Sprawozdanie 1

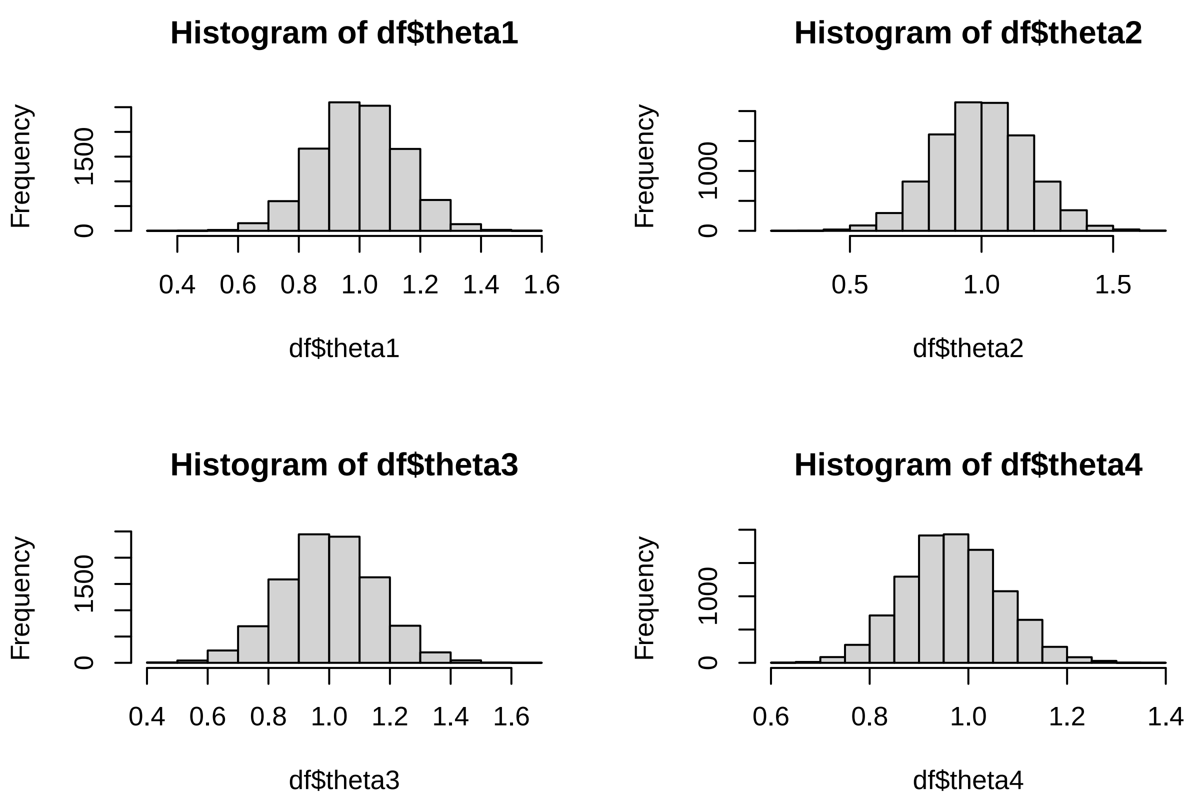
Katarzyna Stasińska

2023-10

Zadanie 1

W przypadku $\hat{\theta}_3$ postanowiłam, że wektorem z wagami, będzie znormalizowany wektor wygenerowany przez $\varphi(\Phi^{-1}(\frac{i-1}{n-1}))$, gdzie oznaczenia są takie same jak w podpunkcie (iv).

a) Rozważmy 50 obserwacji z rozkładu N(1,1), wyliczmy na nich wartości $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ i $\hat{\theta}_4$. Na wykresach możemy zauważyć efekt powtórzenia tej procedury 10 000 razy. Zwróćmy uwagę, że ich wartości są rzeczywiście bliskie $\theta = 1$.

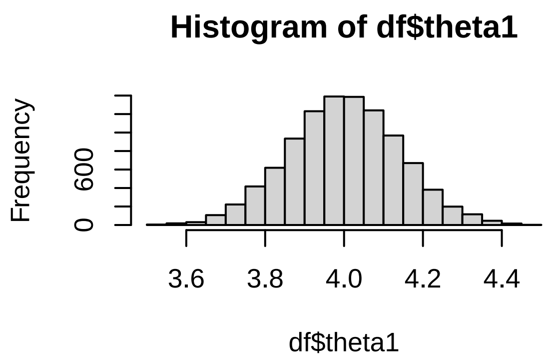


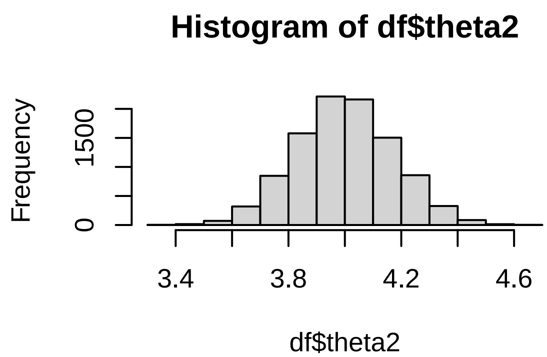
Poniżej prezentuję szacowany błąd średniokwadratowy, wariancję i obciążenie każdego z estymatorów. Wyniki są podobne w przypadku każdego z estymatorów. Obciążenia są bliskie 0.

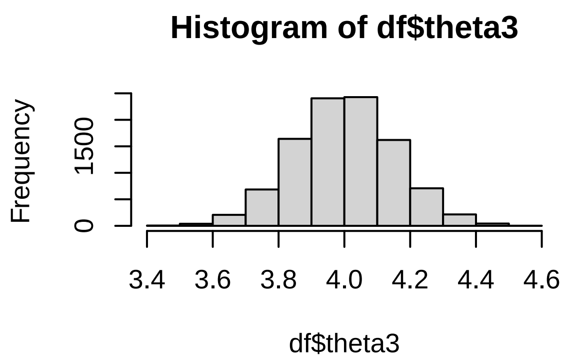
```
## [[1]]
##     theta1     theta2     theta3     theta4
## 0.02015677 0.03074256 0.02405746 0.01051962
##
```

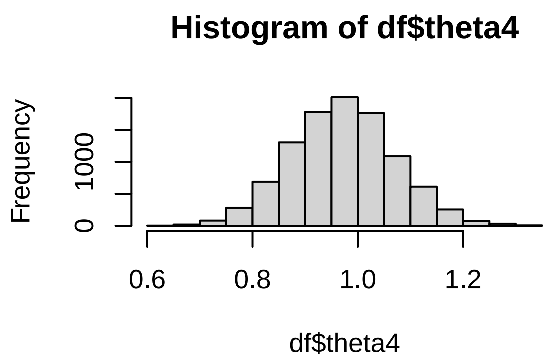
```
## [[2]]
##      theta1      theta2      theta3      theta4
## 0.020156540 0.030741000 0.024056922 0.009668371
##
## [[3]]
##        theta1        theta2        theta3        theta4
## -0.0004834971  0.0012499922 -0.0007348907 -0.0291761253
```

b) Analogicznie rozważmy 50 obserwacji z rozkładu N(4,1), wyliczmy na nich wartości $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ i $\hat{\theta}_4$. Na wykresach możemy zauważyć efekt powtórzenia tej procedury 10 000 razy. Zwróćmy uwagę, że i w tym przypadku wartości są rzeczywiście bliskie $\theta = 4$, poza $\hat{\theta}_4$, której średnia jest bliska 1





