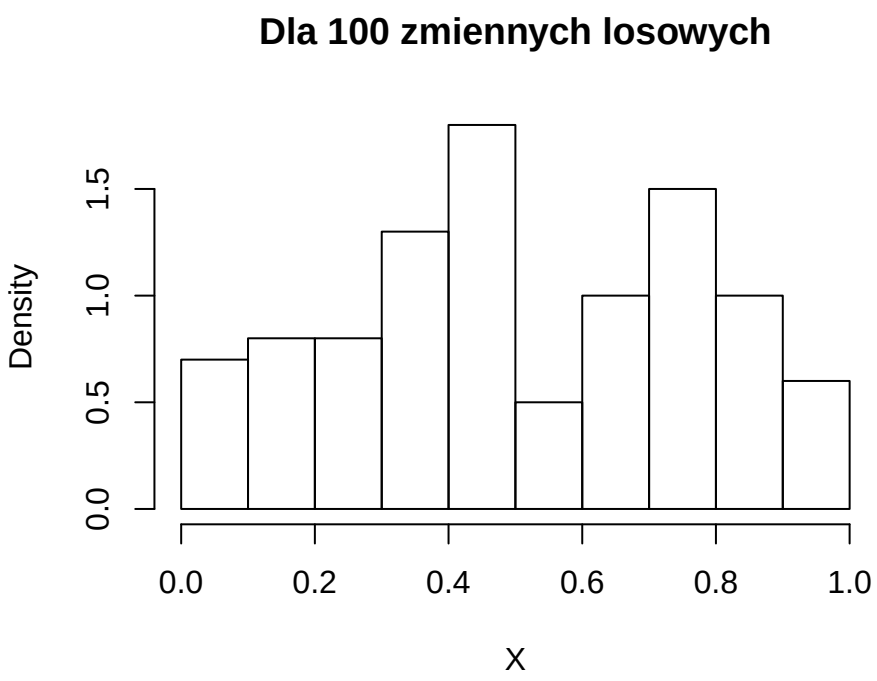
```
X <- runif(5)
names(X) <- c('X1','X2','X3','X4','X5') #nadaję nazwy poszczególnym zmiennym losowym
X
```

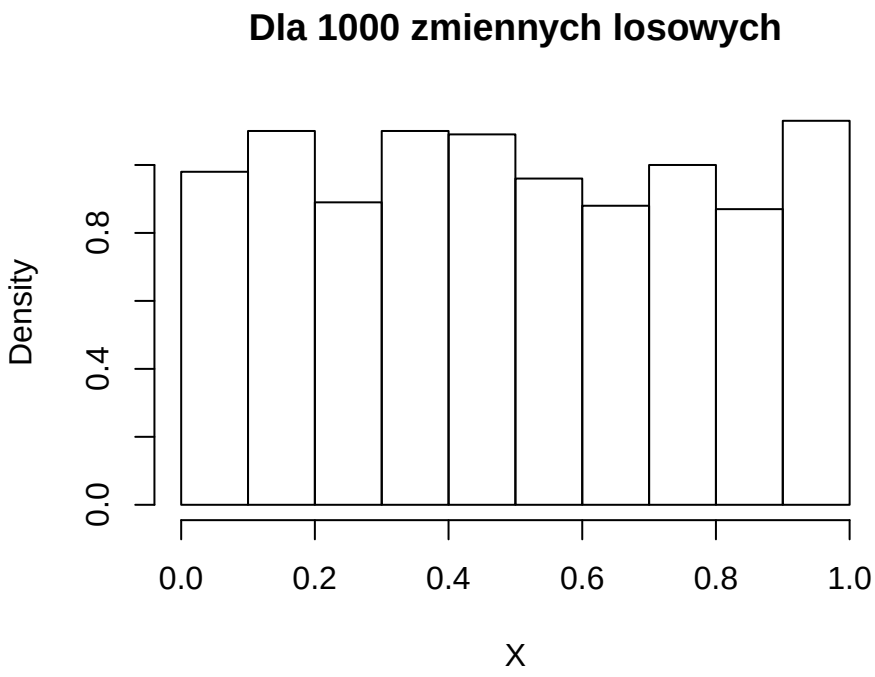
```
##        X1        X2        X3        X4        X5
## 0.2059746 0.1765568 0.6870228 0.3841037 0.7698414
```

