ADPS 2020Z — Laboratorium 1 (rozwiązania)

Jakub Adamowicz

Zadanie 1

Treść zadania

Dla wybranych dwóch spółek

- sporządź wykresy procentowych zmian kursów zamknięcia w zależności od daty,
- wykreśl i porównaj histogramy procentowych zmian kursów zamknięcia,
- wykonaj jeden wspólny rysunek z wykresami pudełkowymi zmian kursów zamknięcia.

Rozwiązanie

Rozpakowanie danych

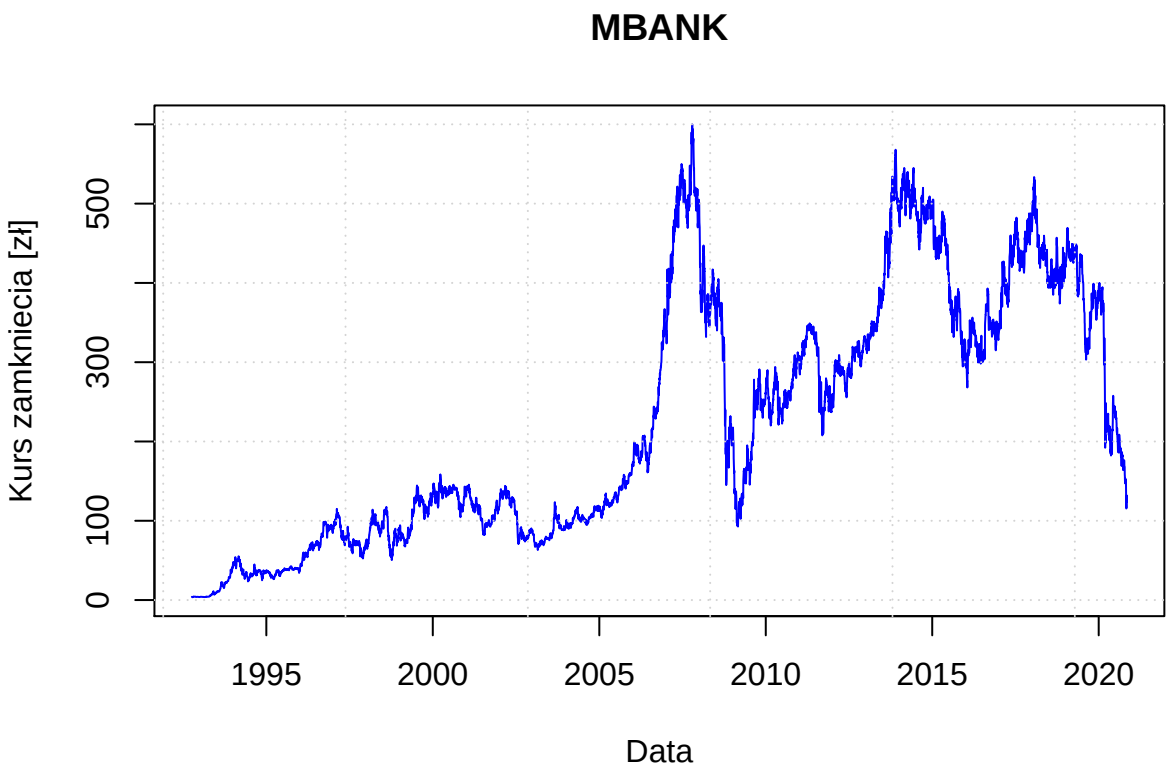
```
unzip('mstall.zip', 'MBANK.mst')
unzip('mstall.zip', 'PEKAO.mst')
```

Wczytanie danych i zmiana nazw kolumn

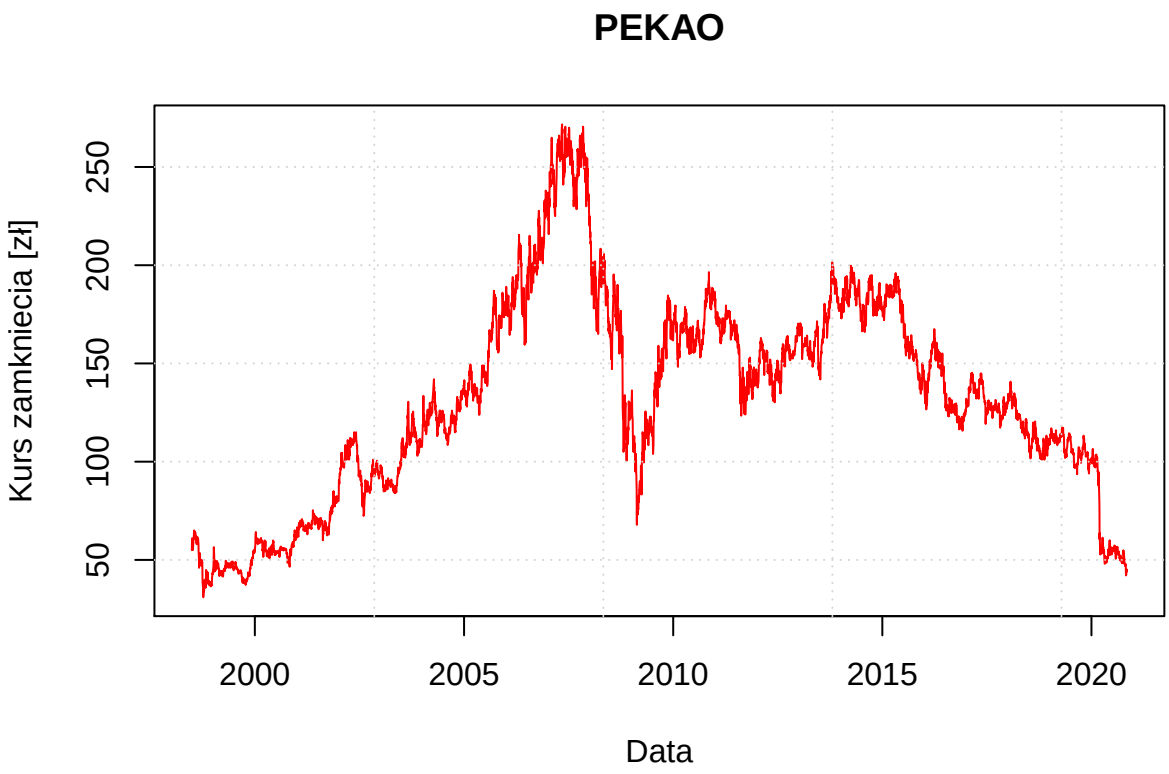
```
df_MBANK = read.csv('MBANK.mst')
df_PEKAO = read.csv('PEKAO.mst')

names(df_MBANK) = c('ticker', 'date', 'open', 'high', 'low', 'close','vol')
names(df_PEKAO) = c('ticker', 'date', 'open', 'high', 'low', 'close','vol')
```

Wykres kursu zamkniecia w zależności od daty

```
df_MBANK$date = as.Date.character(df_MBANK$date, format ='%Y%m%d')
df_PEKAO$date = as.Date.character(df_PEKAO$date, format ='%Y%m%d')
plot(close ~ date, df_MBANK, type = 'l', col = 'blue',
      xlab = 'Data', ylab = 'Kurs zamkniecia [zł]', main = 'MBANK' )
grid()
```

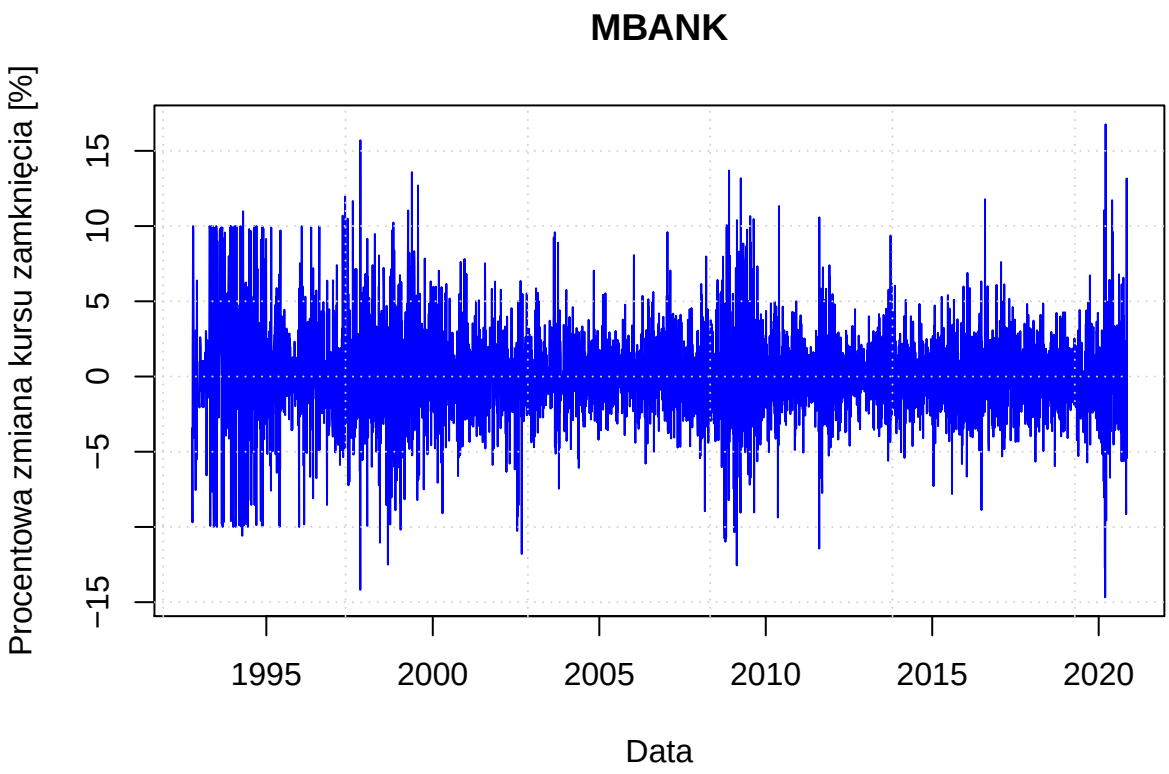



```
plot(close ~ date, df_PEKAO, type = 'l', col = 'red',
     xlab = 'Data', ylab = 'Kurs zamkniecia [zł]', main = 'PEKAO')
grid()
```

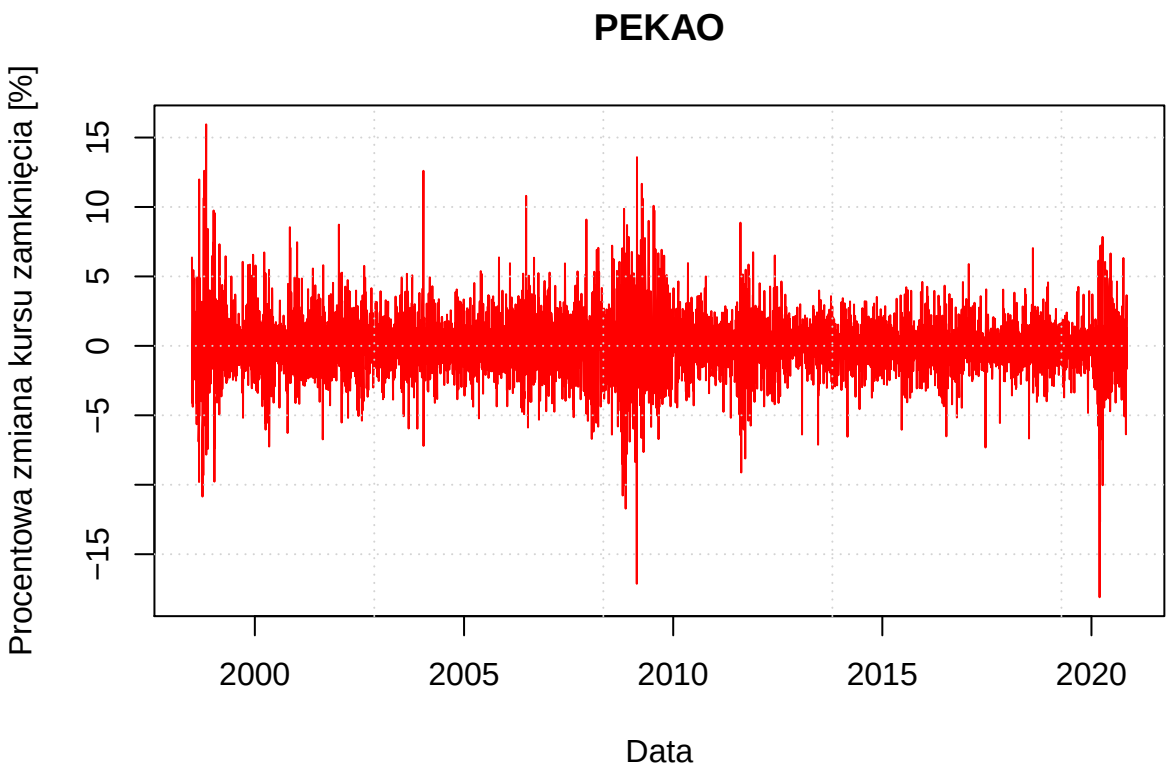


Wykres procentowych zmian kursu zamknięcia

```
df_MBANK$close_ch= with(df_MBANK, c(NA, 100*diff(close)/open[-length(close)]))
plot(close_ch ~ date, df_MBANK, type = 'l', col = 'blue', xlab = 'Data',
     ylab = 'Procentowa zmiana kursu zamknięcia [%]', main = 'MBANK' )
grid()
```




```
df_PEKAO$close_ch= with(df_PEKAO, c(NA, 100*diff(close)/open[-length(close)]))
plot(close_ch ~ date, df_PEKAO, type = 'l', col = 'red', xlab = 'Data',
     ylab = 'Procentowa zmiana kursu zamknięcia [%]', main = 'PEKAO' )
grid()
```