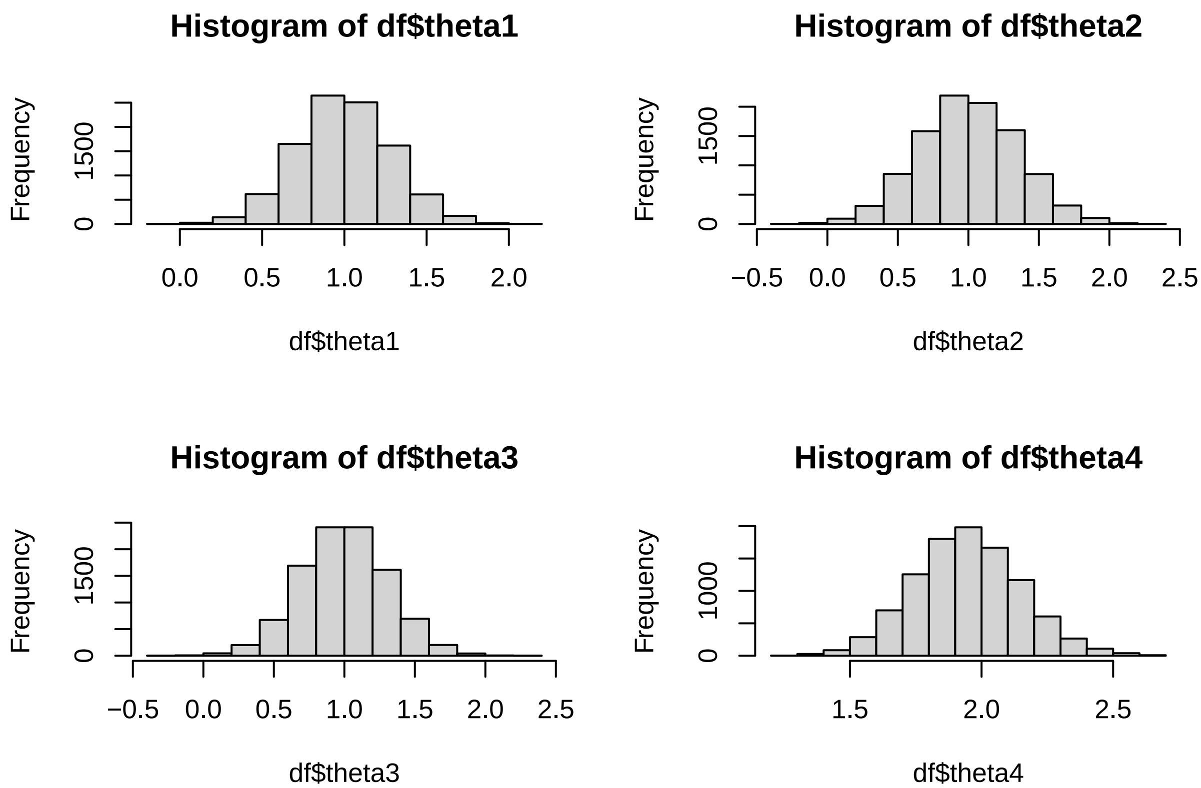




Poniżej prezentuję szacowany błąd średniokwadratowy, wariancję i obciążenie każdego z estymatorów. W tym przypadku wyniki się już różnią, $\hat{\theta}_4$ odstaje od reszty i najsłabiej minimalizuje statystyki.

```
## [[1]]
##     theta1     theta2     theta3     theta4
## 0.01971947 0.03012138 0.02354489 9.18167282
##
## [[2]]
##      theta1      theta2      theta3      theta4
## 0.019719134 0.030121380 0.023544656 0.009674989
##
## [[3]]
##        theta1        theta2        theta3        theta4
##  5.773126e-04  2.214509e-05  4.828489e-04 -3.028531e+00
```

c) Na koniec rozważmy 50 obserwacji z rozkładu N(1,2), wyliczmy na nich wartości $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ i $\hat{\theta}_4$. Na wykresach możemy zauważyć efekt powtórzenia tej procedury 10 000 razy. Zwróćmy uwagę, że i w tym przypadku wartości są rzeczywiście bliskie $\theta = 1$, poza $\hat{\theta}_4$, której średnia jest bliska 2.



Poniżej prezentuję szacowany błąd średniokwadratowy, wariancję i obciążenie każdego z estymatorów. W tym przypadku $\hat{\theta}_4$ znów odstaje od reszty i najsłabiej minimalizuje statystyki.

```
## [[1]]
##     theta1     theta2     theta3     theta4
## 0.08114646 0.12334206 0.09375936 0.93418894
##
## [[2]]
##     theta1     theta2     theta3     theta4
## 0.08114610 0.12334185 0.09375672 0.04044878
##
## [[3]]
##        theta1        theta2        theta3        theta4
## -0.0005983620  0.0004589222 -0.0016237878  0.9453783170
```

Na podstawie powyższych informacji można stwierdzić, że $\hat{\theta}_4$ jest słabym estymatorem.

Zadanie 2

Funkcja `set.seed(1)` inicjuje generator liczb pseudolosowych z ziarnem podanym w argumencie. W praktyce używa się jej, by przy każdym uruchomieniu skryptu otrzymywać takie same wyniki. Dzięki niej próbując konkretnie omówić daną próbkę(tj. wskazując konkretne rekordy) wylosowanych danych, mamy pewność, że przy kolejnym odpaleniu skryptu nasze wnioski wciąż będą w pełni prawdziwe. Możemy też pobrać nasze wylosowane dane i za każdym razem ich nie inicjować (duża oszczędność czasu w przypadku większej ilości danych).

Zadanie 3

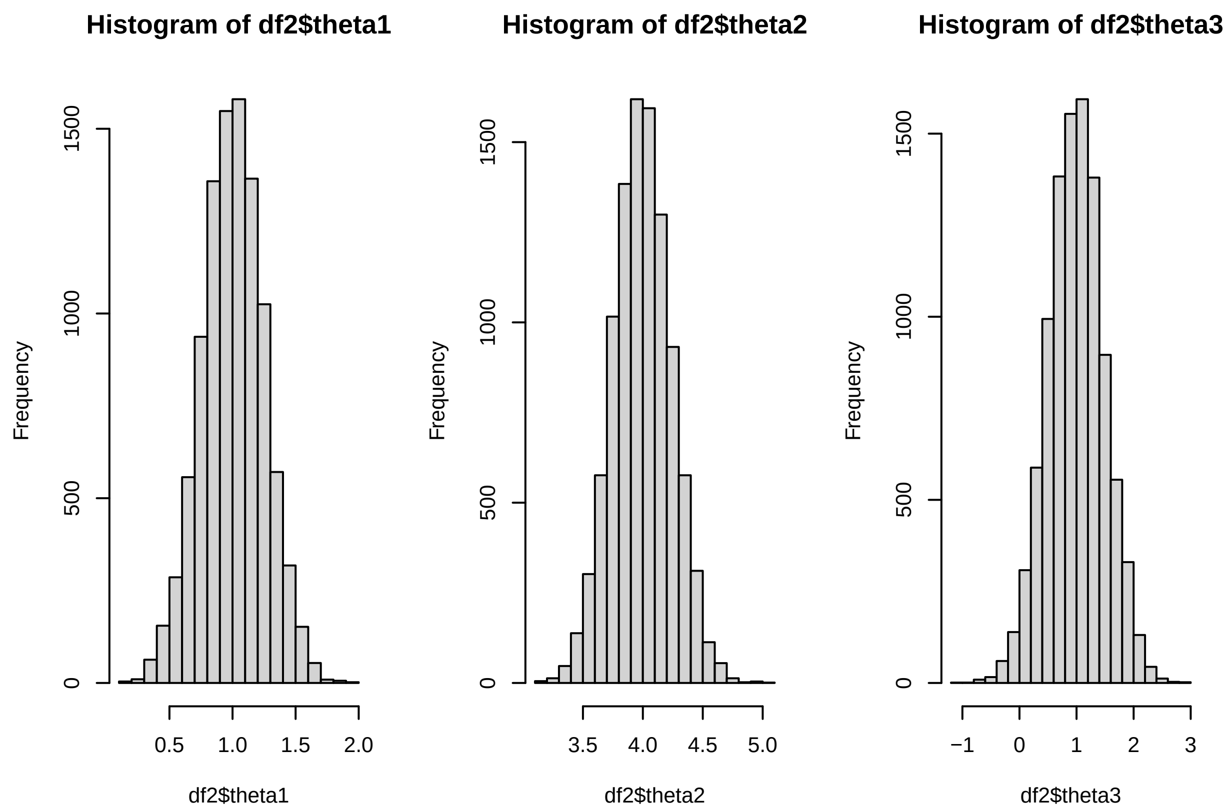
Potencjalne estymatory największej wiarogodności, to miejsca zerowe pochodnej funkcji logwiarogodności, dla rozkładu logistycznego wyraża się ona wzorem: $l'(\theta) = \frac{1}{\sigma}(n - 2\Sigma\frac{e^{\frac{-(x_i-\theta)}{\sigma}}}{1+e^{\frac{-(x_i-\theta)}{\sigma}}})$. Nie jest to łatwe dla tak skomplikowanego wyrażenia. Pokażmy, że nie są to tylko potencjalne, a rzeczywiste estymatory i istnieje tak naprawdę jeden. Zwróćmy uwagę, że druga pochodna funkcji logwiarogodności dana wzorem: $l''(\theta) = \frac{-1}{\sigma^2}(2\Sigma\frac{e^{\frac{-(x_i-\theta)}{\sigma}}}{(1+e^{\frac{-(x_i-\theta)}{\sigma}})^2}$ jest stale mniejsza od zera. Jest tak, ponieważ mianownik i licznik ułamka są dodatnie, przez co wszystkie czynniki iloczynu są dodatnie, poza pierwszym. Oznacza to, że $l'(\theta)$ maleje, a że jest to funkcja określona na całym $\mathbb{R}$, to posiada dokładnie jedno miejsce zerowe, które maksymalizuje wartość funkcji $l(\theta)$. Zatem jest estymatorem największej wiarogodności. Pozostaje wyliczyć to miejsce zerowe i pomogą w tym metody numeryczne.