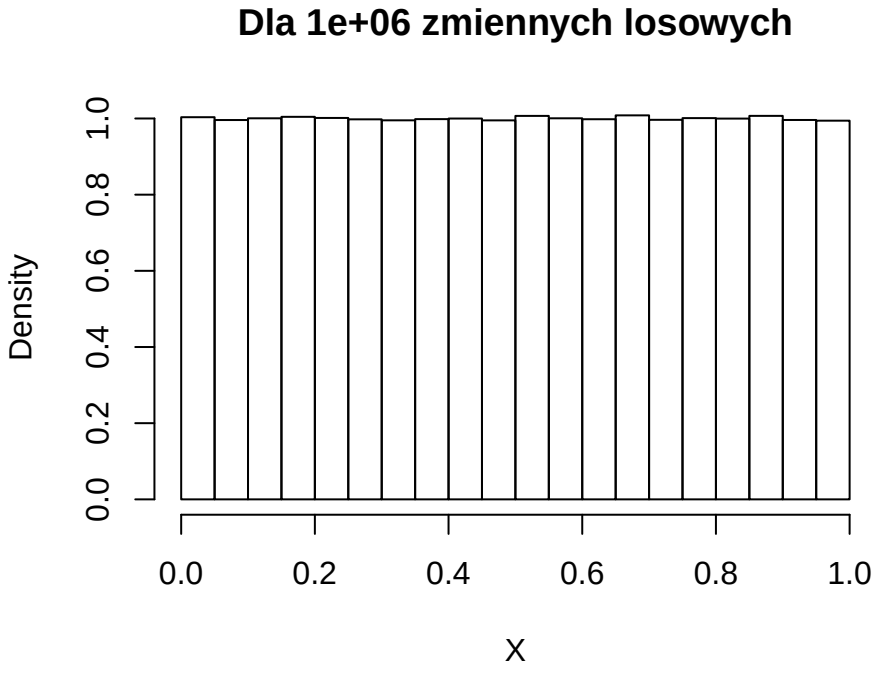
Uwaga. Jak długi by nie był ciąg zmiennych losowych, rozkład zmiennych losowych występujących w ciągu się nie zmienia. Jest cały czas taki, jakim go zdefiniowaliśmy. Oryginalny rozkład do niczego nie zbiega!

Możemy to zaobserwować, rysując histogram z pewnej realizacji ciągu zmiennych losowych. Weźmy ciąg zmiennych losowych z rozkładu jednostajnego na odcinku (0,1). Histogram będzie tak wyskalowany, aby całka z niego wyniosła 1. Zacznijmy od n = 10 i sukcesywnie zwiększajmy n dla kolejnych wykresów:


Dla 10 zmiennych losowych
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
X


Dla 100 zmiennych losowych
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
X





Możemy zaobserwować, że histogram staje się coraz bardziej wygładzony wraz ze wzrostem długości ciągu zmiennych losowych. Wygląda to tak, jakby jednak coś do czegoś zbiegało. I faktycznie, histogram jest estymatorem gęstości rozkładu prawdopodobieństwa z którego zostały wygenerowane zmienne X_i, i w naszym przypadku zbiega do rozkładu jednostajnego na odcinku (0,1). Wyraźnie to widać na ostatnim wykresie.

Średnia z ciągu zmiennych losowych

Na podstawie ciągu zmiennych losowych, jesteśmy w stanie generować pochodne zmienne losowe i badać ich własności. Oczywistymi wielkościami, które moglibyśmy badać są powszechnie używane statystyki, jak średnia, wariancja, min, max, itp.

Rozważając Centralne Twierdzenie Graniczne, wielkością której chcielibyśmy się przyjrzeć jest średnia. Policzmy zatem dwie najczęściej rozpatrywane wielkości rozkładu, czyli wartość oczekiwaną oraz wariancję, udowadniając tym samym (o ile się uda) część CTG.