Obliczenie funkcji gęstości prawdopodobieństwa

```
mb = mean(df_MBANK$close_ch, na.rm = T)
sb = sd(df_MBANK$close_ch, na.rm = T)
```

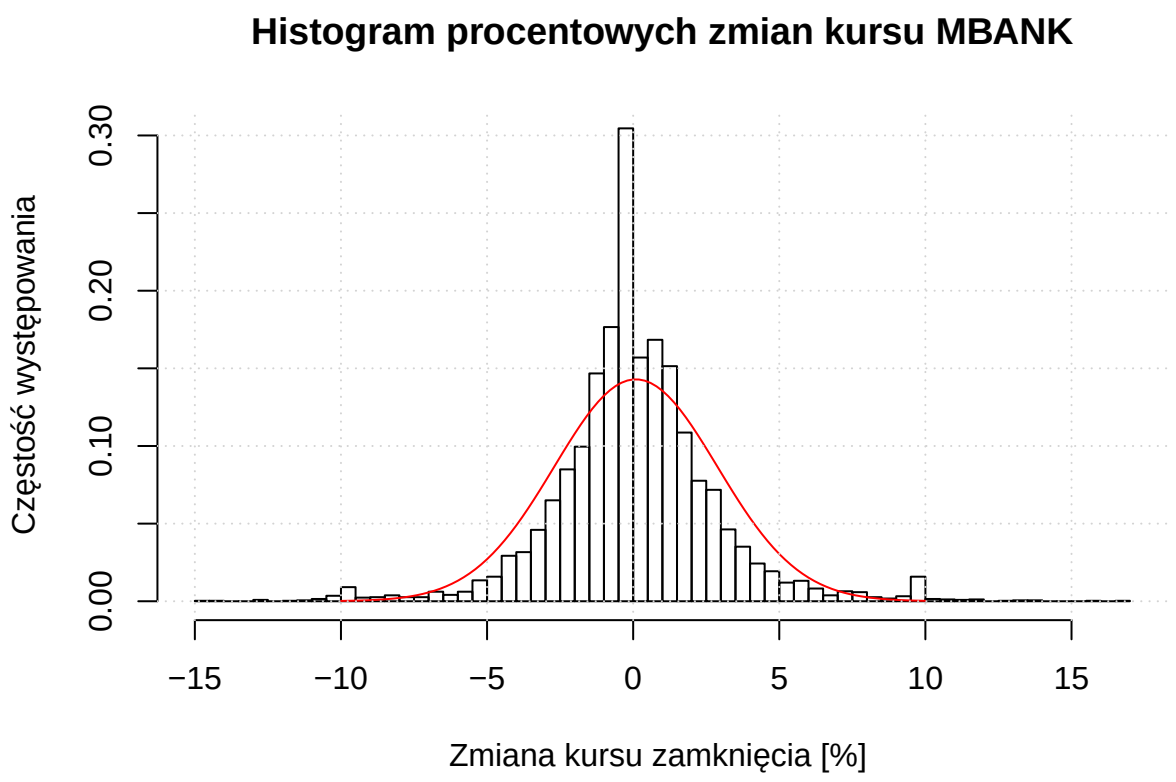
Wartość średnia zmian kursu zamknięcia MBANK wynosi 0.0907, a odchylenie standardowe 2.7913.

```
mp = mean(df_PEKAO$close_ch, na.rm = T)
sp = sd(df_PEKAO$close_ch, na.rm = T)
```

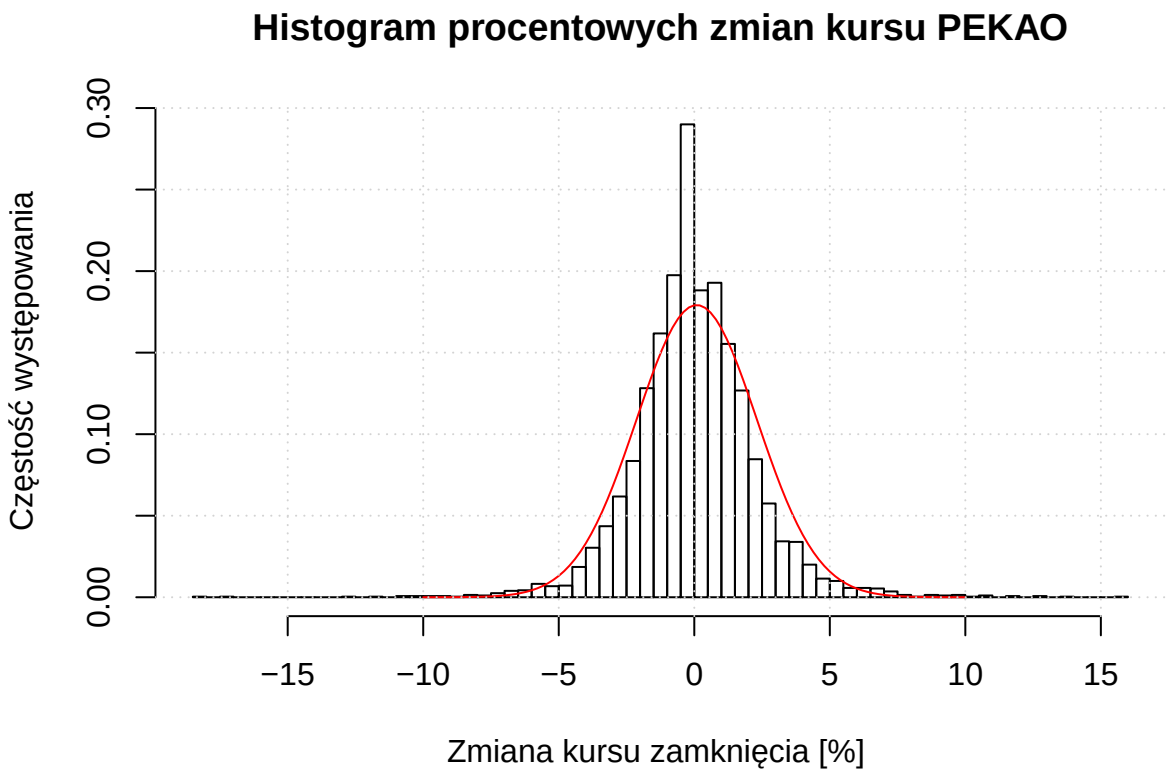
Wartość średnia zmian kursu zamknięcia PEKAO wynosi 0.0211, a odchylenie standardowe 2.2262.

Histogram procentowych zmian kursu zamknięcia

```
hist(df_MBANK$close_ch, breaks = 50, prob = T,
xlab = 'Zmiana kursu zamknięcia [%] ',
ylab = 'Częstość występowania',
main = 'Histogram procentowych zmian kursu MBANK' )
curve(dnorm(x, mean = mb, s = sb), add = T, col = 'red', -10, 10)
grid()
```



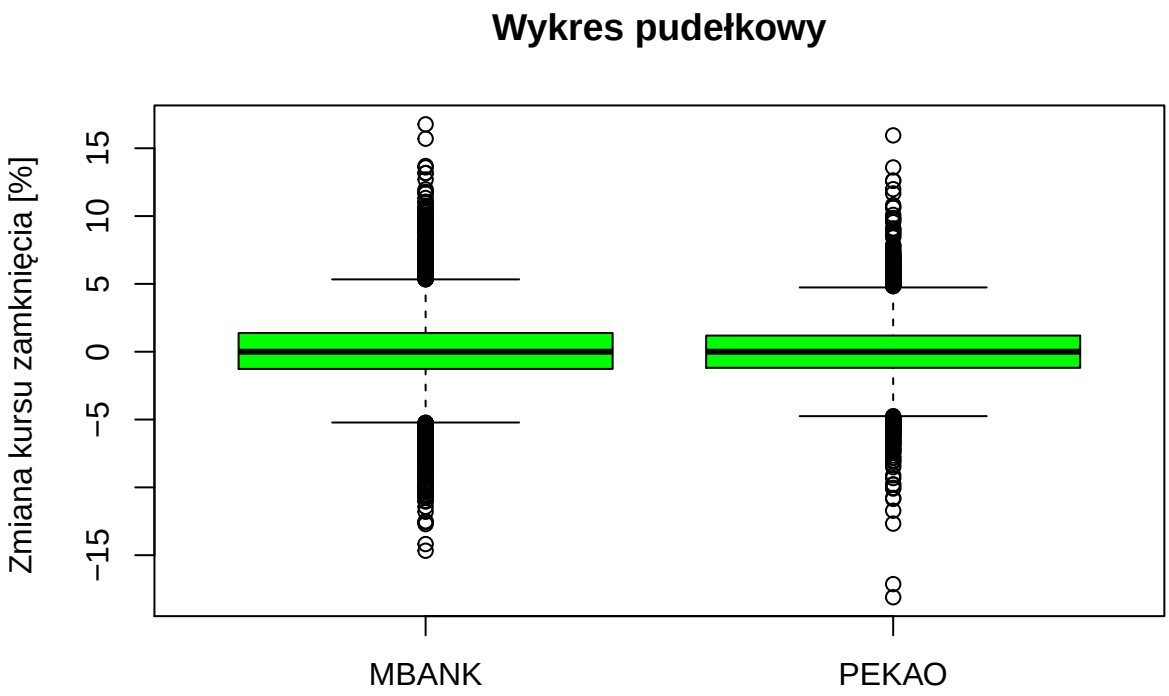

```r
hist(df_PEKAO$close_ch, breaks = 50, prob = T,
xlab = 'Zmiana kursu zamknięcia [%] ',
ylab = 'Częstość występowania',
main = 'Histogram procentowych zmian kursu PEKAO' )
curve(dnorm(x, mean = mb, s = sp), add = T, col = 'red', -10, 10)
grid()
```



Histogramy zmian procentowych spółek MBANK i PEKAO mają bardzo zbliżoną postać. Zarówno histogram spółki MBANK i PEKAO wydaje się być rozkładem normalnym. Cechą odróżniajacą oba histogramy jest zwiększona częstość dla spółki MBANK przy zmianie procentowej około -10% i 10% co nie występuje dla spółki PEKAO.

Wykres pudełkowy obu spółek

```
dwie_spolki = rbind(df_MBANK, df_PEKAO)
boxplot(close_ch ~ ticker, dwie_spolki,
  col = 'green',
  xlab = '', ylab = 'Zmiana kursu zamknięcia [%] ',
  main = 'Wykres pudełkowy' )
```



Zadanie 2

Treść zadania

1. Sporządź wykres liczby katastrof lotniczych w poszczególnych:

- miesiącach,
- dniach,
- dniach tygodnia (weekdays()).

2. Narysuj jak w kolejnych latach zmieniały się:

- liczba osób, które przeżyły katastrofy,
- odsetek osób (w procentach), które przeżyły katastrofy.