Zadanie 4

Rozważmy metodę Newtona (zwaną również metodą stycznych), jako narzędzie pozwalające wyliczyć miejsce zerowe funkcji $l'(\theta)$. Polega ona na tym, że zaczynamy od pewnej wartości θ_0 (ważne, aby mieściła się w przedziale, w którym chcemy poszukiwać miejsca zerowego, w naszym przypadku jednak nie ma to znaczenia, bo miejsce zerowe jest tylko jedno). Jest to metoda iteracyjna, a kolejne iteracje wyglądają następująco: wyznaczamy punkt przecięcia OX i stycznej do funkcji $l'(\theta)$ w punkcie $l'(\theta_i)$, jest on równy θ_{i+1}. Możemy to zapisać wzorem $\theta_{i+1} = \theta_i - \frac{l'(\theta+i)}{l''(\theta+i)}$. Iteracje kończymy na przykład, gdy $|l'(\theta_i)| < \epsilon$ albo $|\theta_{i+1} - \theta_i| < \epsilon$, albo gdy liczba iteracji jest wystarczająca duża. Szukane miejsce zerowe jest bliskie θ_k, gdzie k to ostatnia iteracja.

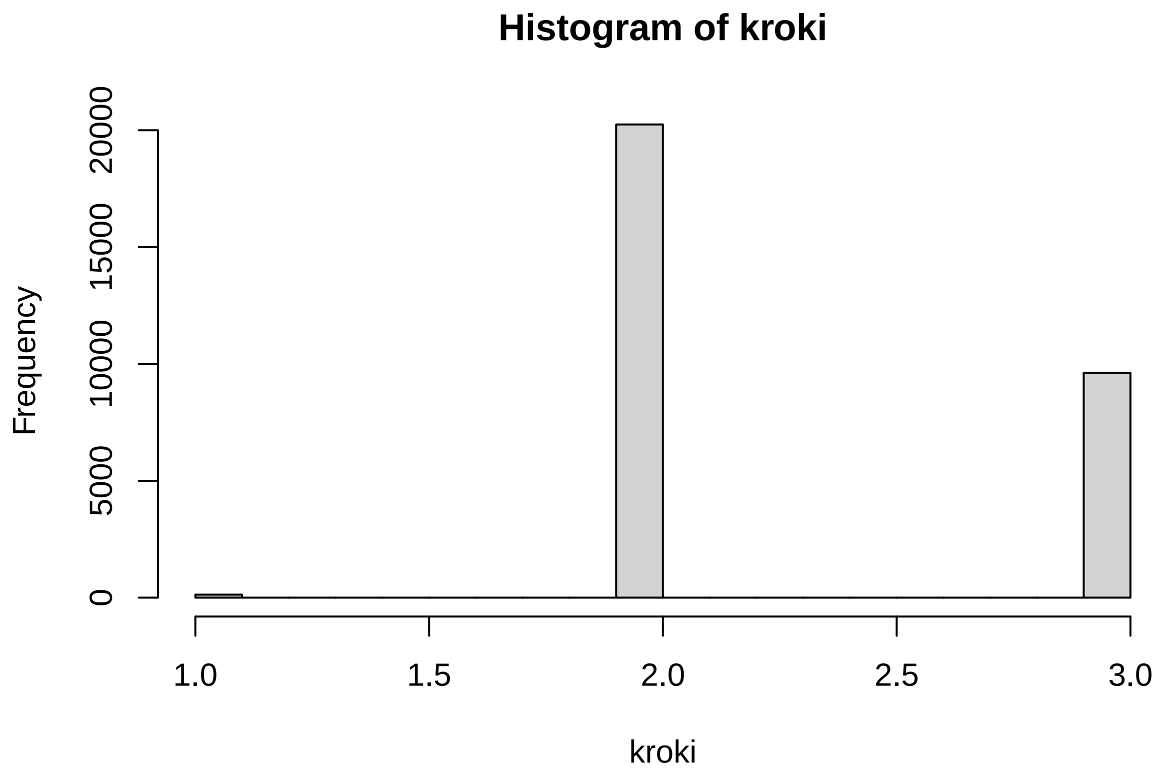
Zadanie 5

Szukane miejsca zerowe są w pobliżu podanej θ, dlatego właśnie ten punkt jest punktem początkowym w metodzie Newtona. Liczbę kroków ograniczam przez 5000, raczej nigdy tyle nie następuje, bo przy ustalonym jak powyżej punkcie startowym, dość szybko są znajdowane punkty, dla których funkcja l' przyjmuje wartości bliskie zeru. Poniżej przedstawiam histogramy z wyliczonymi estymatorami θ oraz szacowany błąd średniokwadratowy, wariancję i obciążenie dla każdego z podpunktów. W każdym przypadku wyniki są bliskie zeru (w podpunkcie *c*) MSE i var przekraczają 2, ale powinny być 4 razy większe niż w *a*) i *b*), co się zgadza.



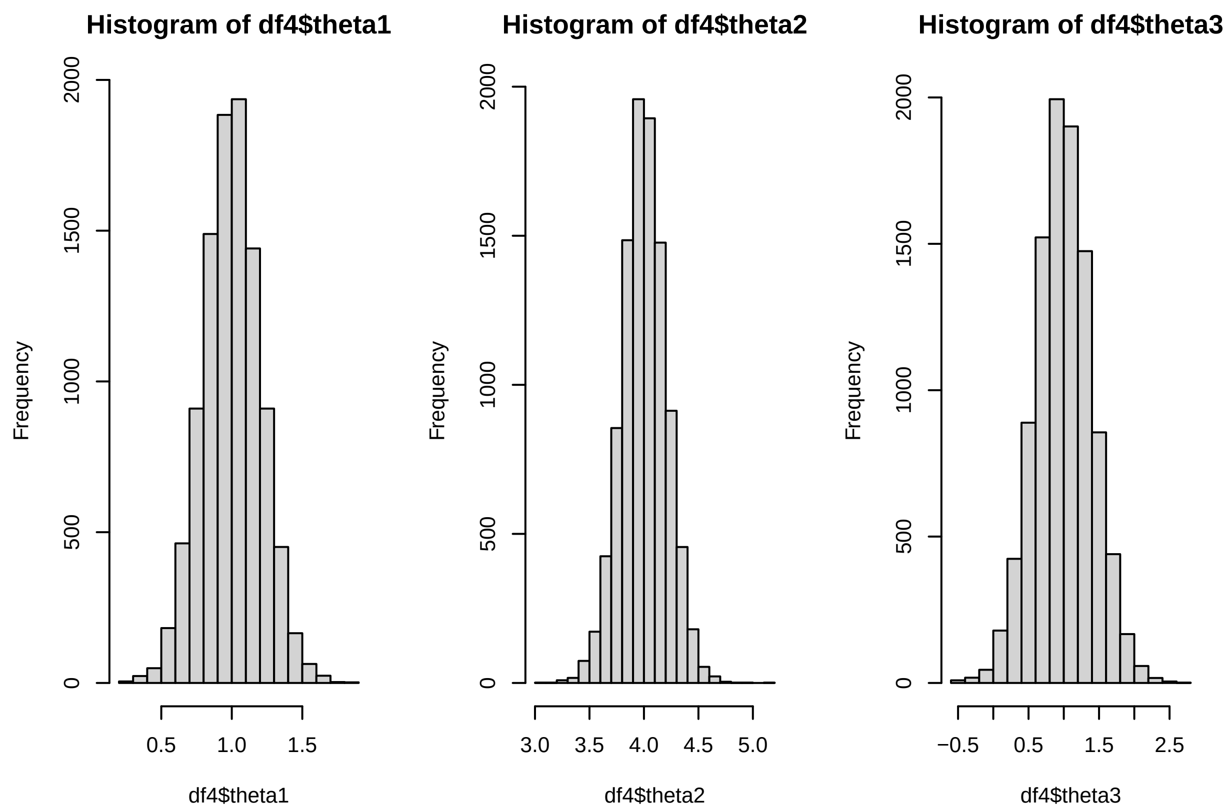

```
##          MSE       var         bias
## 1 0.6052853 0.6051381  0.003836375
## 2 0.5890368 0.5888771 -0.003996478
## 3 2.4165327 2.4155701 -0.009811088
```

Można lepiej przyjrzeć się liczbie kroków potrzebnych do zakończenia metody Newtona na histogramie poniżej.