Dla wartości oczekiwanej będziemy mieli:

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) =$$

Stałą można wyciągnąć przed wartość oczekiwaną, a wartość oczekiwana sumy to suma wartości oczekiwanych. Możemy zatem z wartością oczekiwaną wejść pod znak sumy. Wartość oczekiwana każej zmiennej losowej z ciągu ma wartość oczekiwaną μ, a zatem ostatecznie otrzymamy

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

Analogiczne dla wariancji

$$\mathbb{V}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}X_i = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Pokazaliśmy zatem, że rozkład średniej ma wartośc oczekiwaną i odchylenie standardowe takie jak w CTG. Zainteresowanych dowodem że rozkład średniej zbiega do rozkładu normalnego odesyłam do literatury.

Zwrócę tutaj uwagę na jedną rzecz, która może nie jest konieczna do zrozumienia CTG, ale myślę że warto o niej wspomnieć. Patrząc na postać wariancji średniej, widać że będzie się ona zmieniać wraz ze zmianą n. Wzrost n, powoduje zmniejszenie wariancji.

Wynik ten jest zgodny z intuicją. Chcąc policzyć średnią wagę 5-cio latków w Polsce, wybieramy z całej populacji pewną próbę n 5-cio latków i w oparciu o tę próbę wyliczamy średnią. Przy dwóch pięciolatkach czujemy, że nasza średnia dość mocno może odbiegać od rzeczywistej średniej - stąd duża wariancja. Przy 100 będziemy już bardziej przekonani co do zgodności z rzeczywistą średnią, by przy 10 tys. z dużą pewności stwierdzić że jesteśmy blisko rzeczywistej średniej na populacji, co oznacza że nasza średnia ma małą wariancję.

CTG na przykładach

Zobrazujmy sobie teraz działanie twierdzenia na przykładach.

Rozkład jednostajny Weźmy ciąg niezależnych zmiennych losowych o rozkładzie jednostajnym na odcinku (0,1). Aby pokazać rozkład średniej (zależny od n jak pokazał poprzedni punkt) z tego ciągu, musimy wygenerować ciąg zmiennych losowych Z_j będących średnimi.

$$Z_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}$$

gdzie $i = 1, 2, ..., n$ oraz $j = 1, 2, ..., N$ oraz $X_{i,j}$ jest ciągiem niezlaeżnych zmiennych losowych z rozkładu jednostajnego na odcinu (0,1).

Przyjrzyjmy się przykładowym realizacjom zmiennych losowych $X_{i,j}$ oraz Z_j.

```r
N<-5 #liczba średnich
n <- 2 #liczba zmiennych losowych
X <- matrix(runif(N*n), nrow=n, ncol=N) #wygenerowane zmienne z rozkładu jednostajnego
Z <- colMeans(X) #wyliczam ciąg średnich

X
```

```
##           [,1]       [,2]      [,3]      [,4]      [,5]
## [1,] 0.1249295 0.27368702 0.6987412 0.4076154 0.6407130
## [2,] 0.9423007 0.01745772 0.2311964 0.3892212 0.5894214
```

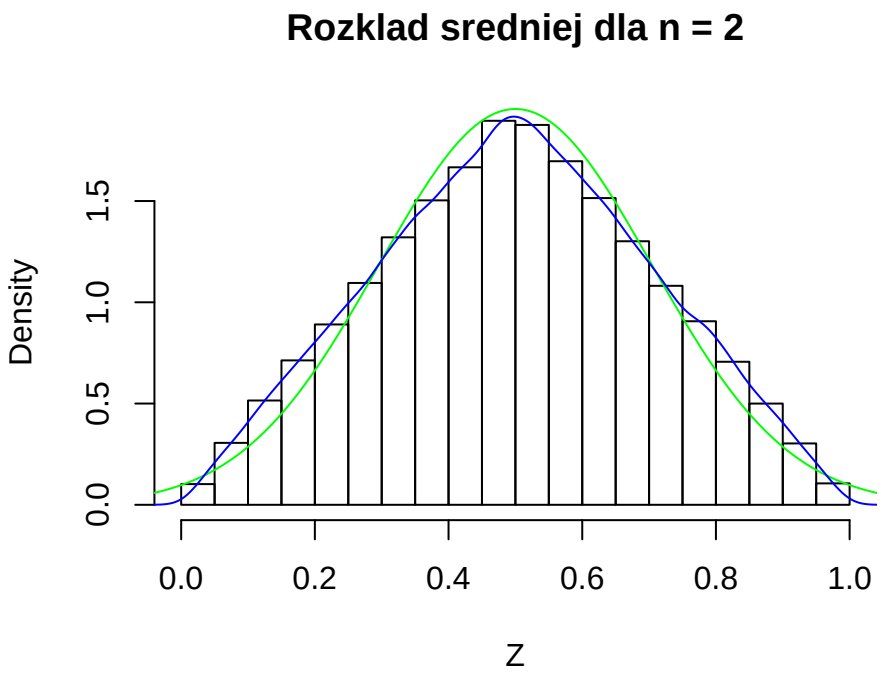
```r
Z
```

```
## [1] 0.5336151 0.1455724 0.4649688 0.3984183 0.6150672
```