Rozwiązanie

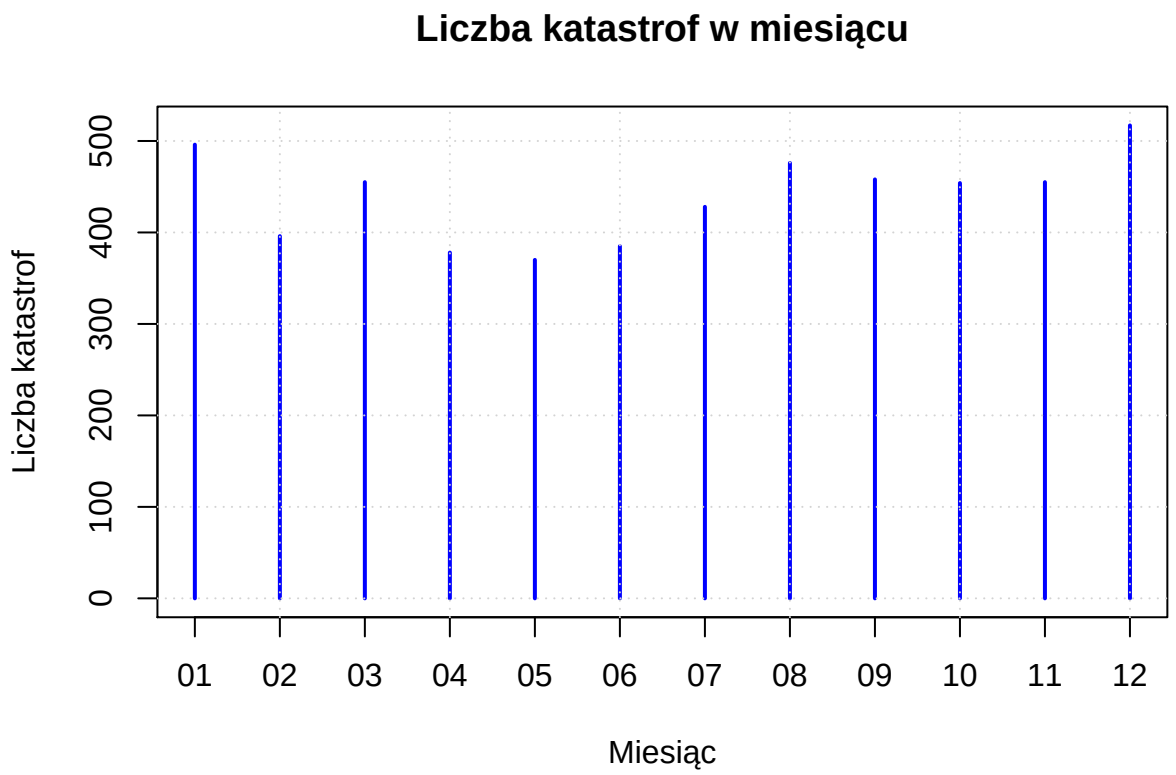
Załadowanie danych i stworzenie kolumn z dniem miesiąca, miesiącem i dniem tygodnia

```
cat =read.csv('crashes.csv')
```

```
cat$Month = strftime(as.Date(cat$Date, '%m/%d/%Y'), '%m')
cat$Day = strftime(as.Date(cat$Date, '%m/%d/%Y'), '%d')
cat$Year = strftime(as.Date(cat$Date, '%m/%d/%Y'), '%Y')
cat$Weekday = weekdays(as.Date(cat$Date,'%m/%d/%Y'))
```

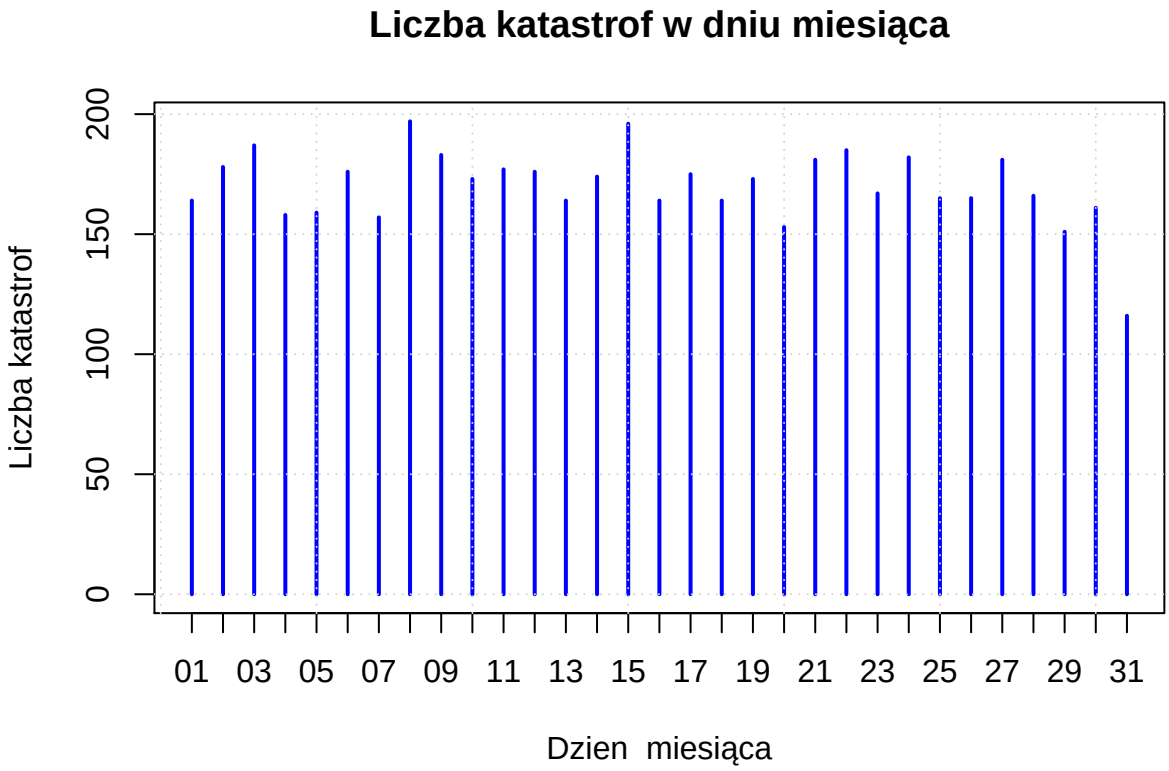
Wykres liczby wypadków w danym miesiącu

```
plot(table(cat$Month), type = 'h', col = 'blue', xlab = 'Miesiąc',
ylab = 'Liczba katastrof', main = 'Liczba katastrof w miesiącu' )
grid()
```



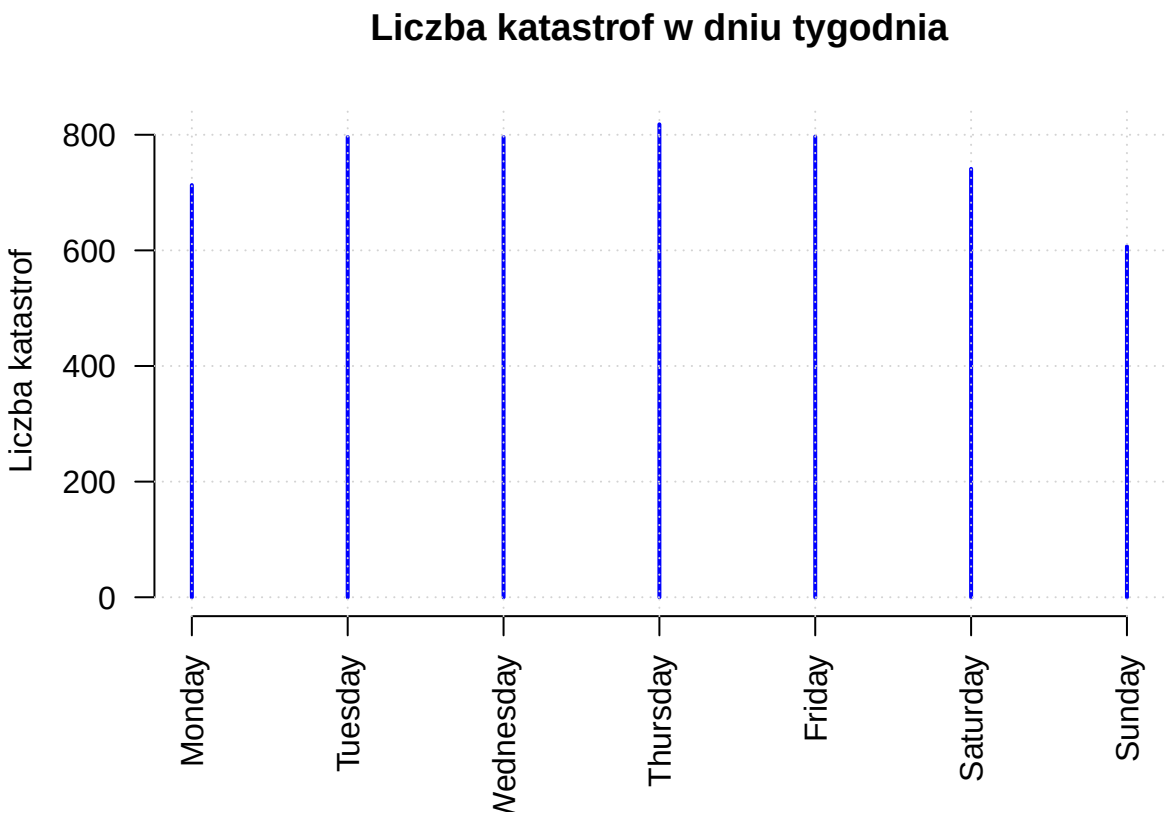
Wykres liczby wypadków w danym dniu miesiąca

```
plot(table(cat$Day), type = 'h', col = 'blue', xlab = 'Dzien  miesiąca',
ylab = 'Liczba katastrof', main = 'Liczba katastrof w dniu miesiąca' )
grid()
```



Wykres liczby wypadków w danym dniu tygodnia

```
x1  = factor(cat$Weekday, levels=c("Monday", "Tuesday", "Wednesday","Thursday","Friday","Saturday","Sun
plot(table(x1), type = 'h', col = 'blue', xlab = '',
ylab = 'Liczba katastrof', main = 'Liczba katastrof w dniu tygodnia',las=2 )
grid()
```

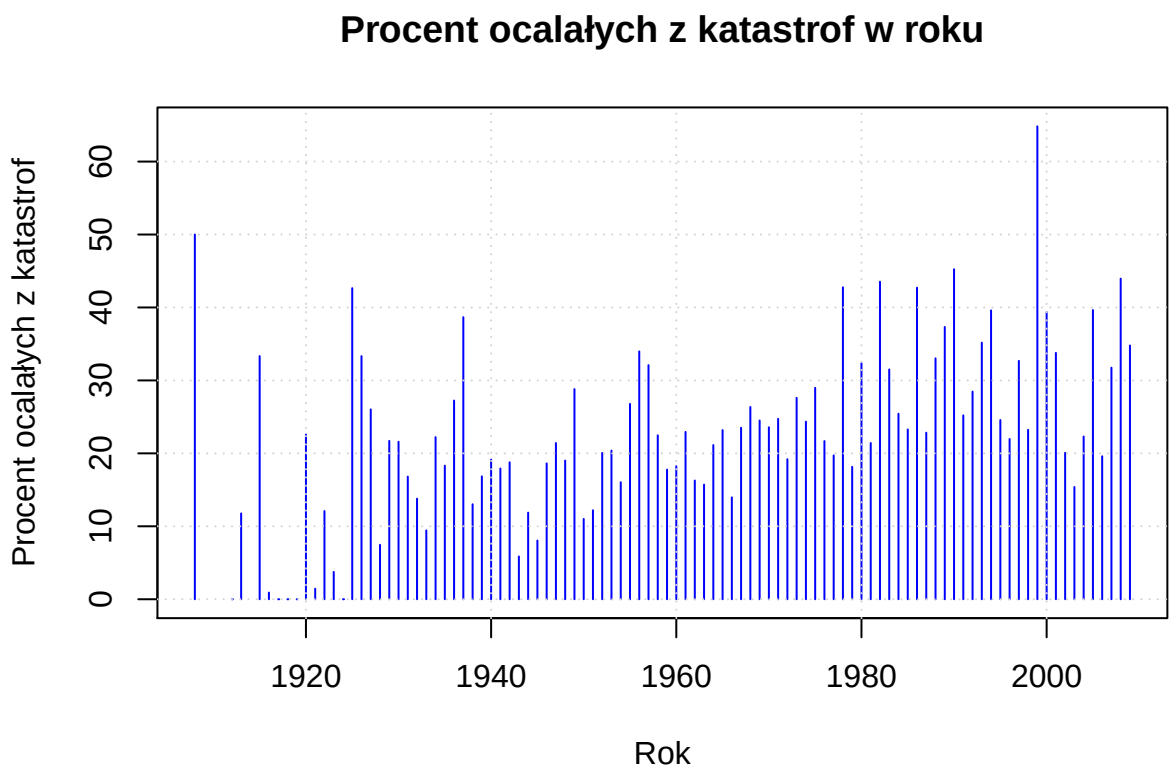


Agregacja danych po latach

```
cat$Surv = cat$Aboard - cat$Fatalities
Surv_agr = aggregate(Surv ~ Year, cat, FUN = sum)
All_agr = aggregate(Aboard ~ Year,cat,FUN = sum)
All_agr$Surv_agr = aggregate(Surv ~ Year, cat, FUN = sum)[, c(2)]
All_agr$Surv_perc = (All_agr$Surv_agr / All_agr$Aboard) * 100
```

