Zadanie 1. W pierwszym kroku generujemy 50 obserwacji z rozkładu $N(1, 1)$. Następnie na ich podstawie obliczamy wartości estymatorów parametru θ postaci

(i) $\hat{\theta}_1 = \bar{X} = (1/n) \sum_{i=1}^{n} X_i$,

```
theta_1 <- function(n, mean, sd) return(sum(rnorm(n, mean, sd))/n)
```

(ii) $\hat{\theta}_2 = \mathrm{Me}\{X_1, \ldots, X_n\}$,

```
theta_2 <- function(n, mean, sd) return(median(rnorm(n, mean,
                                                     sd)))
```

(iii) $\hat{\theta}_3 = \sum_{i=1}^{n} w_i X_i$, $\sum_{i=1}^{n} w_i = 1$, $0 \leqslant w_i \leqslant 1$, $i = 1, \ldots, n$, z losowym wyborem wag,

```
theta_3 <- function(n, mean, sd) {
    unnormed_weights <- runif(n, 0, 1)
    weights <- unnormed_weights/sum(unnormed_weights)
    return(sum(rnorm(n, mean, sd) * weights))
}
```

(iv) $\hat{\theta}_4 = \sum_{i=1}^{n} w_i X_{i:n}$, gdzie $X_{1:n} \leqslant \cdots \leqslant X_{n:n}$ są uporządkowanymi obserwacjami $X_1, \ldots, X_n$,

$$w_i = \varphi(\Phi^{-1}(\frac{i-1}{n})) - \varphi(\Phi^{-1}(\frac{i}{n})),$$

przy czym φ jest gęstością. a Φ dystrybuantą standardowego rozkładu normalnego $N(0, 1)$

```
theta_4 <- function(n, mean, sd) {
    weights <- sapply(c(1:n),
                      function(i) dnorm(qnorm((i-1)/n)) -
                      dnorm(qnorm(i/n)))
    return(sum(rnorm(n, mean, sd) * weights))
}
```

W wyniku otrzymujemy następujące wartości

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
0.936	1.225	1.072	0.078

Na podstawie uzyskanych wyników możemy zauważyć, że wartości estymatorów $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ są zbliżone do 1 tj. prawdziwej wartości parametru θ, natomiast wartość estymatora $\hat{\theta}_4$ znacznie odbiega od 1.

Powtarzamy powyższe doświadczenie dla rozkładu $N(4,1)$. W wyniku otrzymujemy następujące wartości

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
4.069	3.645	4.073	0.266

Widzimy, że podobnie jak w przypadku rozkładu $N(1,1)$, dla rozkładu $N(4,1)$ najlepsze przybliżenia θ otrzymujemy dla estymatorów $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$. Dla estymatora $\hat{\theta}_4$ wyliczona wartość znacznie różni się od 4.

Powtarzamy wcześniejsze doświadczenie dla rozkładu $N(1,2)$. W wyniku otrzymujemy następujące wartości

$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
1.13	0.86	0.838	-0.045

Analogicznie do wcześniejszych doświadczeń, dla rozkładu $N(1,2)$ najlepsze przybliżenia θ są osiągane przez estymatory $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$, zaś wartość estymatora $\hat{\theta}_4$ jest znacznie odbiega od 1.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów.

```
stats <- function(n, mean, sd, fun) {
  trials <- rep(NA, 10000)
  for (i in 1:10000) {
    trials[i] = fun(n, mean, sd)
  }
  return(c(var(trials), sum((trials - rep(mean, 10000))^2)/10000,
           mean(trials) - mean))
}


stats_for_estim <- function(x) return(data.frame(stats(50, 1, 1, x),
                                                 stats(50, 4, 1, x),
                                                 stats(50, 1, 2, x)))
```

Dla estymatora $\hat{\theta}_1$ otrzymujemy następujące statystyki

	$N(1,1)$	$N(4,1)$	$N(1,2)$
wariancja	0.020216	0.019764	0.081134
błąd średniokwadratowy	0.020217	0.019764	0.081158
obciążenie	-0.001769	0.001403	0.005640

Dla estymatora $\hat{\theta}_2$ otrzymane statystyki wynoszą

	$N(1,1)$	$N(4,1)$	$N(1,2)$
wariancja	0.030796	0.031619	0.120086
błąd średniokwadratowy	0.030793	0.031615	0.120125
obciążenie	0.000253	-0.000141	0.007091

Natomiast dla estymatora $\hat{\theta}_3$ wartości statystyk są równe

	$N(1,1)$	$N(4,1)$	$N(1,2)$
wariancja	0.026560	0.026731	0.108723
błąd średniokwadratowy	0.026557	0.026729	0.108725
obciążenie	0.000030	0.000653	0.003520

Widzimy zatem, że w przypadku estymatorów $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ obciążenie jest bliskie 0.