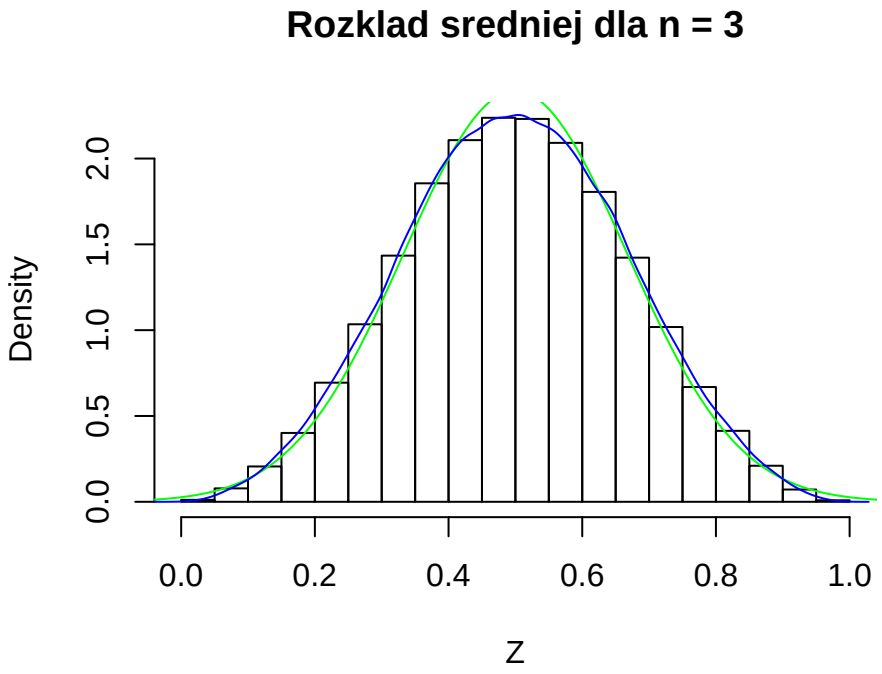
Zwiększmy liczbę średnich, aby móc wyrysować histogram. Pozostańmy przy dwóch zmiennych losowych. Dodatkowo dodajmy do histogramu gęstość rozkładu normalnego wyznaczonego na podstawie CTG (kolor zielony) oraz nieparametryczną estymację gęstości (kolor niebieski). Gdyby rozkład średnich faktycznie był rozkładem normalnym o parametrach określonych w CTG, niebieski wykres pokryłby się z wykresem zielonym. Jeśli wykresy nie pokrywają się, oznacza to że rozkład średniej **nie jest rozkładem normalnym**.