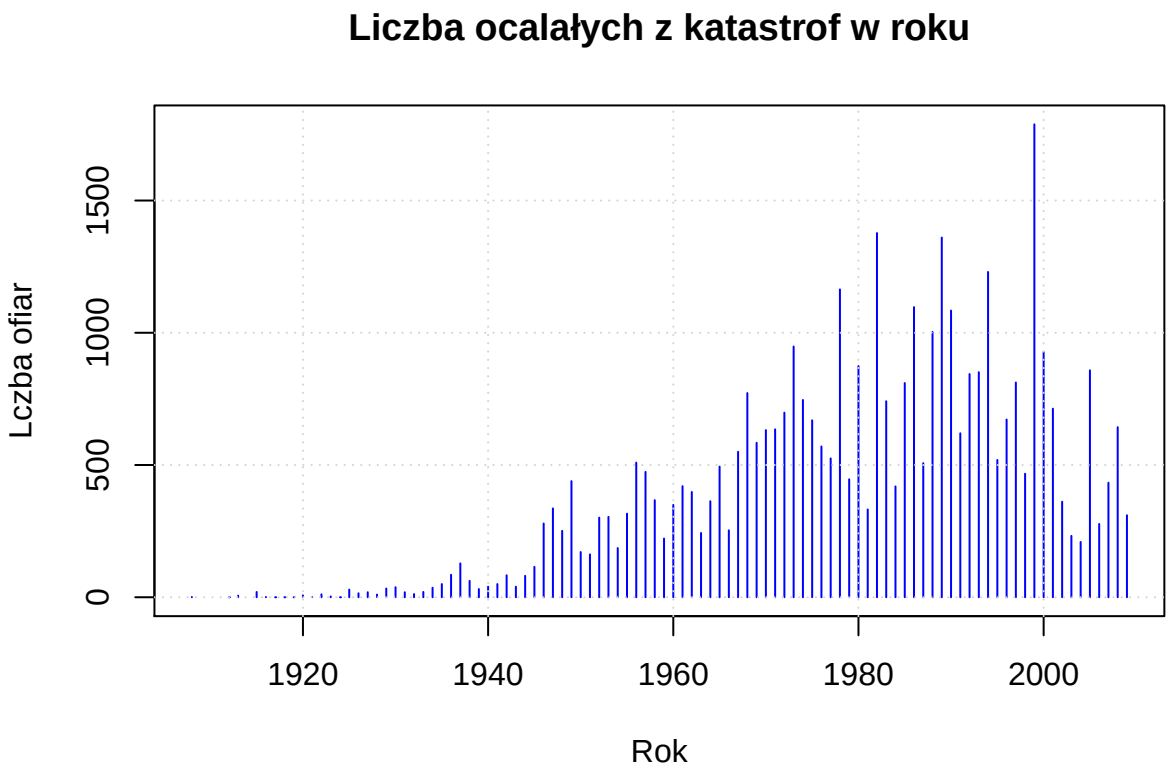
Wykres procentu ocalałych z katastrof

```
plot(All_agr$Surv_perc ~All_agr$Year, type = 'h', col = 'blue',xlab = 'Rok',
     ylab = 'Procent ocalałych z katastrof', main = 'Procent ocalałych z katastrof w roku' )
grid()
```



Wykres liczby osób ocalałych z katastrof

```
plot(Surv_agr, type = 'h', col = 'blue',xlab = 'Rok',
     ylab = 'Lczba ofiar', main = 'Liczba ocalałych z katastrof w roku' )
grid()
```



Zadanie 3

Treść zadania

1. Dla dwóch różnych zestawów parametrów rozkładu dwumianowego (rbinom):

- Binom(20,0.2)
- Binom(20,0.8)

wygeneruj próby losowe składające się z M = 1000 próbek i narysuj wartości wygenerowanych danych.

2. Dla poszczególnych rozkładów (zestawów parametrów) narysuj na jednym rysunku empiryczne i teoretyczne (dbinom) funkcje prawdopodobieństwa, a na drugim rysunku empiryczne i teoretyczne (pbinom) dystrybuanty. W obu przypadkach wyskaluj oś odciętych od 0 do 20.

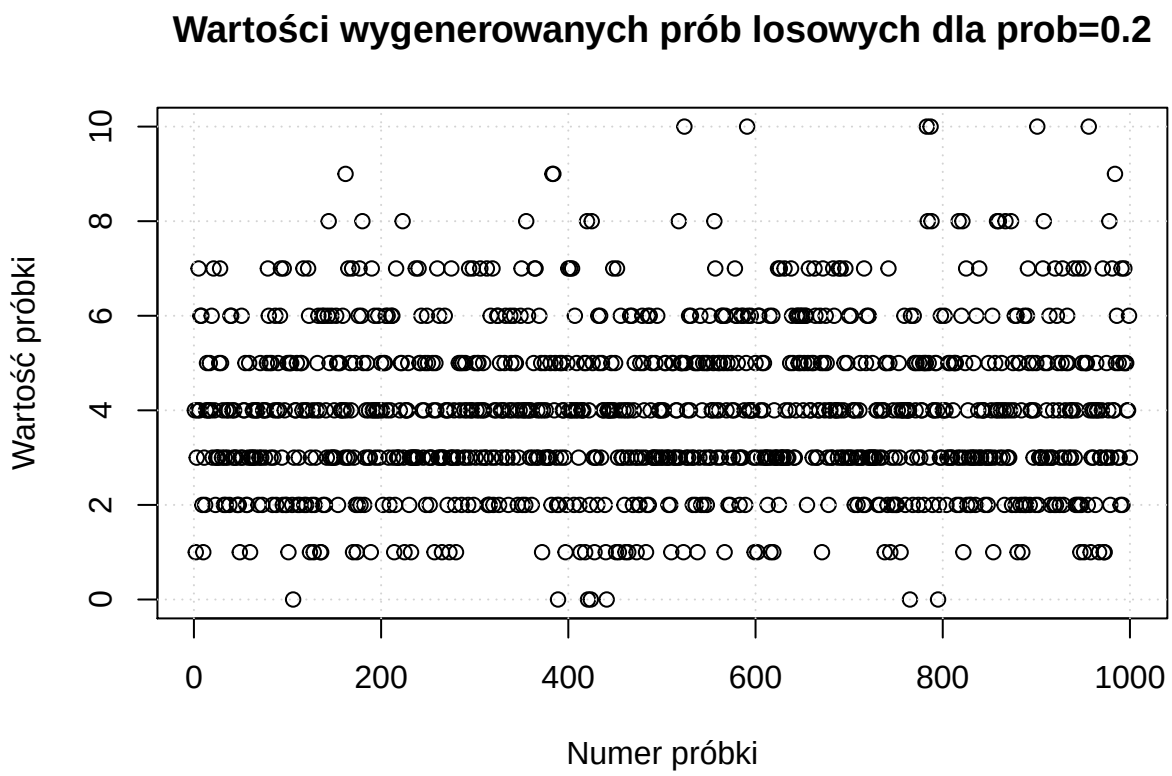
Rozwiązanie

Wygenerowanie prób losowych

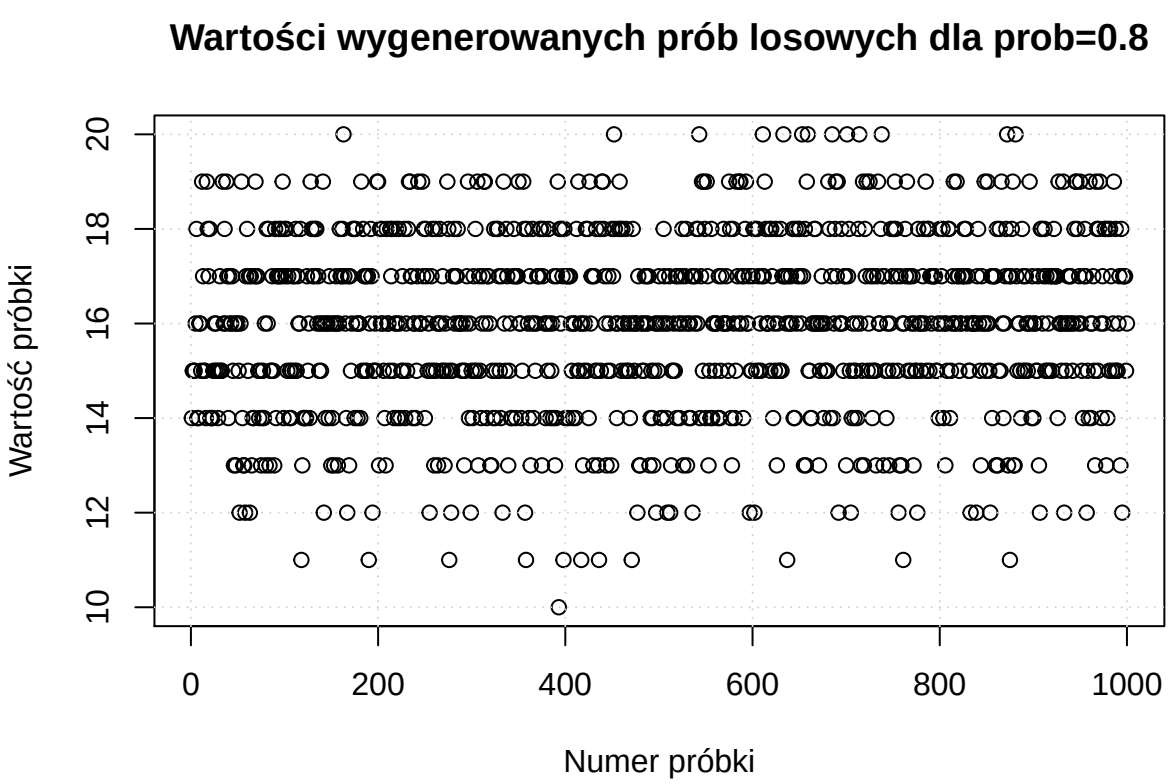
```
M=1000
proba02 = rbinom(M,20,0.2)
proba08 = rbinom(M,20,0.8)
```

Wykres wygenerowanych prób losowych

```
plot(proba02, xlab = 'Numer próbki', ylab = 'Wartość próbki',
     main = 'Wartości wygenerowanych prób losowych dla prob=0.2')
grid()
```

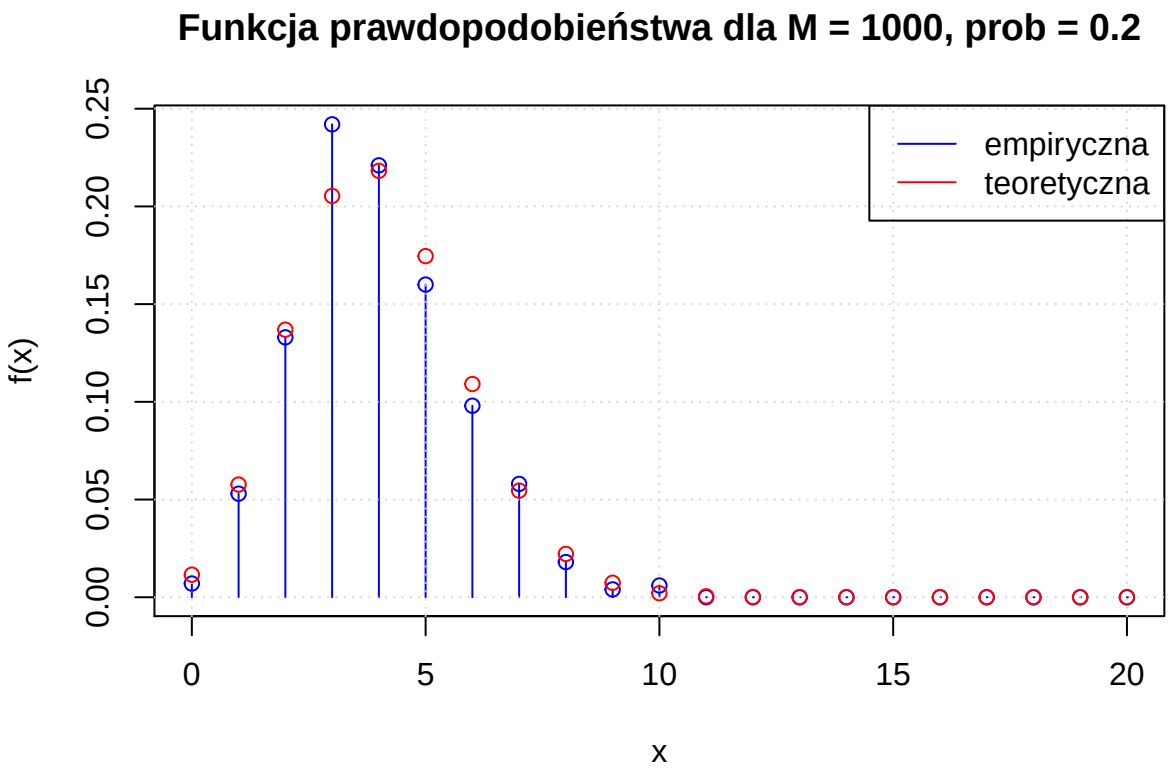



```
plot(proba08, xlab = 'Numer próbki', ylab = 'Wartość próbki',
     main = 'Wartości wygenerowanych prób losowych dla prob=0.8')
grid()
```