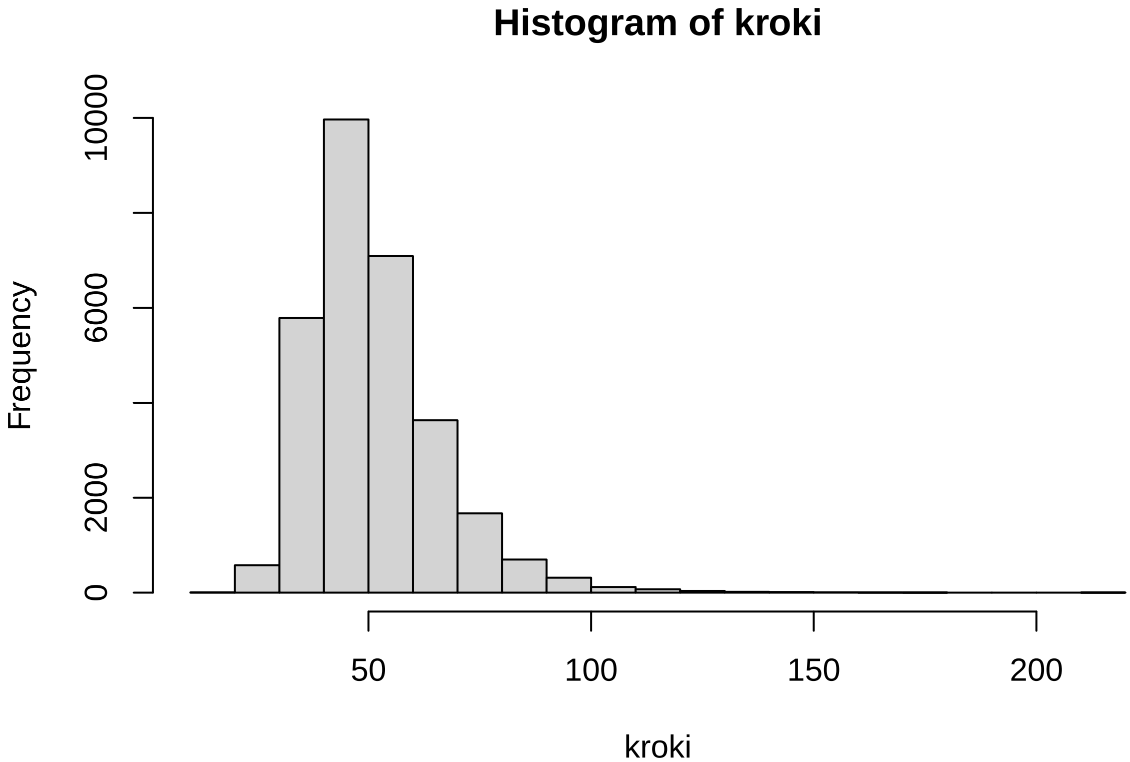
zadanie 6

Jak w zadaniu wyżej, podaną θ uznaję za punkt początkowy w metodzie Newtona. Liczbę kroków ograniczam przez 5000, raczej nigdy tyle nie następuje, bo przy ustalonym w ten sposób punkcie startowym, dość szybko są znajdowane punkty, dla których funkcja l' przyjmuje wartości bliskie zeru. Poniżej przedstawiam histogramy z wyliczonymi estymatorami θ oraz szacowany błąd średniokwadratowy, wariancję i obciążenie dla każdego z podpunktów. W każdym przypadku wyniki są bliskie zeru, znowu trzecia wartość MSE i var jest 4 razy większa niż poprzednie dwie, co zgadza się, bo mamy $\sigma = 2$.



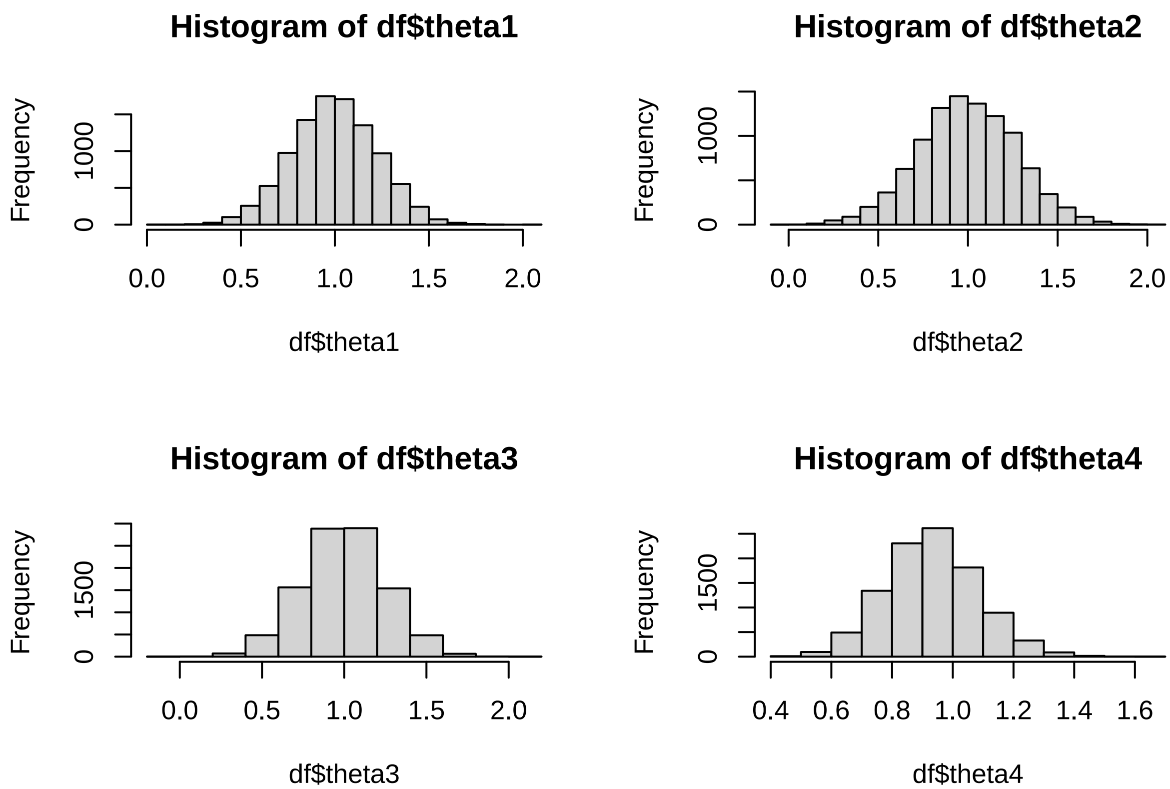

```
##         MSE       var          bias
## 1 0.4233177 0.4233134 -0.0006548678
## 2 0.4253594 0.4253243  0.0018750379
## 3 1.6396146 1.6395516 -0.0025107490
```

Analogicznie jak w zadaniu 5, poniżej przedstawiam histogram liczby kroków potrzebnych do zakończenia funkcji Newton.