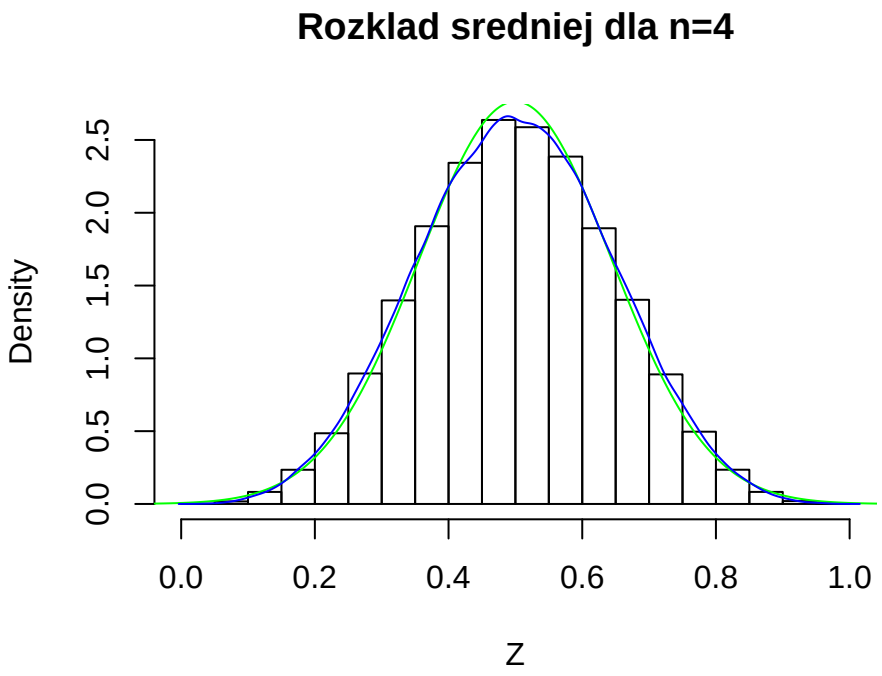
```r
N<-100000 #liczba średnich
n <- 2 #liczba zmiennych losowych
X <- matrix(runif(N*n), nrow=n, ncol=N) #wygenerowane zmienne z rozkładu jednostajnego
Z <- colMeans(X) #wyliczam ciąg średnich
hist(Z, prob=TRUE, main=paste("Rozkład średniej dla n =",n))

x <- seq(-0.1, 1.1, length=1000)
hx <- dnorm(x, mean=0.5, sd=sqrt(1/12/n))
lines(x,hx, col='green')
lines(density(Z), col='blue')
```