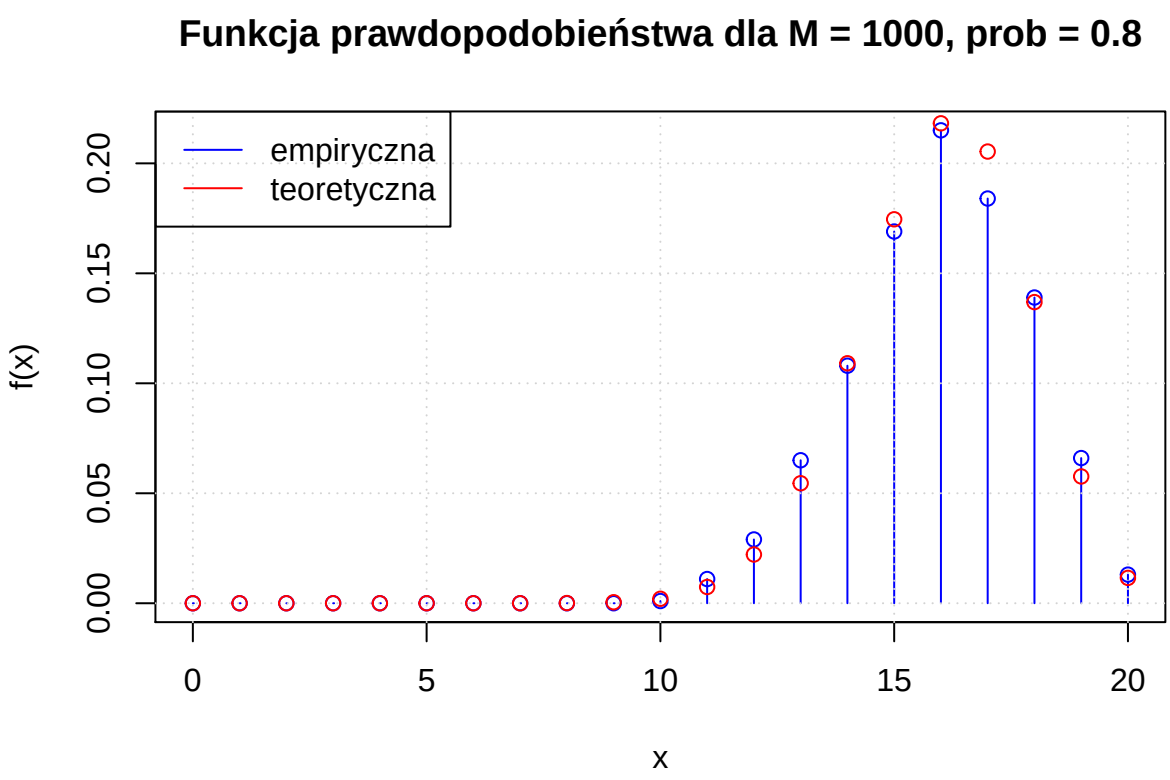
Empiryczna funkcja prawdopodobieństwa dla prob = 0.2

```
Arg02 = 0:20
teor02 = dbinom(Arg02,20,0.2)
Freq02 = as.numeric(table(factor(proba02, levels = Arg02))) / M
plot(Freq02 ~ Arg02, type = 'h', col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = paste0('Funkcja prawdopodobieństwa dla M = ', M, ', prob = 0.2'))
grid()
points(Freq02 ~ Arg02, col = 'blue')

points(teor02 ~ Arg02, col = 'red')
legend('topright', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



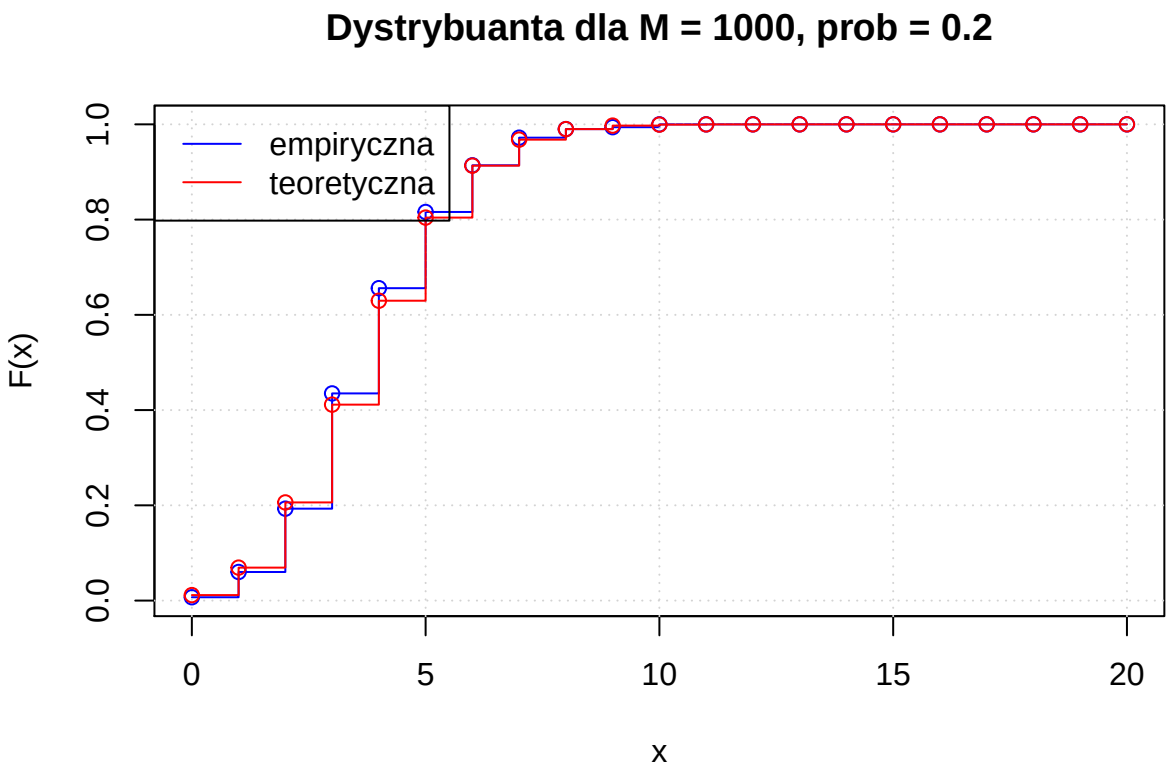
Empiryczna funkcja prawdopodobieństwa dla prob = 0.8

```
Arg08 = 0:max(proba08)
teor08 = dbinom(Arg08,20,0.8)
Freq08 = as.numeric(table(factor(proba08, levels = Arg08))) / M
plot(Freq08 ~ Arg08, type = 'h', col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = paste0('Funkcja prawdopodobieństwa dla M = ', M, ', prob = 0.8'))
grid()
points(Freq08 ~ Arg08, col = 'blue')

points(teor08 ~ Arg08, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



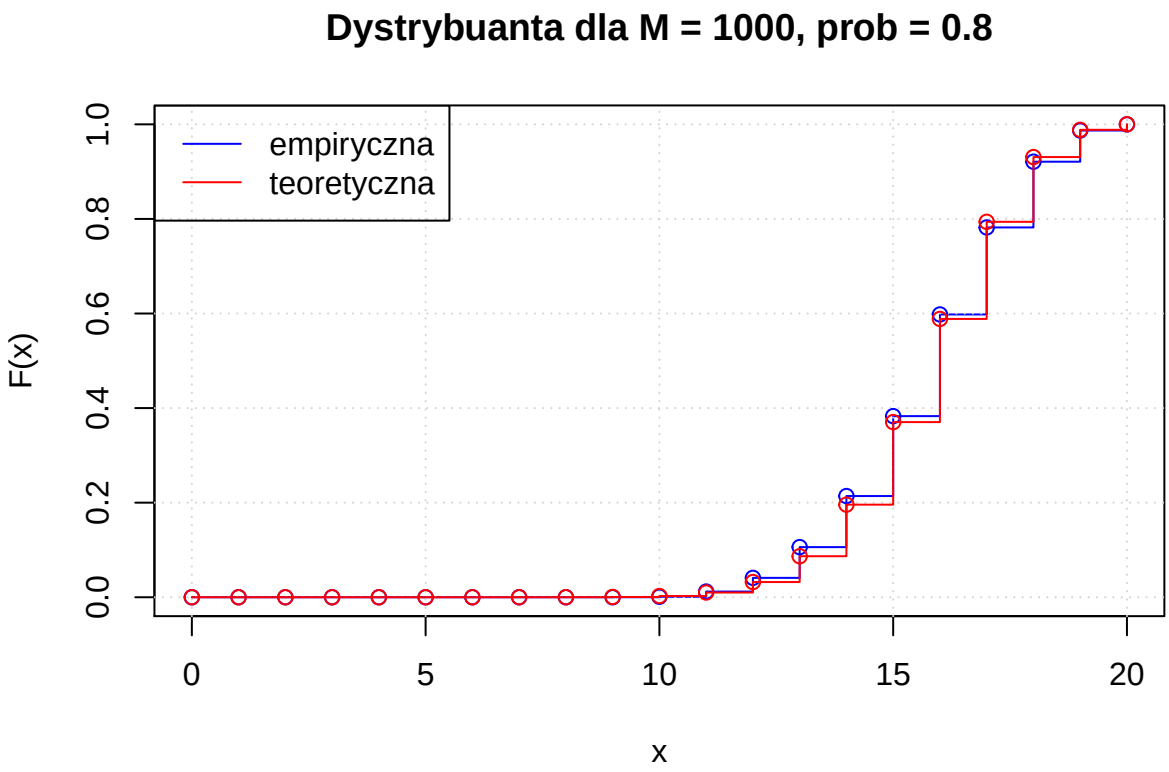
Dystrybuanta empiryczna i teoretyczna dla prob = 0.2

```
plot(cumsum(Freq02) ~ Arg02, type = 's', col = 'blue',
    xlab = 'x', ylab = 'F(x)', main = paste0('Dystrybuanta dla M = ', M,', prob = 0.2'))
grid()
points(cumsum(Freq02) ~ Arg02, col = 'blue')

dist02 = pbinom(Arg02,20,0.2)
lines(dist02 ~ Arg02, type = 's', col = 'red',
      xlab = 'x', ylab = 'F(x)')
points(dist02 ~ Arg02, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



Dystrybuanta empiryczna i teoretyczna dla prob = 0.8

```
plot(cumsum(Freq08) ~ Arg08, type = 's', col = 'blue',
    xlab = 'x', ylab = 'F(x)', main = paste0('Dystrybuanta dla M = ', M,', prob = 0.8'))
grid()
points(cumsum(Freq08) ~ Arg08, col = 'blue')

dist08 = pbinom(Arg02,20,0.8)
lines(dist08 ~ Arg08, type = 's', col = 'red',
      xlab = 'x', ylab = 'F(x)')
points(dist08 ~ Arg08, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



Zadanie 4

Treść zadania

1. Dla rozkładu dwumianowego Binom(20, 0.8) wygeneruj trzy próby losowe składające się z M = 100, 1000 i 10000 próbek.
2. Dla poszczególnych prób wykreśl empiryczne i teoretyczne funkcje prawdopodobieństwa, a także empiryczne i teoretyczne dystrybuanty.
3. We wszystkich przypadkach oblicz empiryczne wartości średnie i wariancje. Porównaj je z wartościami teoretycznymi dla rozkładu Binom(20, 0.8).

Rozwiązanie

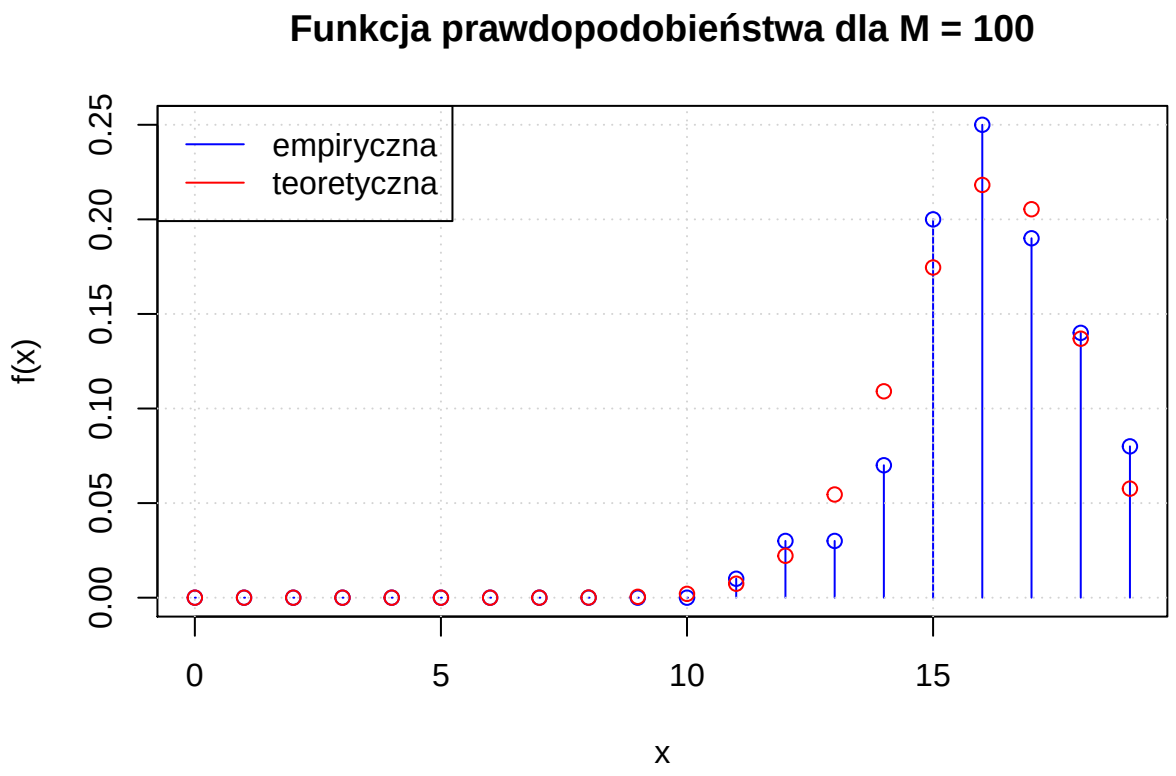
Wygenerowanie trzech prób losowych i teoretycznej funkcji prawdopodobieństwa i teoretycznej dystrybuanty

```
M1 = 100
M2 = 1000
M3 = 10000

prob1 = rbinom(M1, 20,0.8)
prob2 = rbinom(M2, 20,0.8)
prob3 = rbinom(M3, 20,0.8)
```

```r
Arg = 0:max(prob1)
teor = dbinom(Arg,20,0.8)
dist = pbinom(Arg,20,0.8)
```