Histogram of kroki
Frequency
0
2000
6000
10000
50
100
150
200
kroki

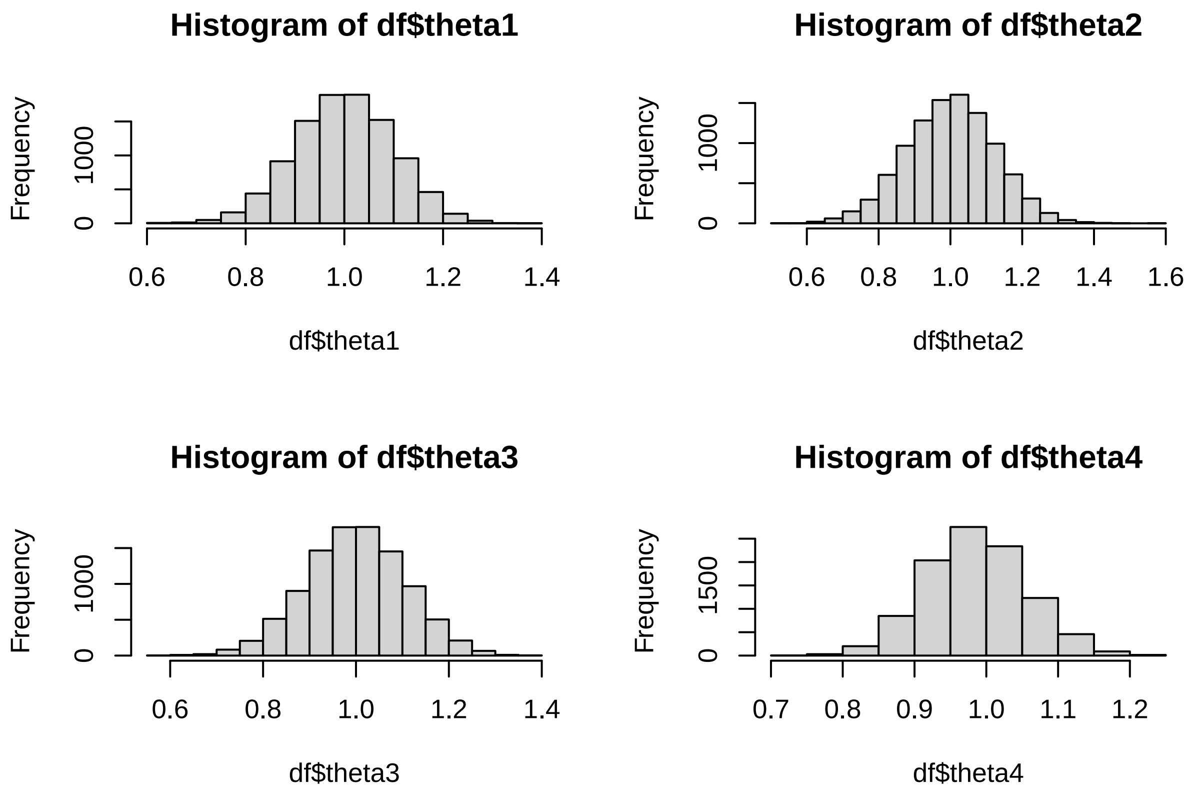
zadanie 7

zadanie 1a) n=20




```
## [[1]]
##     theta1     theta2     theta3     theta4
## 0.05036254 0.07422788 0.06164512 0.02775485
##
## [[2]]
##     theta1     theta2     theta3     theta4
## 0.05035623 0.07422432 0.06164426 0.02296466
##
## [[3]]
##        theta1        theta2        theta3        theta4
## -0.0025117113 -0.0018882604 -0.0009324537 -0.0692111613
```

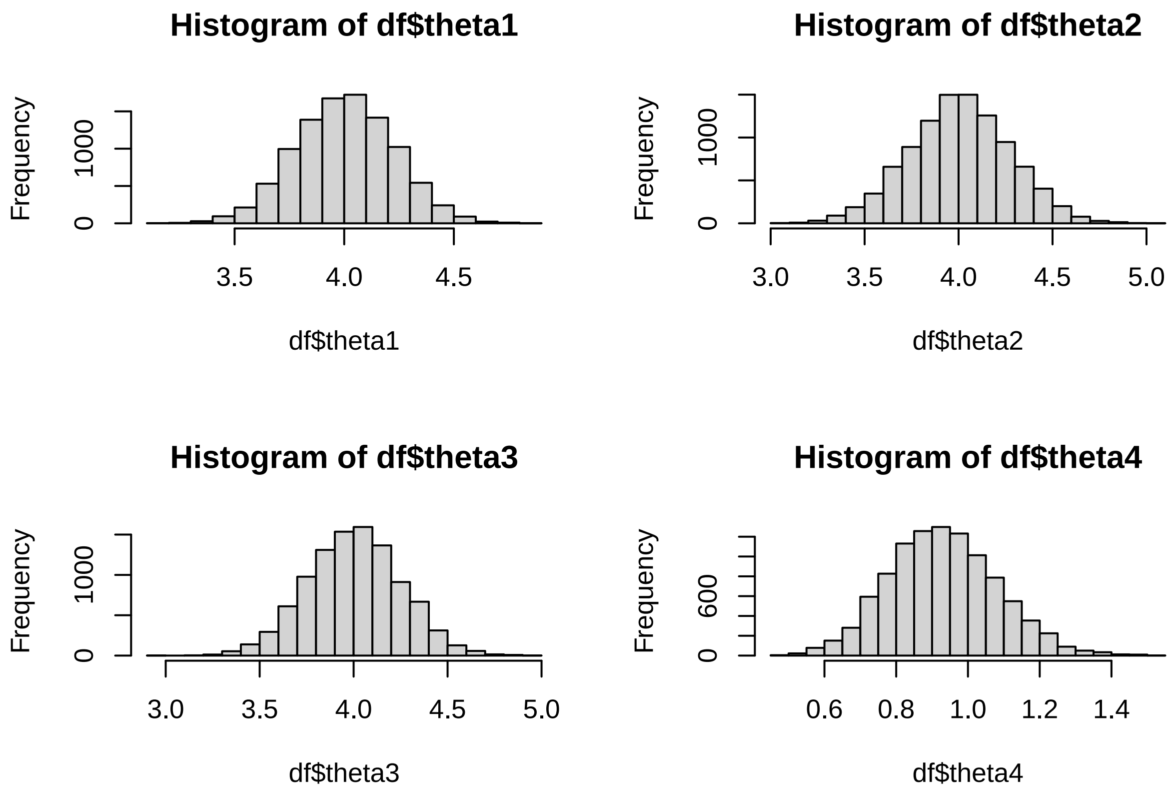
zadanie 1a) n=100




```
## [[1]]
##      theta1      theta2      theta3      theta4
## 0.009856718 0.015206839 0.011473092 0.005118363
##
## [[2]]
##      theta1      theta2      theta3      theta4
## 0.009856689 0.015206638 0.011472835 0.004909625
##
## [[3]]
##        theta1        theta2        theta3        theta4
## -0.0001679528  0.0004483666 -0.0005067876 -0.0144477427
```

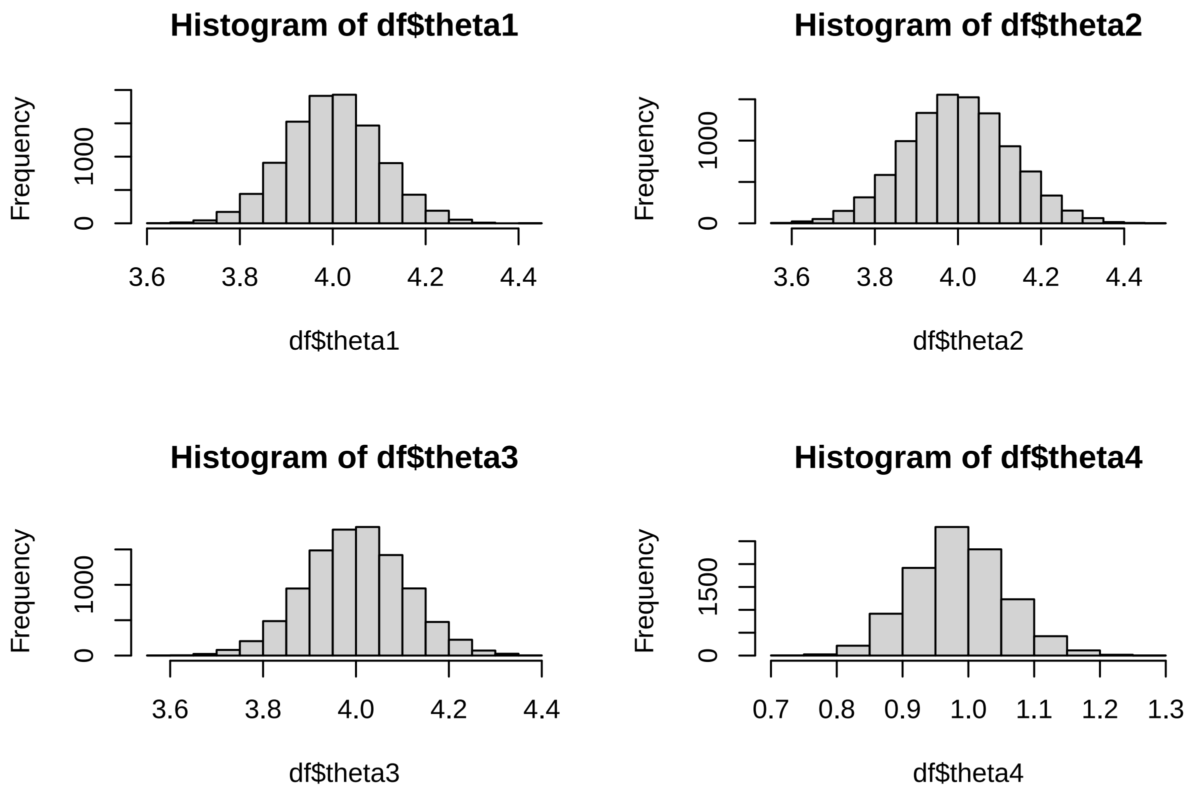
Wnioski: MSE i var są najbliższe zeru dla $n = 100$, potem dla $n = 50$, a na końcu $n = 20$. Przy biasie nie widać aż takich różnic.