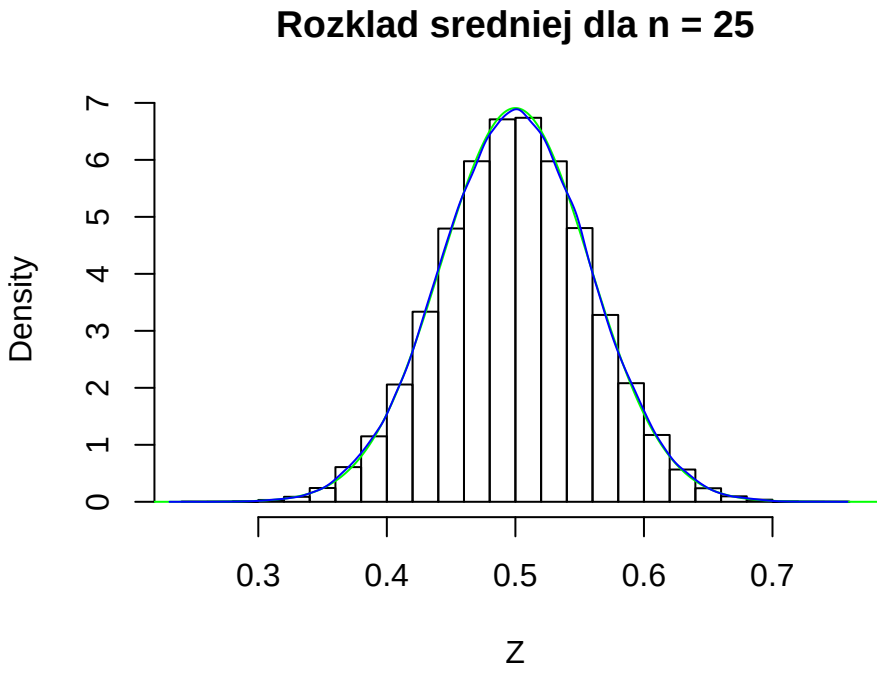


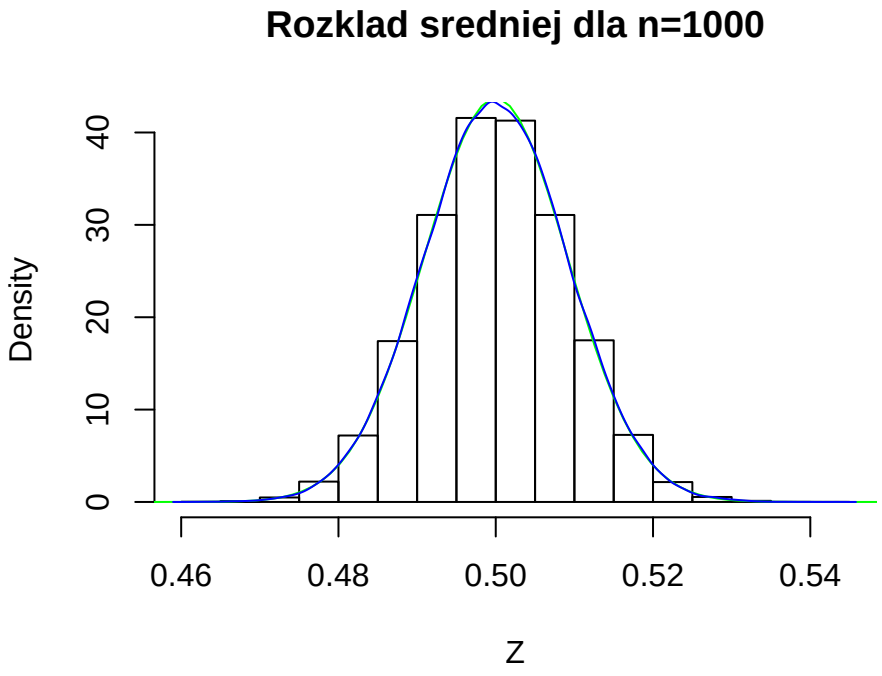
Jak widać, już dla $n = 2$ rozkład średniej znacznie odbiega od rozkładu jednostajnego, z którego były wylosowane wartości zmiennych losowych w ciągu. Porównując jednak wykresy gęstości, widać że gęstość rozkładu teoretycznego (zielony) znacznie różni się od gęstości wyestymowanej z danych (niebieski). Spójrzmy jak to będzie się przedstawiać dla kolejnych n. Zwróćmy jeszcze uwagę, że na kolejnych wykresach zmieniają się skale na osiach. Zawężanie skali na osi OX wskazuje na zmniejszającą się wariancję średniej.





Gdzieś kiedyś widziałem, że przyjmuje się że średnia z 25-ciu zmiennych losowych z rozkładu jednostajnego ma rozkład normalny.





Możemy jeszcze potwierdzić normalność ostatniego rozkładu średniej testem statystycznym. Interpretację pozostawiam już Tobie.

```
  shapiro.test(Z[1:5000])
```

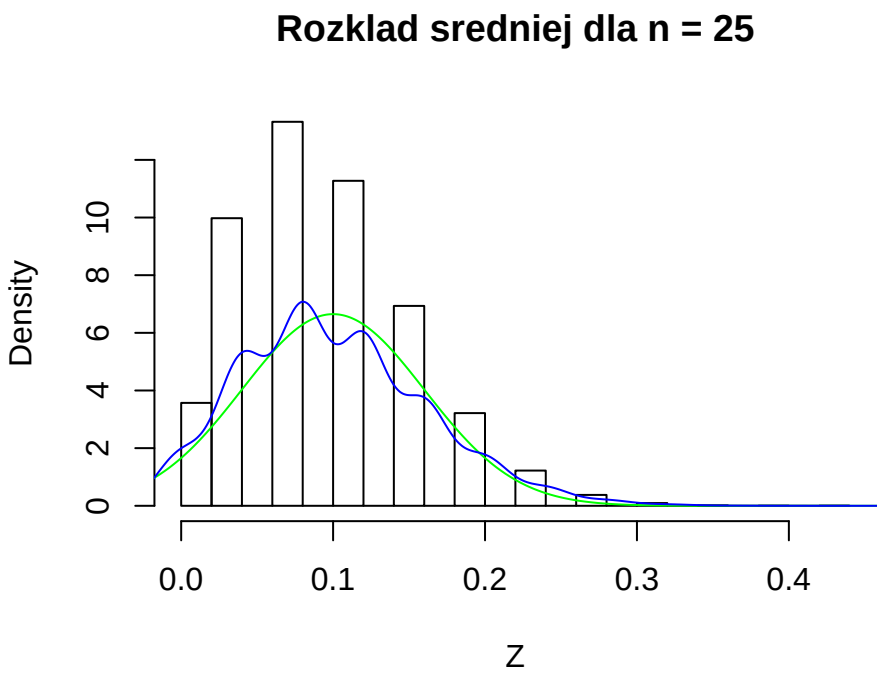
```
##
##  Shapiro-Wilk normality test
##
## data:  Z[1:5000]
## W = 0.9998, p-value = 0.8734
```