Wykreślenie empirycznej i teoretycznej funkcji prawdopodobieństwa dla M = 100

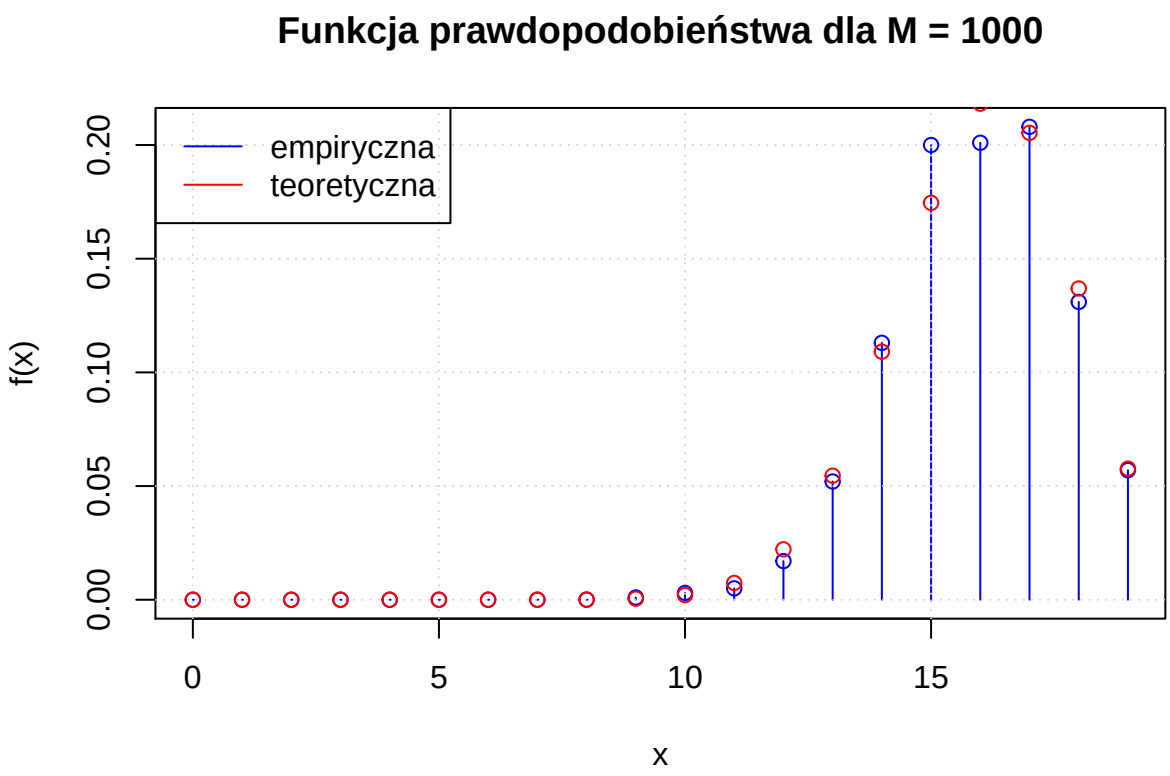
```r
Freq1 = as.numeric(table(factor(prob1, levels = Arg))) / M1
plot(Freq1 ~ Arg, type = 'h', col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = paste0('Funkcja prawdopodobieństwa dla M = ', M1))
grid()
points(Freq1 ~ Arg, col = 'blue')

points(teor ~ Arg, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



Wykreślenie empirycznej i teoretycznej funkcji prawdopodobieństwa dla M = 1000

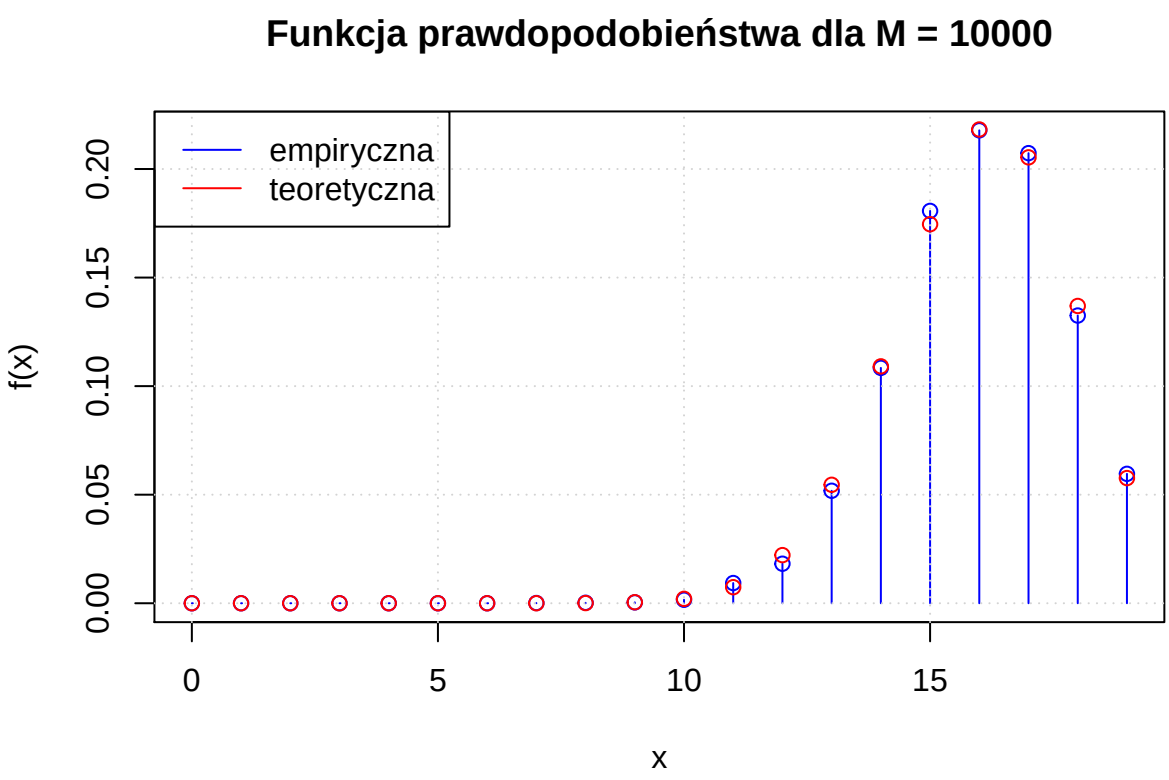
```r
Freq2 = as.numeric(table(factor(prob2, levels = Arg))) / M2
plot(Freq2 ~ Arg, type = 'h', col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = paste0('Funkcja prawdopodobieństwa dla M = ', M2))
grid()
points(Freq2 ~ Arg, col = 'blue')
```

```
points(teor ~ Arg, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



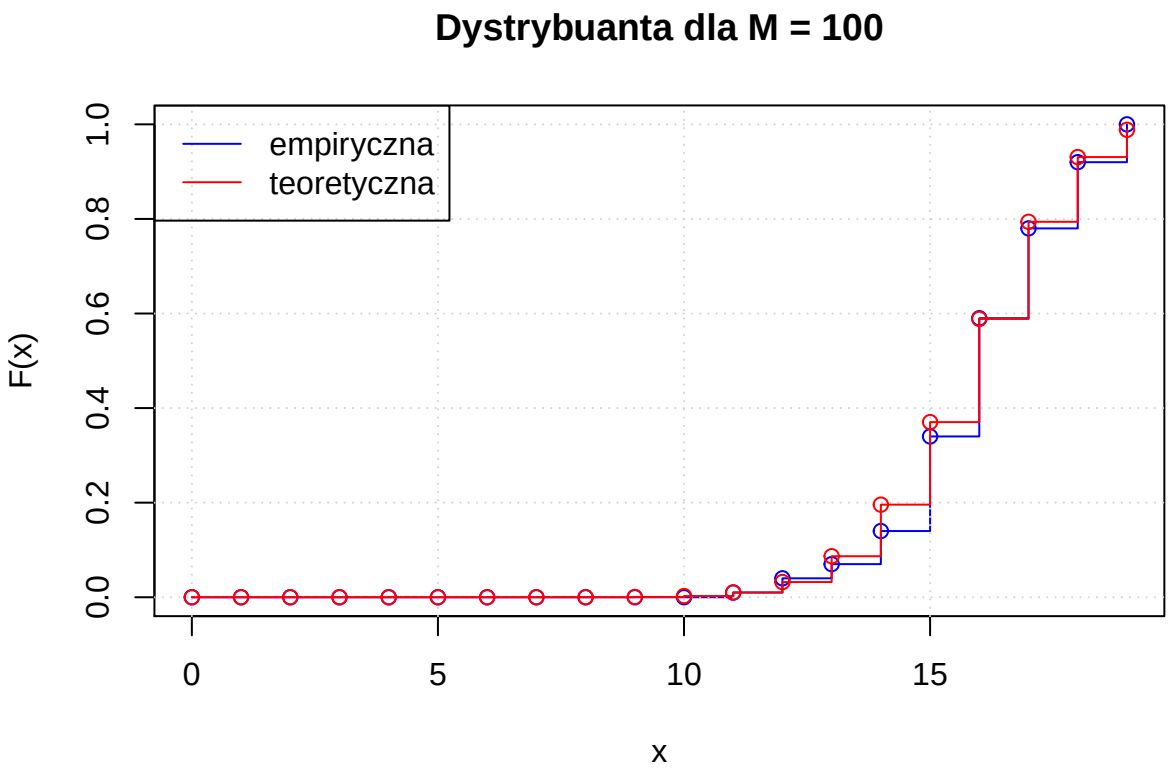
Wykreślenie empirycznej i teoretycznej funkcji prawdopodobieństwa dla M = 10000

```
Freq3 = as.numeric(table(factor(prob3, levels = Arg))) / M3
plot(Freq3 ~ Arg, type = 'h', col = 'blue', xlab = 'x', ylab = 'f(x)',
    main = paste0('Funkcja prawdopodobieństwa dla M = ', M3))
grid()
points(Freq3 ~ Arg, col = 'blue')

points(teor ~ Arg, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



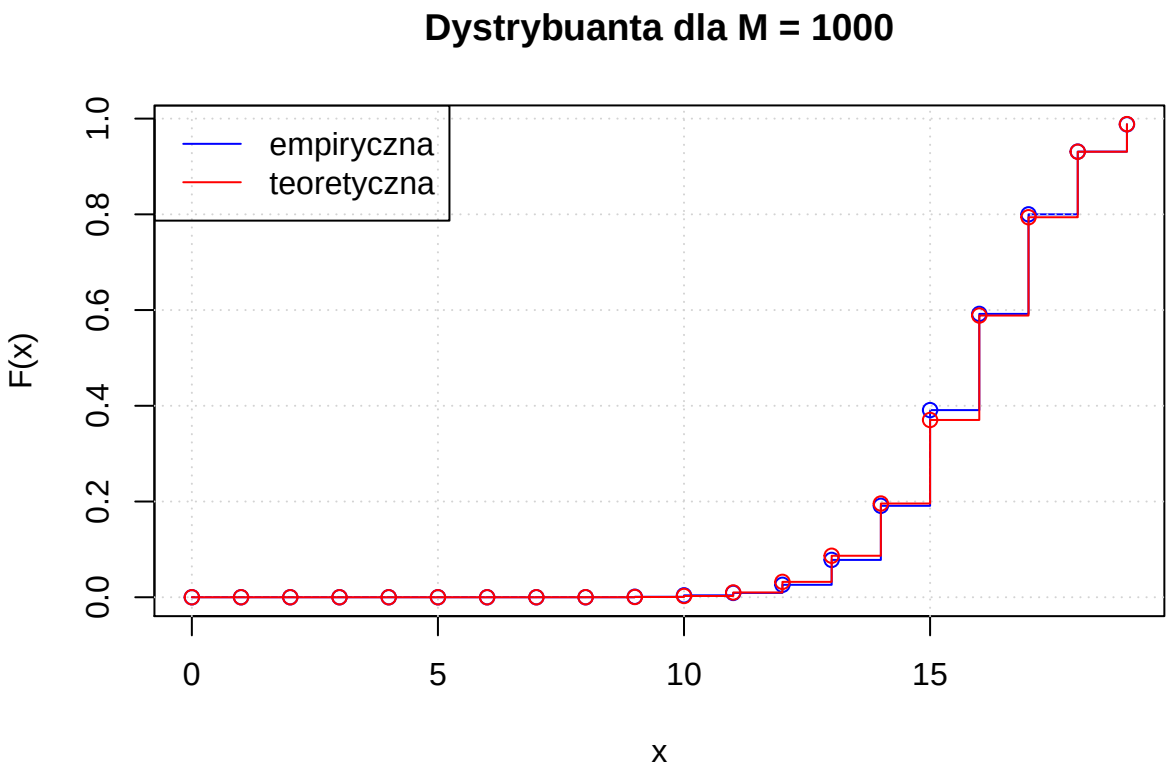
Dystrybuanta empiryczna i teoretyczna dla M = 100