zadanie 1b) n=20




```
## [[1]]
##     theta1     theta2     theta3     theta4
## 0.05008064 0.07249274 0.06114820 9.44590155
##
## [[2]]
##     theta1     theta2     theta3     theta4
## 0.05007364 0.07247331 0.06114045 0.02272933
##
## [[3]]
##       theta1       theta2       theta3       theta4
##  0.002645563  0.004408631  0.002783069 -3.069718589
```

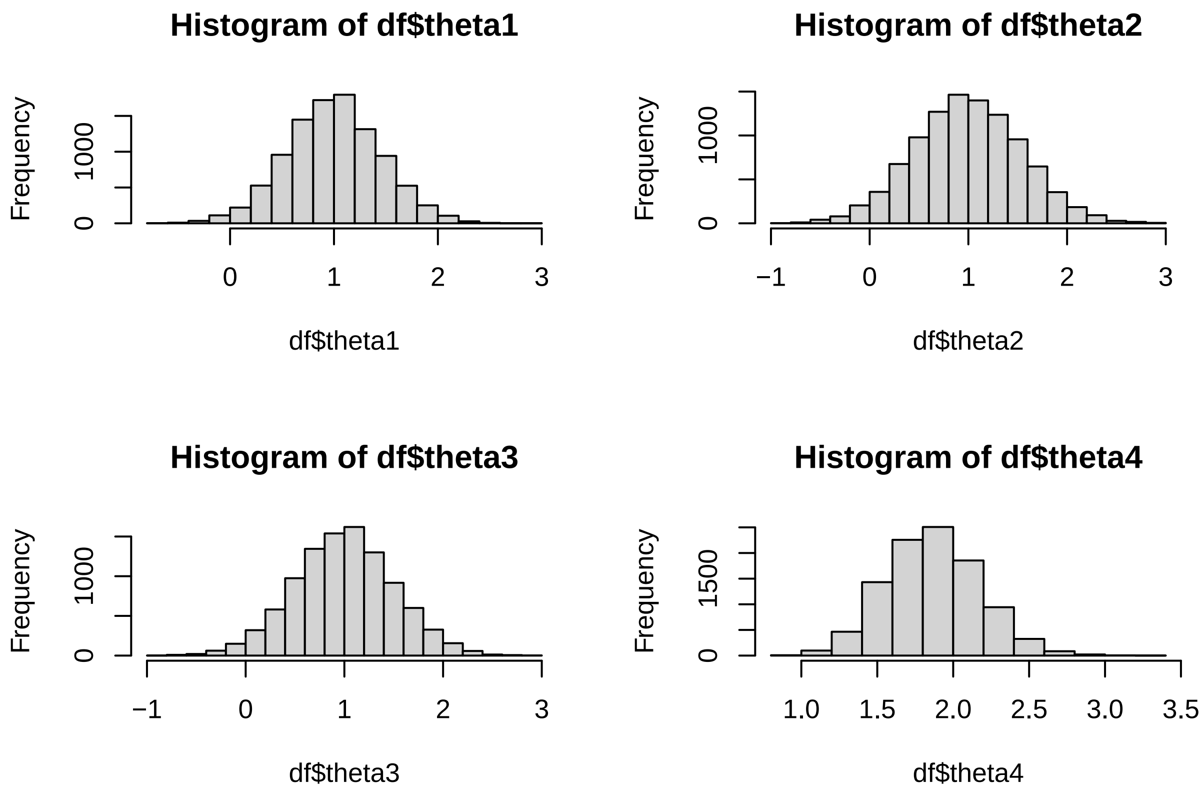
zadanie 1b) n=100




```
## [[1]]
##     theta1     theta2     theta3     theta4
## 0.01008459 0.01572784 0.01152461 9.09199960
##
## [[2]]
##      theta1      theta2      theta3      theta4
## 0.010084566 0.015727696 0.011524458 0.005029478
##
## [[3]]
##        theta1        theta2        theta3        theta4
## -0.0001588990  0.0003787100 -0.0003899568 -3.0144601713
```

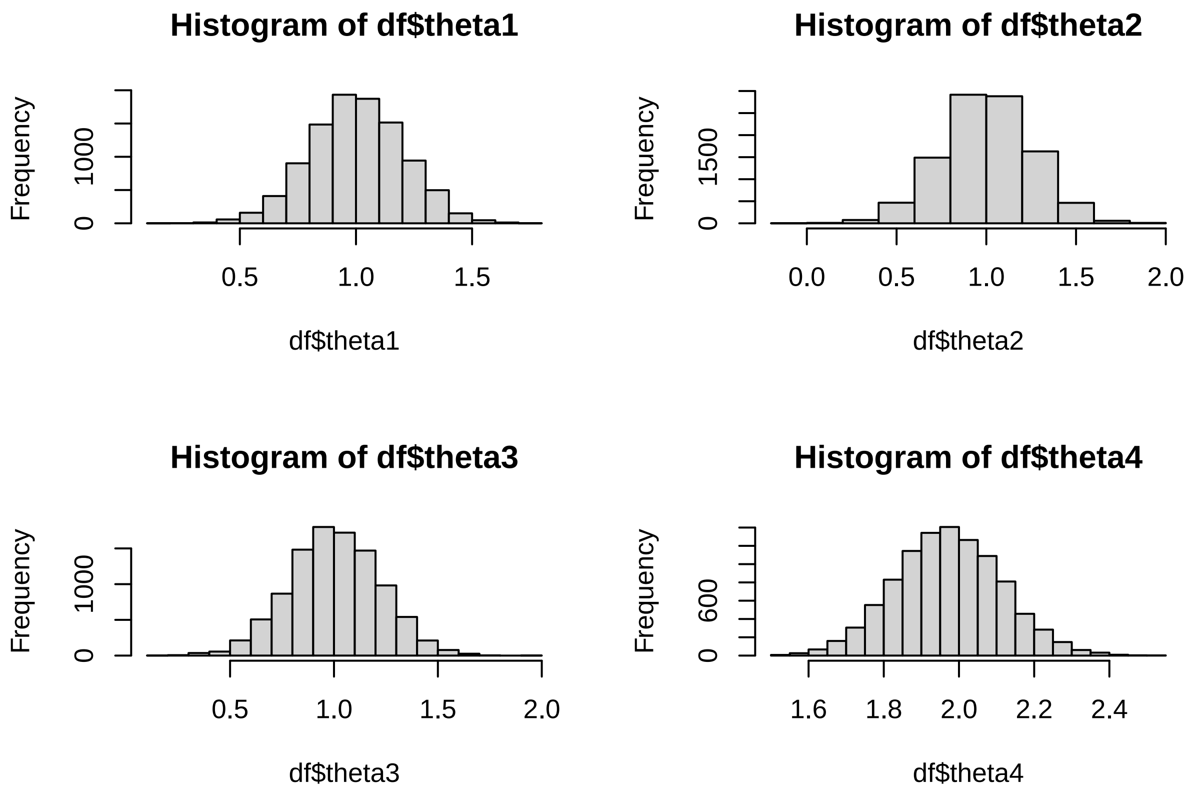
Wnioski: Tak samo jak wyżej, to znaczy MSE i var są najbliższe zeru dla $n = 100$, potem dla $n = 50$, a na końcu $n = 20$. Przy biasie nie widać aż takich różnic.

zadanie 1c) n=20




```
## [[1]]
##    theta1    theta2    theta3    theta4
## 0.2039708 0.2944147 0.2506513 0.8424573
##
## [[2]]
##     theta1     theta2     theta3     theta4
## 0.20396228 0.29437234 0.25064054 0.09374098
##
## [[3]]
##       theta1       theta2       theta3       theta4
## -0.002918601 -0.006505678 -0.003273307  0.865283927
```