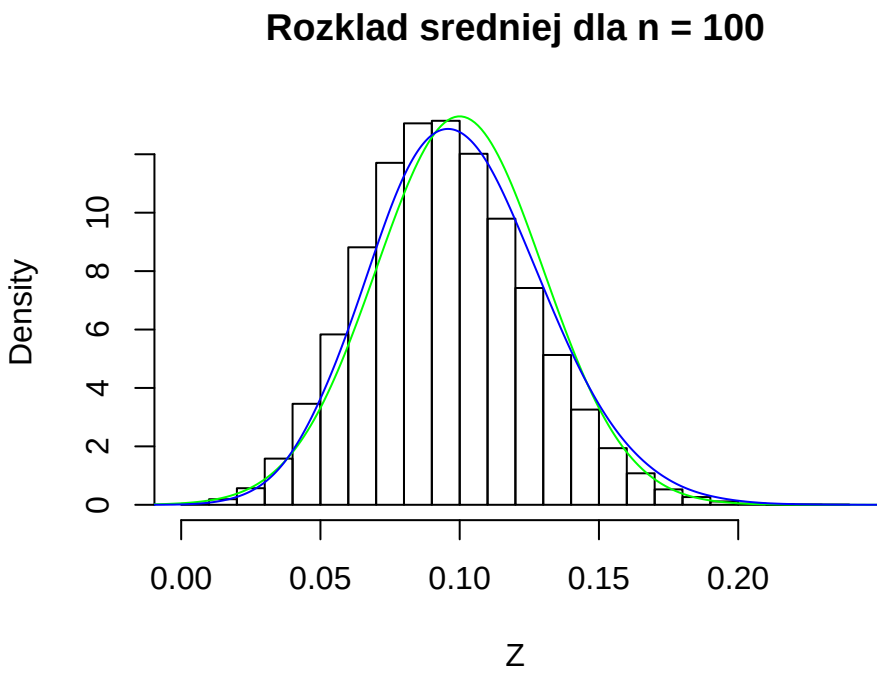
Rozkład dwumianowy Intuicyjnie zapewne czujemy, że trudniej będzie osiągnąć rozkład normalny średniej z rozkładu dwumianowego. Zamiast całego spektrum wartości jakie może osiągnąć zmienna losowa w przypadku rozkładu jednostajneto, tutaj są możliwe tylko dwie wartości. Co więcej, zbieżność do rozkładu normalnego jest zależna od parametru p. Im p jest dalej od $1/2$, tym zbieżność będzie słabsza. Przyjrzyjmy się zatem przykładom dla $p = 0.1$. O ile w rozkładzie jednostajnym praktycznie kończyliśmy przy 25 zmiennych losowych, tutuj proponuję od nich rozpocząć.