```
plot(cumsum(Freq1) ~ Arg, type = 's', col = 'blue',
    xlab = 'x', ylab = 'F(x)', main = paste0('Dystrybuanta dla M = ', M1))
grid()
points(cumsum(Freq1) ~ Arg, col = 'blue')


lines(dist ~ Arg, type = 's', col = 'red',
      xlab = 'x', ylab = 'F(x)')
points(dist ~ Arg, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



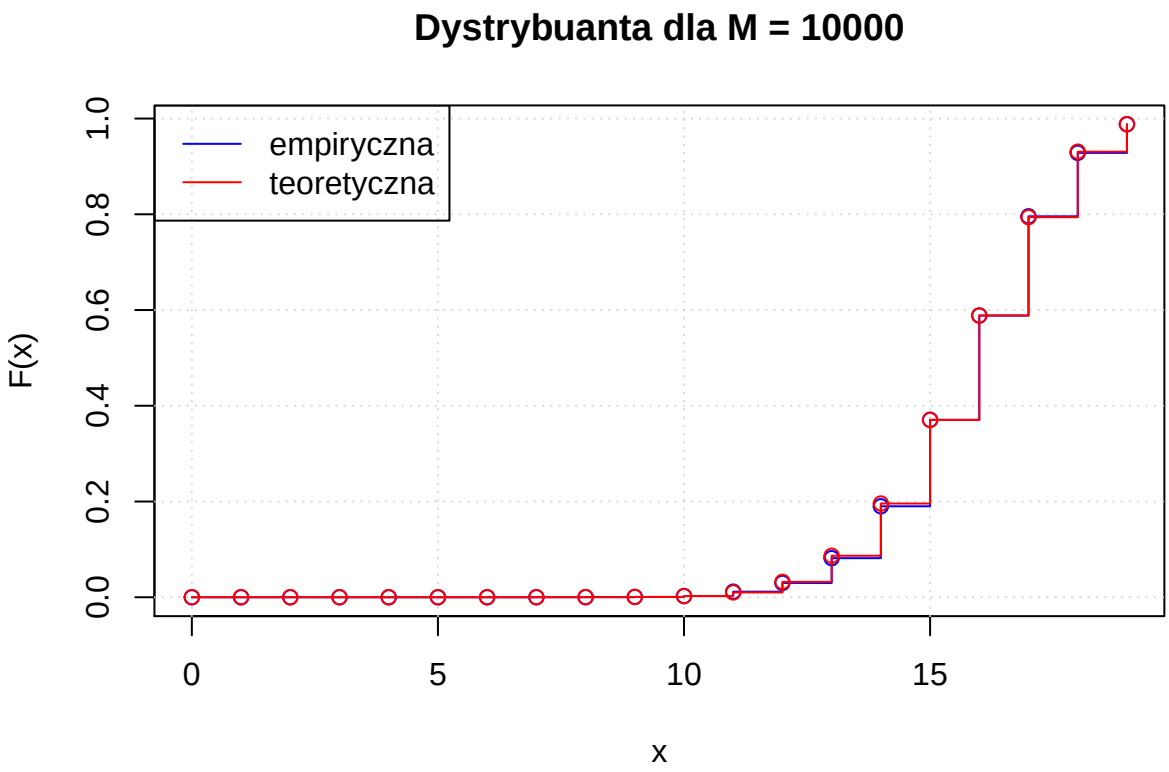
Dystrybuanta empiryczna i teoretyczna dla M = 1000

```
plot(cumsum(Freq2) ~ Arg, type = 's', col = 'blue',
    xlab = 'x', ylab = 'F(x)', main = paste0('Dystrybuanta dla M = ', M2))
grid()
points(cumsum(Freq2) ~ Arg, col = 'blue')


lines(dist ~ Arg, type = 's', col = 'red',
      xlab = 'x', ylab = 'F(x)')
points(dist ~ Arg, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



Dystrybuanta empiryczna i teoretyczna dla M = 10000

```
plot(cumsum(Freq3) ~ Arg, type = 's', col = 'blue',
    xlab = 'x', ylab = 'F(x)', main = paste0('Dystrybuanta dla M = ', M3))
grid()
points(cumsum(Freq3) ~ Arg, col = 'blue')


lines(dist ~ Arg, type = 's', col = 'red',
      xlab = 'x', ylab = 'F(x)')
points(dist ~ Arg, col = 'red')
legend('topleft', c('empiryczna', 'teoretyczna'),
    col = c('blue', 'red'), lwd = 1)
```



Wartości parametrów z prób i wartości teoretyczne średniej i wariancji

```
m1 = mean(prob1); v1 = var(prob1)
m2 = mean(prob2); v2 = var(prob2)
m3 = mean(prob3); v3 = var(prob3)
mt = 20 * 0.8; vt = 20 * 0.8 * (1 - 0.8)
```

Wartość średniej dla rozkładu teoretycznego wynosi: 16, wartość średniej dla M=100 wynosi: 16.11, wartość średniej dla M=1000 wynosi: 15.989, wartość sredniej dla M=10000 wynosi: 16.0124.

Wartość wariancji dla rozkładu teoretycznego wynosi: 3.2, wartość wariancji dla M=100 wynosi: 2.9676, wartość wariancji dla M=1000 wynosi: 3.114, wartość wariancji dla M=10000 wynosi: 3.1698.

Zadanie 5

Treść zadania

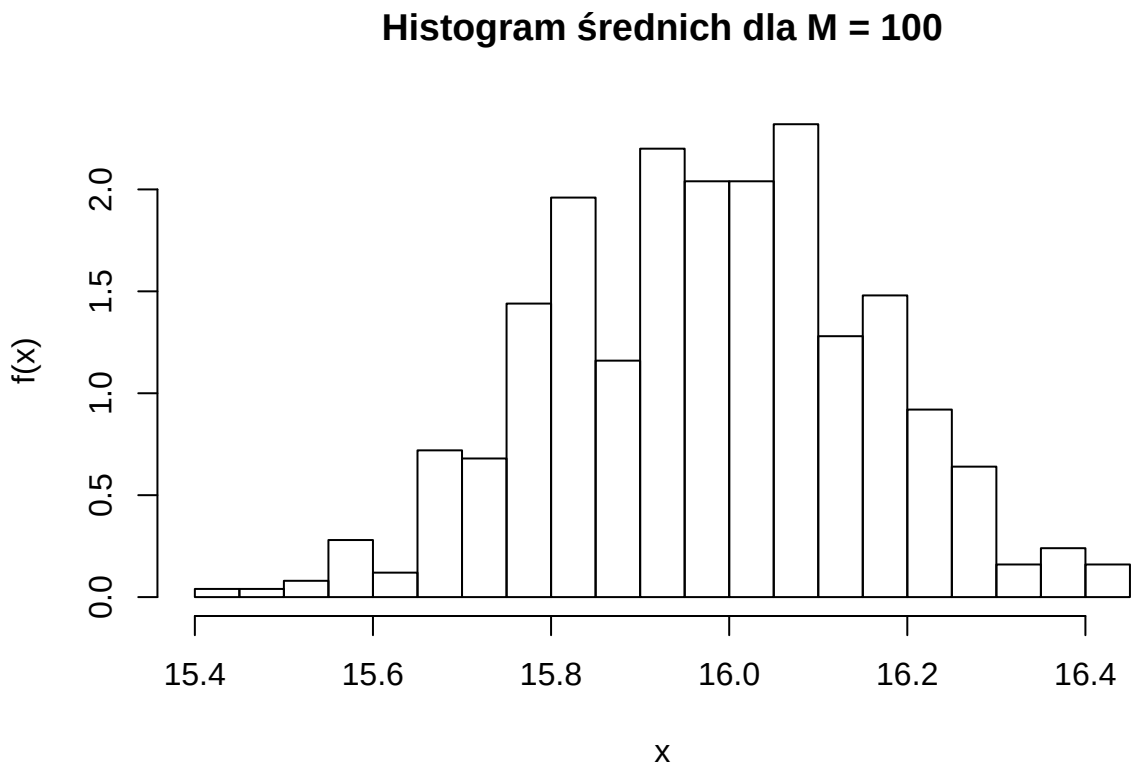
1. Wygeneruj K = 500 realizacji (powtórzeń) prób losowych składających się z M = 100 próbek pochodzących z rozkładu Binom(20, 0.8).
2. Dla wszystkich realizacji oblicz wartości średnie i wariancje. Następnie narysuj histogramy wartości średnich i histogramy wariancji.
3. Powtórz eksperymenty dla M = 1000 i M = 10000. Jak zmieniają się histogramy ze zmianą liczby próbek?

Wskazówka: mm = replicate(500, mean(rbinom(M, 20, 0.8)))

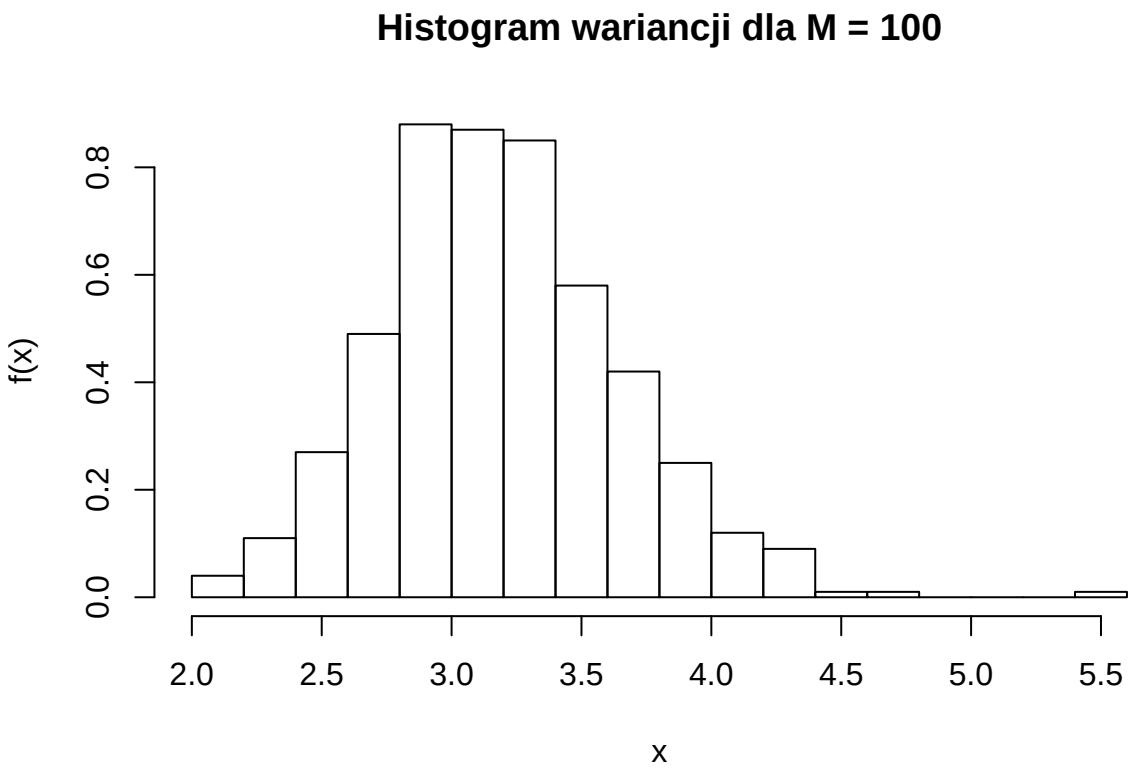
Rozwiązanie