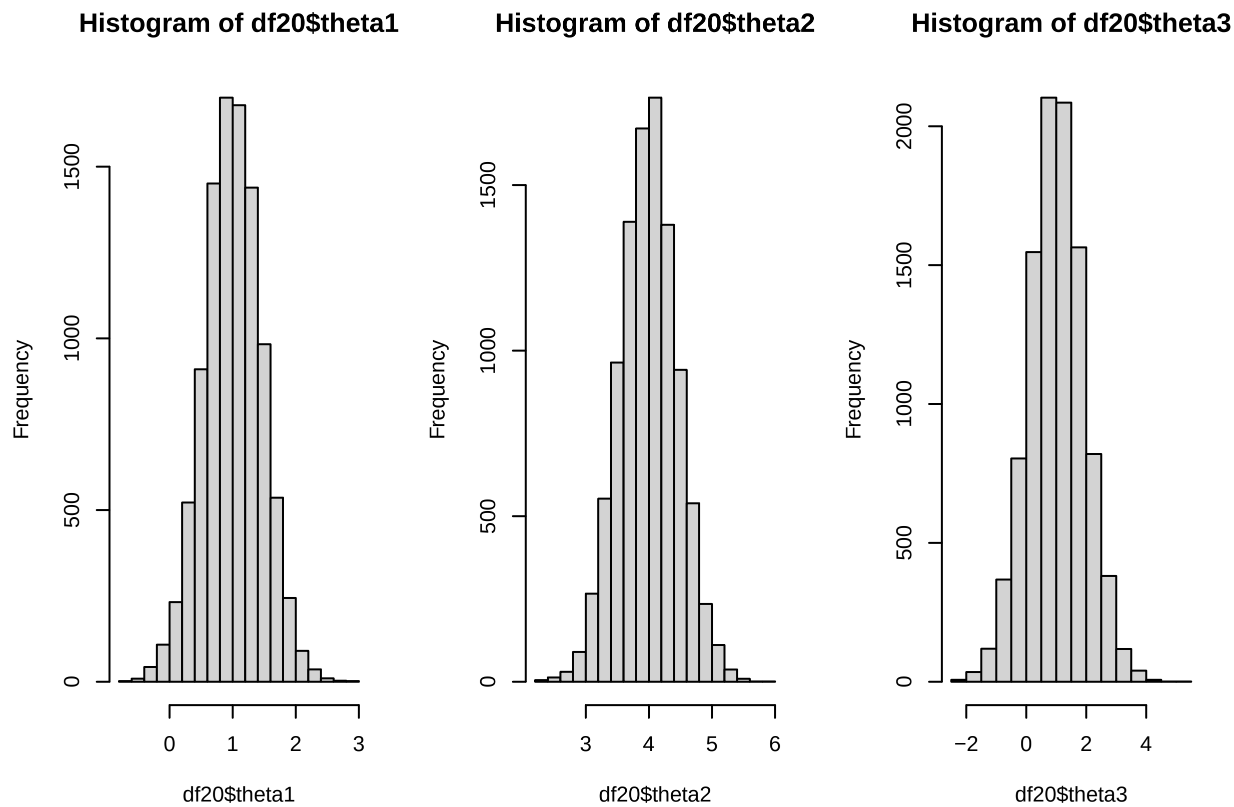
zadanie 1c) n=100




```
## [[1]]
##     theta1     theta2     theta3     theta4
## 0.04025848 0.06235408 0.04691779 0.96316656
##
## [[2]]
##     theta1     theta2     theta3     theta4
## 0.04025095 0.06235113 0.04690150 0.01975347
##
## [[3]]
##      theta1      theta2      theta3      theta4
## 0.002744222 0.001716548 0.004035922 0.971294548
```

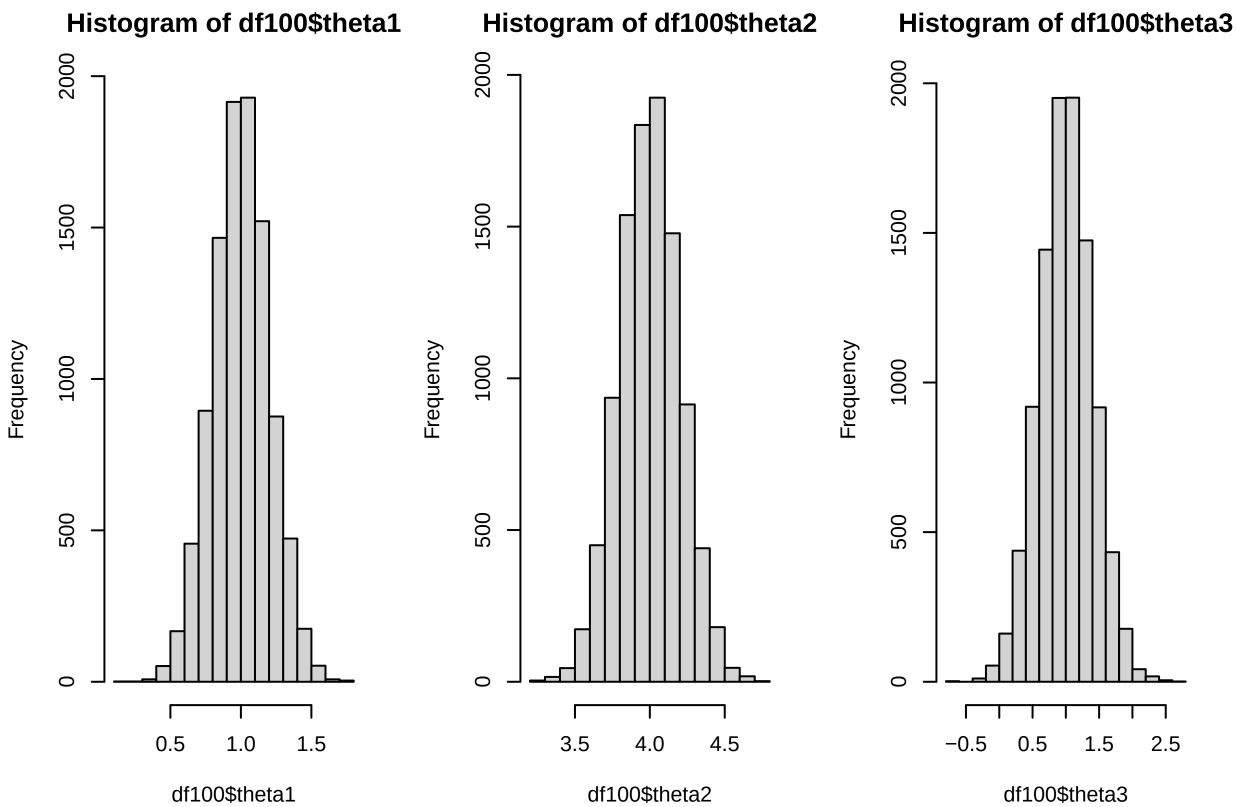
Wnioski zebrane: Im większy rozmiar próbki, tym dane bardziej zachowują się w sposób, w jaki przewidujemy korzystając z teorii, minimalizowane jest MSE, var oraz bias.

zadanie 5 n=20




```
##         MSE      var         bias
## 1 2.072821 2.072729  0.003029093
## 2 2.097284 2.097251 -0.001827109
## 3 8.572335 8.571962  0.006106124
```

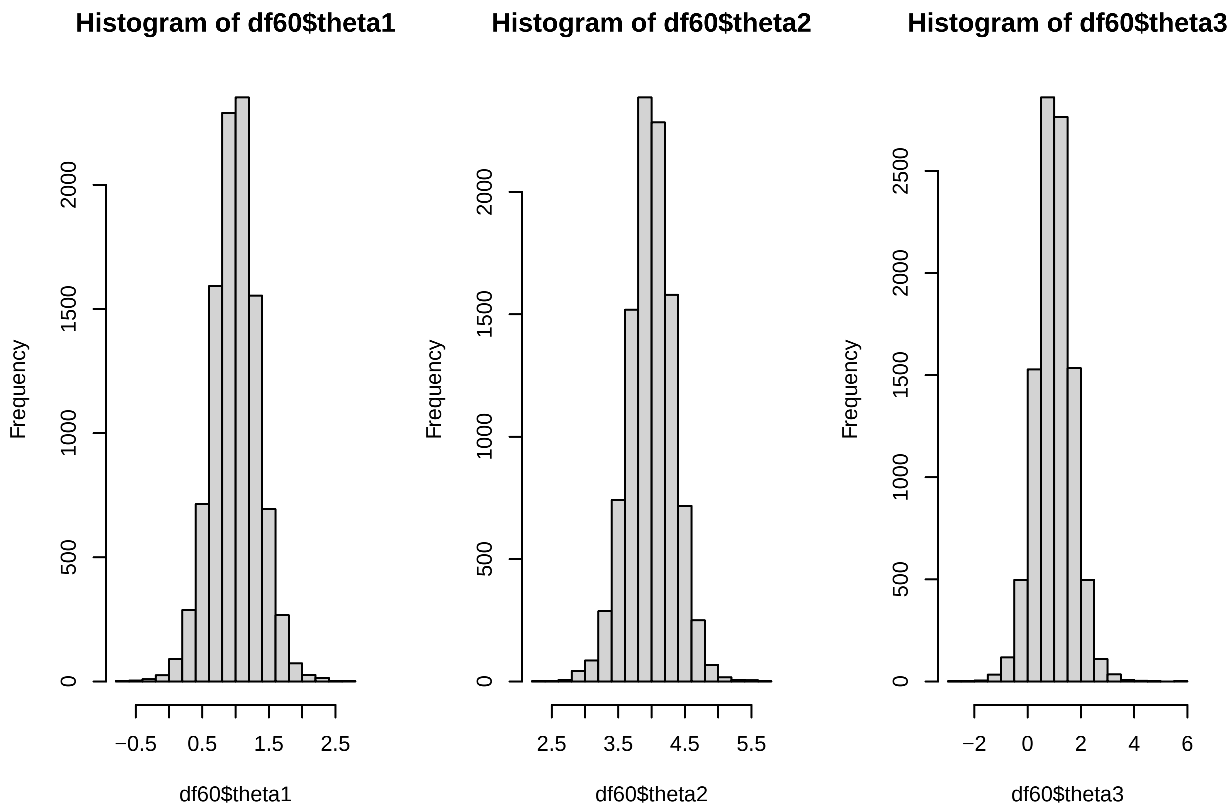
zadanie 5 n=100




```
##         MSE       var          bias
## 1 0.4026076 0.4025843  0.0015252656
## 2 0.4103320 0.4103264 -0.0007475108
## 3 1.6133788 1.6133323  0.0021564528
```