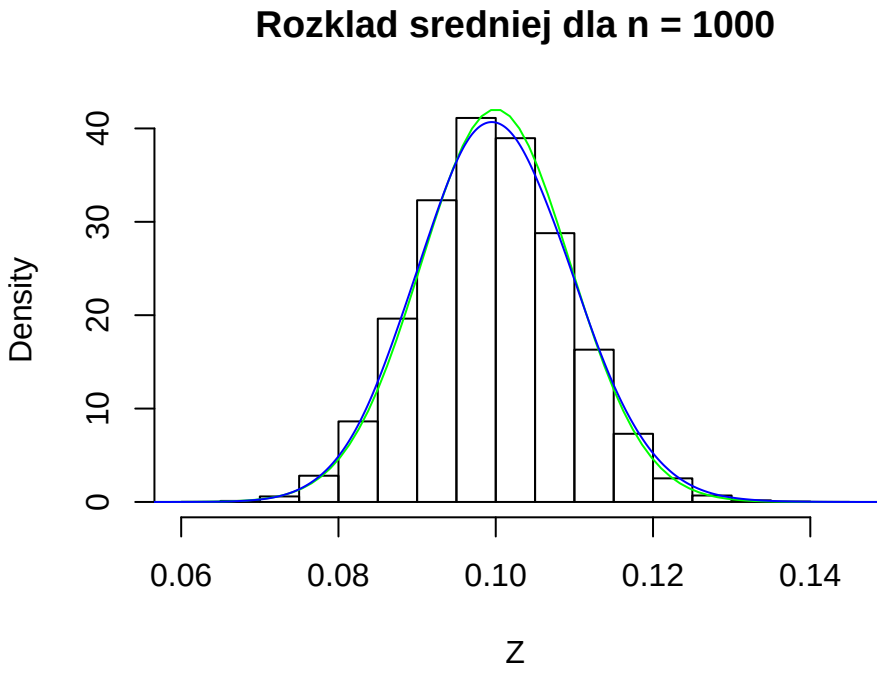
```
N<-100000 #liczba średnich
n <- 25 #liczba zmiennych losowych
p <- 0.1

X <- matrix(rbinom(N*n, size=1, p=p), nrow=n, ncol=N) #wygenerowane zmienne z rozkładu dwumianowego
Z <- colMeans(X)
hist(Z, prob=TRUE, main=paste("Rozkład średniej dla n =",n))


x <- seq(-0.1, 1.1, length=1000)
hx <- dnorm(x, mean=0.1, sd=sqrt(p*(1-p)/n))
lines(x,hx, col='green')
lines(density(Z, adjust=3), col='blue')
```







Jak widać, w przypadku rozkładu dwumianowego nawet dla liczności $n = 1000$ i $p = 0.1$, rozkład wyestymowany na podstawie danych w widoczny sposób odbiega od rozkładu normalnego.

Zbieżność

W poprzednim punkcie widzieliśmy, że fakt zbieżności asymptotycznej nie mówi nam nic o prędkości tej zbieżności. Zbieżność asymptotyczna niekoniecznie musi pociągać za sobą zbieżność w naszym konkretnym zagadnieniu. Do podjęcia decyzji czy faktycznie możemy skorzystać z CTG, możemy podejść na dwa sposoby. Jako najszybszy i najpowszechniej stosowany, to oparcie się o ogólnie przyjęte kryteria np. że dla rozkładu dwumianowego możemy skorzystać z CTG gdy ..... Jako ćwiczenie proponuję sprawdzić, czy faktycznie powyższe kreterium pozwala nam osiągnąć zadowalającą zbieżność rozkładu dla różnych p.

Jeśli jednak chcielibyśmy podejść do tematu bardziej świadomie, możemy zrobić symulację Monte Carlo, czyli dokładnie takie symulacje jak powyżej aby stwierdzić, czy rozkład średniej jest rozkładem normalnym. Oczywiście symulację taką możemy wykonać tylko w przypadku, gdy znamy rozkład początkowy. Jeśli natomiast dysponujemy próbą, to rozkładu tego nie znamy i w zasadzie znać nie musimy. Możemy się w takim przypadku oprzeć o metody bootstrap.

Uwaga 1 Pamiętajmy, że stosując test statystyczny o poziomie istotności np. 5% do określenia czy zasymulowany rozkład średniej jest rozkładem normalnym, w 5% symulacji test pokaże że nie jest nawet jeśli byłby!

Uwaga 2 Nawet jeśli wszystkie znaki na ziemi i niebie wskazują, że rozkład średniej w Twoim przypadku nie jest rozkładem normalnym, to różnice w rozkładach mogą być tak niewielkie, że z praktycznego punktu widzenia w rozpatrywanym przez Ciebie zagadnieniu są one pomijalne. Jeśli faktycznie tak jest, nic nie stoi na przeszkodzie (poza ewentualnie narzuconymi z góry wymogami), aby podjąć decyzję (świadomą!) o zastosowaniu przyliżenia normalnego.