```
M1 = 100
M2=1000
M3=10000
mm1 = replicate(500, mean(rbinom(M1, 20, 0.8)))
mm2 = replicate(500, mean(rbinom(M2, 20, 0.8)))
mm3 = replicate(500, mean(rbinom(M3, 20, 0.8)))

mmv1 = replicate(500, var(rbinom(M1, 20, 0.8)))
mmv2 = replicate(500, var(rbinom(M2, 20, 0.8)))
mmv3 = replicate(500, var(rbinom(M3, 20, 0.8)))
```

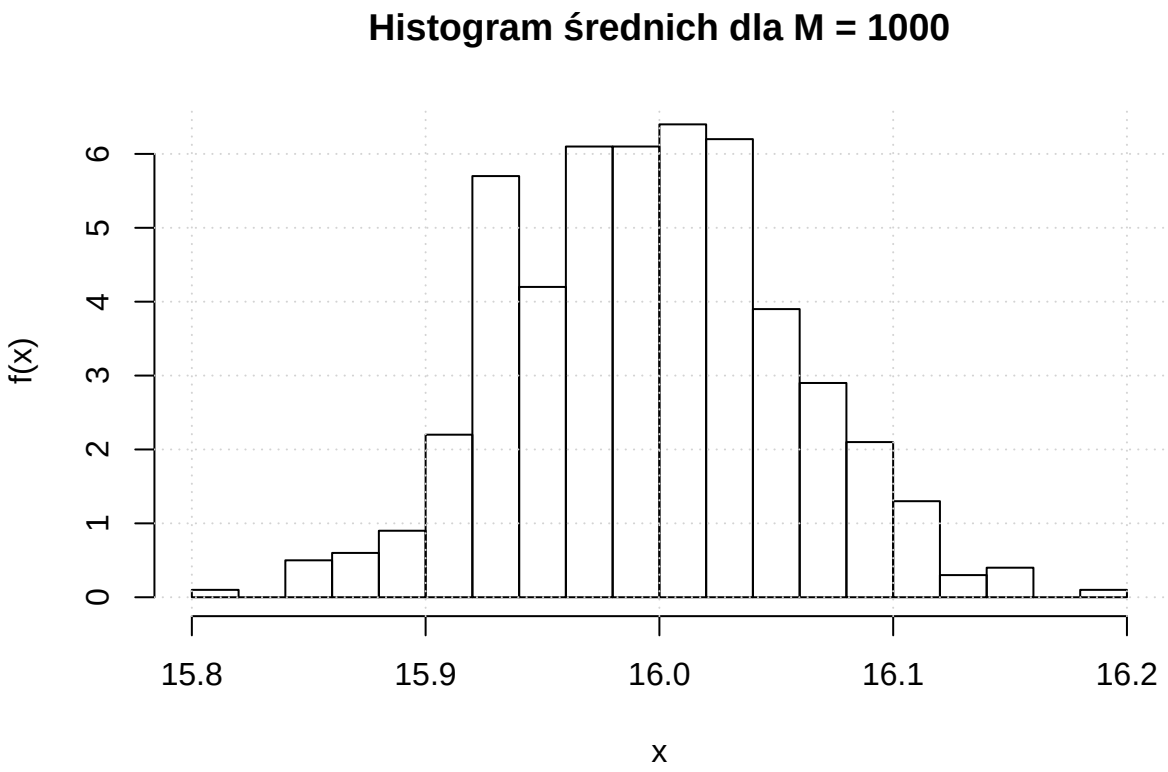
```
hist(mm1, breaks = 20, prob = T, xlab = 'x', ylab = 'f(x)',
     main = 'Histogram średnich dla M = 100')
```



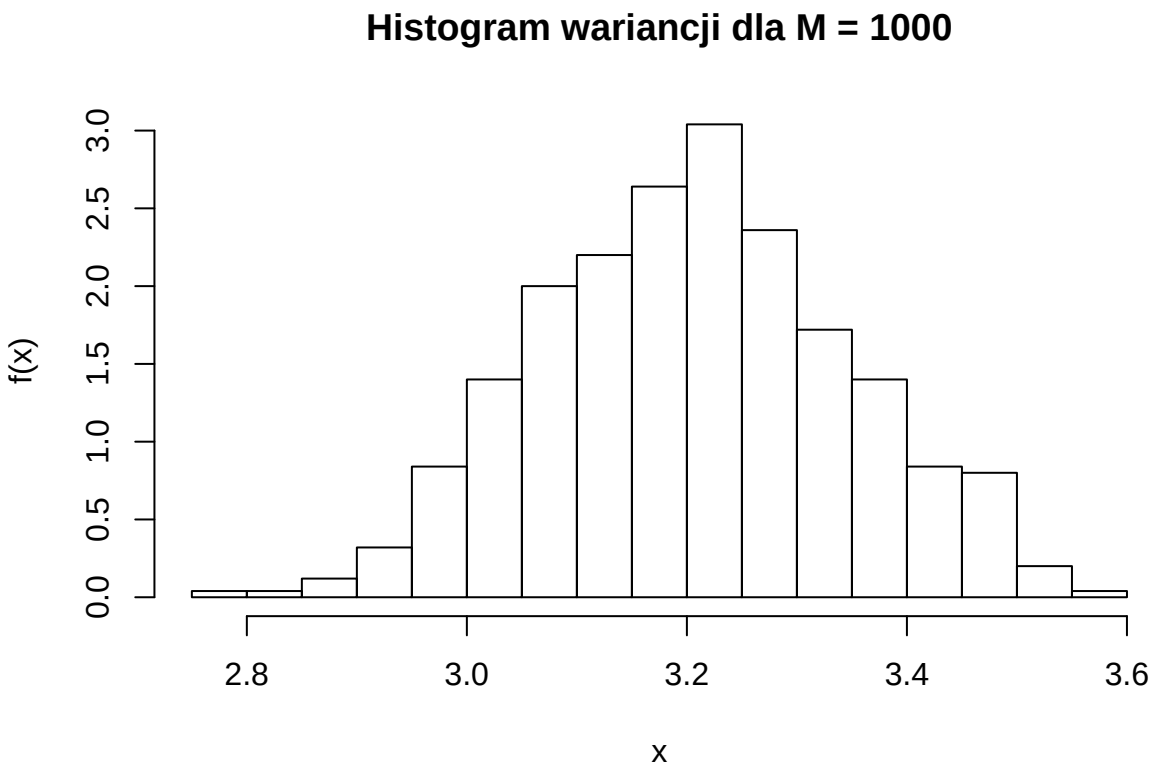

```
hist(mmv1, breaks = 20, prob = T, xlab = 'x', ylab = 'f(x)',
     main = 'Histogram wariancji dla M = 100')
```



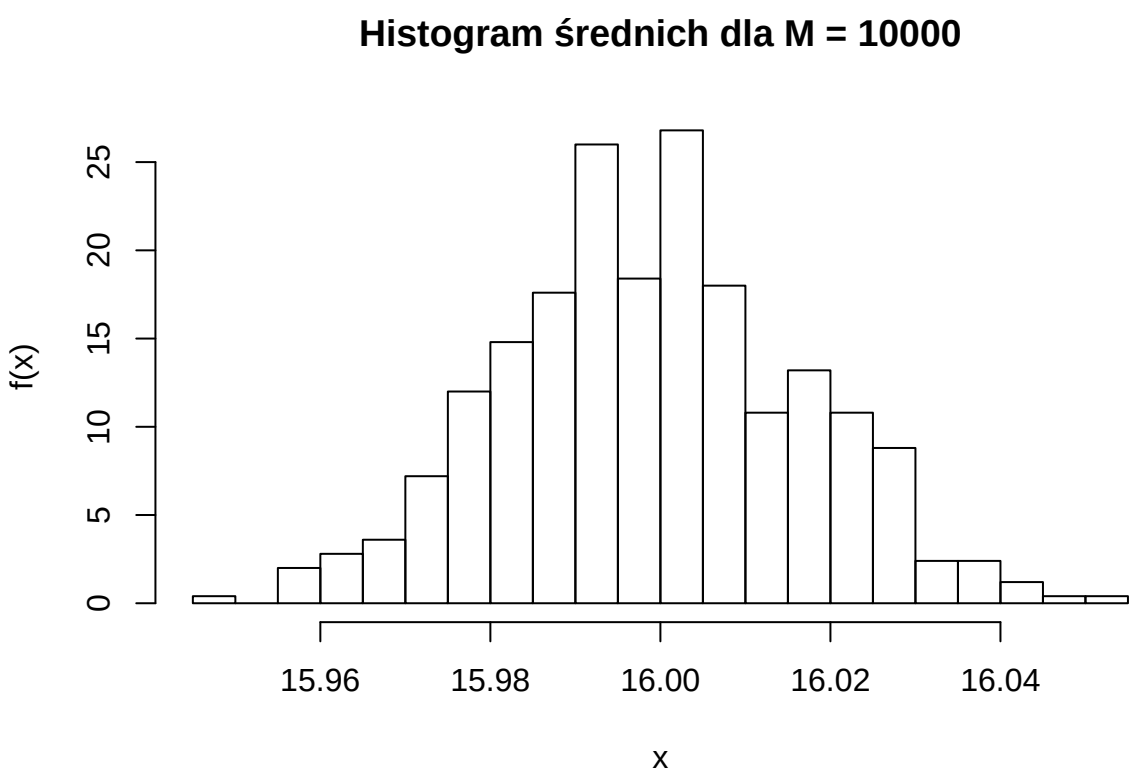

```
hist(mm2, breaks = 20, prob = T, xlab = 'x', ylab = 'f(x)',
     main = 'Histogram średnich dla M = 1000')
curve(dnorm(x, mean = mean(mm2), sd = sqrt(var(mm2))), add = T, col = 'red', -15, 15)
grid()
```



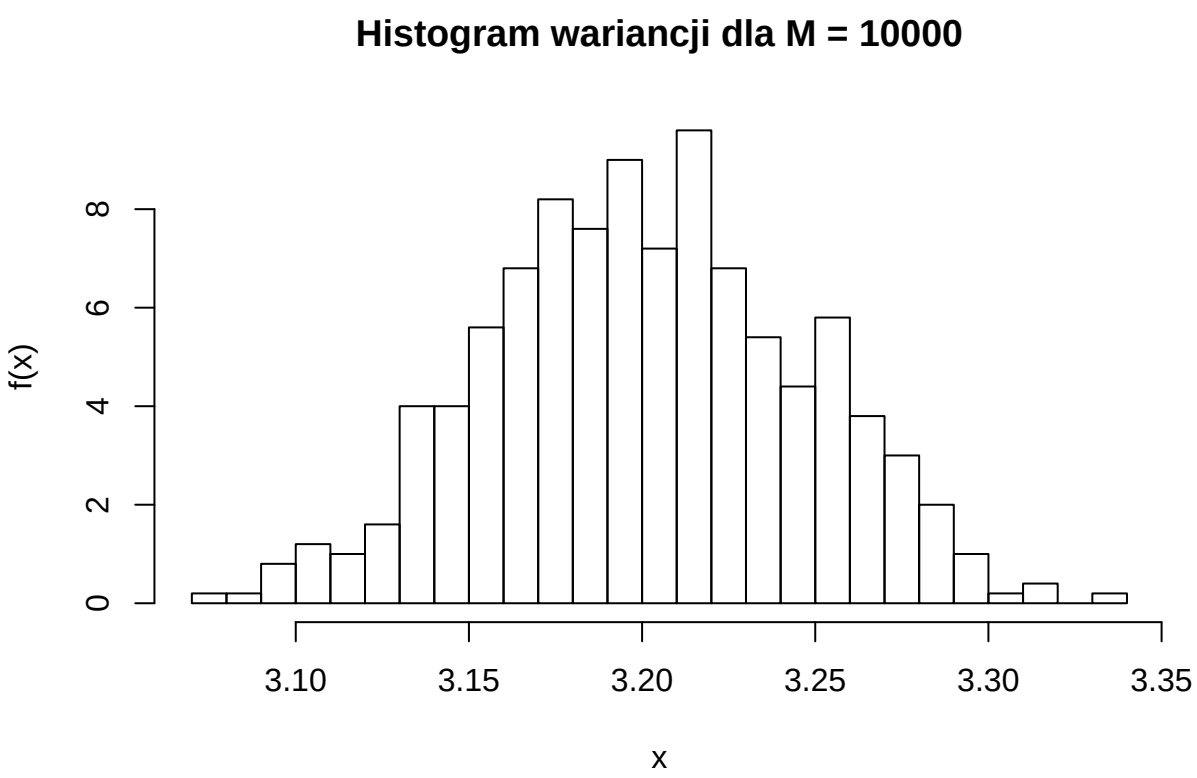

```
hist(mmv2, breaks = 20, prob = T, xlab = 'x', ylab = 'f(x)',
   main = 'Histogram wariancji dla M = 1000')
```




```
hist(mm3, breaks = 20, prob = T, xlab = 'x', ylab = 'f(x)',
     main = 'Histogram średnich dla M = 10000')
```




```
hist(mmv3, breaks = 20, prob = T, xlab = 'x', ylab = 'f(x)',
     main = 'Histogram wariancji dla M = 10000')
```



Wraz ze wzrostem liczby próbek z M = 100 do M = 10000 zmniejsza się rozstęp średnich w próbach losowych jak i wariancji w próbach losowych. Wraz ze wzrostem M wartości są bardziej skoncentrowane, wzrasta kurtoza.