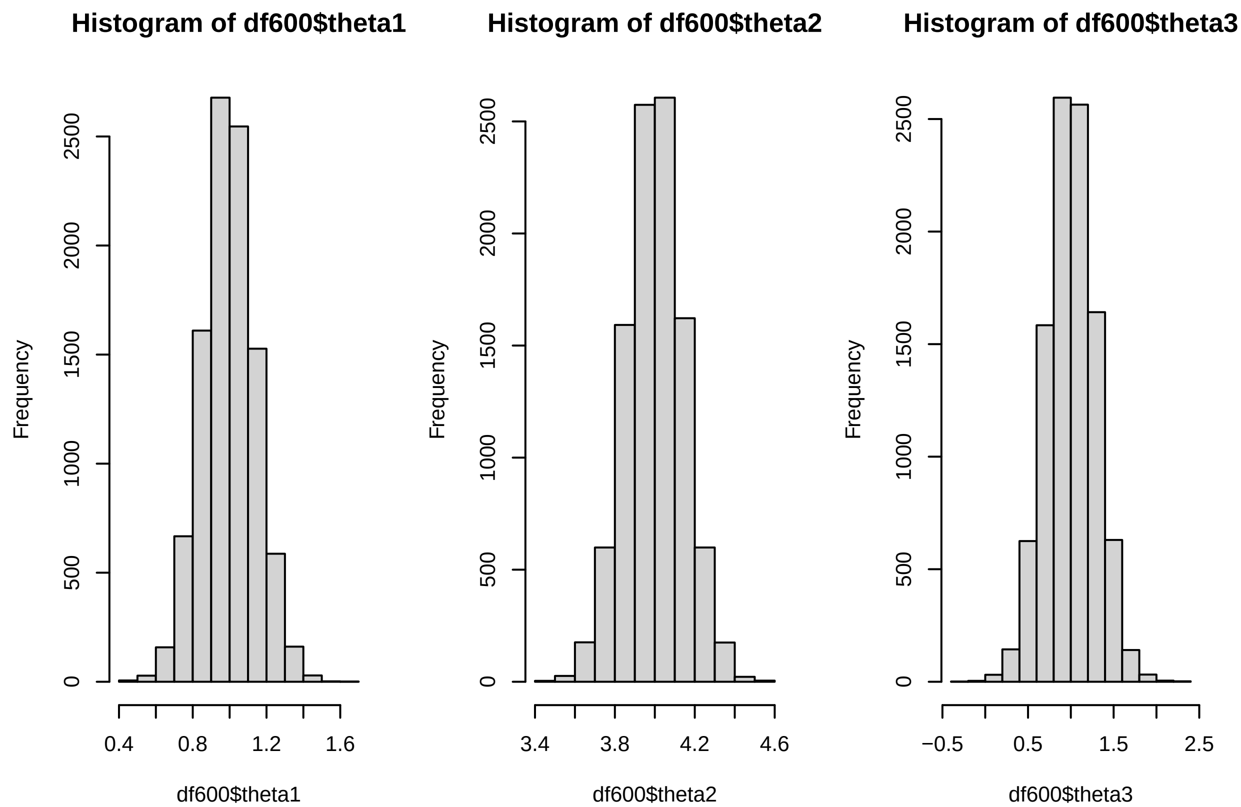
Wnioski: jak w zadaniu 1, im większe n, tym MSE, var i bias są bliższe 0.

zadanie 6 n=20




```
##         MSE      var          bias
## 1 1.174207 1.174060 -0.0038455581
## 2 1.159868 1.159421 -0.0066863895
## 3 4.695533 4.695528  0.0006969934
```

zadanie 6 n=100




```
##         MSE       var          bias
## 1 0.2070163 0.2068782 -0.0037163412
## 2 0.2073195 0.2073125  0.0008319452
## 3 0.8319905 0.8319384  0.0022821608
```

Wnioski: Również i w tym przypadku większa próbka skutkowała mniejszymi wartościami MLE i var. Przy biasie nie widać specjalnej różnicy.

Większa próbka sprawia, że mniejsza jej część to dane mocno zaburzone.

Rachunki

Rozkład logistyczny:

$$f(x) = \frac{e^{-(x-\theta)/\sigma}}{\sigma(1+e^{-(x-\theta)/\sigma})^2}$$

$$l(\theta) = \sum_{i=1}^{n} logf(x_i,\theta) = \sum_{i=1}^{n} \frac{-(x_i-\theta)}{\sigma} - (log\sigma + 2log(1+e^{-(x-\theta)/\sigma}) \tag{1}$$

$$= \frac{-n(\bar{x}-\theta)}{\sigma} - nlog\sigma - 2\sum_{i=1}^{n} log(1+e^{-(x-\theta)/\sigma}) \tag{2}$$

$$l'(\theta) = \frac{n}{\sigma} - \frac{2}{\sigma}\sum_{i=1}^{n} \frac{e^{-(x-\theta)/\sigma}}{1+e^{-(x-\theta)/\sigma}} \tag{3}$$

$$l''(\theta) = \frac{-2}{\sigma}\sum_{i=1}^{n}\frac{\frac{-1}{\sigma}e^{-(x-\theta)/\sigma}(e^{-(x-\theta)/\sigma}+1)-(\frac{-1}{\sigma}e^{-(x-\theta)/\sigma}e^{-(x-\theta)/\sigma})}{(\frac{-1}{\sigma}+1)^2} \quad (4)$$

$$= \frac{-2}{\sigma}\sum_{i=1}^{n}\frac{e^{-(x-\theta)/\sigma}}{(e^{-(x-\theta)/\sigma}+1)^2} \quad (5)$$

Rozkład Cauchy'ego:

$$f(x) = \frac{1}{\pi\sigma(1+\frac{x-\theta}{\sigma})^2}$$

$$l(\theta) = \sum_{i=1}^{n} log f(x_i,\theta) = \sum_{i=1}^{n} log\frac{1}{\pi\sigma(1+\frac{x_i-\theta}{\sigma})^2} = -\sum_{i=1}^{n} log(\pi\sigma(1+\frac{x_i-\theta}{\sigma})^2) \quad (6)$$

$$= -\sum_{i=1}^{n} log(\pi\sigma)+log((1+\frac{x-\theta}{\sigma})^2) = -nlog(\pi\sigma)-\sum_{i=1}^{n} log((1+\frac{x-\theta}{\sigma})^2) \quad (7)$$

$$l'(\theta) = -\sum_{i=1}^{n}\frac{\frac{-2(x-\theta)}{\sigma^2}}{(1+\frac{x-\theta}{\sigma})^2} = \frac{2}{\sigma}\sum_{i=1}^{n}\frac{\frac{x-\theta}{\sigma}}{1+(\frac{x-\theta}{\sigma})^2} \quad (8)$$

$$l''(\theta) = \frac{2}{\sigma}\sum_{i=1}^{n}\frac{\frac{-1}{\sigma}((1+(\frac{x-\theta}{\sigma})^2-2\frac{x-\theta}{\sigma}\frac{x-\theta}{\sigma})}{(1+(\frac{x-\theta}{\sigma})^2)^2} = \frac{-2}{\sigma^2}\sum_{i=1}^{n}\frac{1-(\frac{x-\theta}{\sigma})^2}{(1+(\frac{x-\theta}{\sigma})^2)^2} \quad (9)$$