Ponadto możemy zaobserwować, że w przypadku rozkładów $N(1,1)$ oraz $N(4,1)$, dla każdego z powyższych estymatorów wariancja, podobnie jak MSE wynosi w przybliżeniu co najwyżej 0.03, a zatem zróżnicowanie wyników jest stosunkowo małe oraz są one bliskie wartości estymowanemu parametrowi θ. Jedynie dla rozkładu $N(1,2)$ wyniki są bardziej rozproszone wokół średniej, gdyż w tym przypadku wariancja osiąga wartość 0.12. Możemy również stwierdzić, że dla tego rozkładu wartości estymatora są bardziej oddalone od wartości θ, ponieważ MSE wynosi w przybliżeniu 0.11.

Przejdziemy do analizy statystyk otrzymanych dla estymatora θ_4, które wynoszą

	$N(1,1)$	$N(4,1)$	$N(1,2)$
wariancja	0.020115	0.020244	0.079853
błąd średniokwadratowy	1.024498	16.017924	1.086911
obciążenie	-1.002190	-3.999710	-1.003527

Widzimy zatem, że dla każdego z rozkładów $\hat{\theta}_4$ jest obciążony. Dodatkowo jego wartości są istotnie rozproszone oraz znacznie oddalone od prawdziwej wartości estymowanego parametru θ. Jako wniosek otrzymujemy więc, że estymator $\hat{\theta}_4$ nie stanowi dobrego przybliżenia θ.

Zadanie 5. W tym zadaniu rozważać będziemy rozkład logistyczny $L(\theta, \sigma)$. Naszym celem będzie oszacowanie wartości estymatora największej wiarygodności parametru θ na podstawie generowanych przez nas prób. W tym celu korzystać będziemy z metody Newtona.

Przypomnijmy, że rozkład logistyczny jest zbliżony do rozkładu normalnego, dlatego na wejściu przyjmujemy średnią z próby. Zakładamy, że zatrzymanie algorytmu następuje wtedy i tylko wtedy, gdy odległość pomiędzy kolejnymi przybliżeniami jest mniejsza niż 10^{-7}

```
MLE_for_logistic <- function(x, sigma, theta = mean(x),
                             epsilon = .0000001, stop = FALSE) {
  while(stop == FALSE) {
    first_deriv <- length(x)/sigma - 2*sum((exp((-(x-theta))/sigma))/
    (sigma*(1+exp((-(x-theta))/sigma))))
    sec_deriv <- -2*sum((exp((-(x-theta))/sigma))/
                 (sigma^2*(1+exp((-(x-theta))/sigma))^2))
    theta = theta - first_deriv/sec_deriv
    stop = abs(first_deriv/sec_deriv) < epsilon
  }
  return(theta)
}
```

W pierwszym kroku generujemy 50 obserwacji pochodzących kolejno z $L(1,1)$, $L(4,1)$, $L(1,2)$. Estymatory $\hat{\theta}$ obliczone w opisany powyżej sposób są wówczas równe

$L(1,1)$	$L(4,1)$	$L(1,2)$
1.063	3.982	0.849

Możemy zauważyć, że w przypadku każdego z badanych rozkładów różnica między wartością estymatora $\hat{\theta}$, a rzeczywistą wartością parametru θ nie przekracza 0.151, zatem uzyskane przez nas przybliżenie jest dostateczne.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów.

```
stats_for_logistic <- function(n, theta, sigma) {
  trials <- rep(NA, 10000)
  for (i in 1:10000) {
    trials[i] = MLE_for_logistic(rlogis(n, theta, sigma), sigma)
  }
  return(c(var(trials), sum((trials - theta)^2)/10000,
           mean(trials) - theta))
}
```

W wyniku otrzymujemy następujące statystyki

	$L(1,1)$	$L(4,1)$	$L(1,2)$
wariancja	0.061209	0.060207	0.243880
błąd średniokwadratowy	0.061207	0.060201	0.243928
obciążenie	-0.001901	-0.000480	0.008537

Zauważmy, że na podstawie otrzymanych statystyk dla rozkładów $L(1,1), L(4,1)$ możemy wnioskować o skuteczności przyjętej przez nas metody estymacji θ. Istotnie, wariancja oraz MSE nie przekraczają 0.0613, zatem zróżnicowanie wyników jest stosunkowo niskie oraz bliskie są one wartości estymowanego parametru θ. Również obciążenie jest nieznaczne.

Dla rozkładu $L(1,2)$ wariancja i MSE osiągają wartość 0.244, co jest istotnym wynikiem biorąc pod uwagę rzeczywistą wartość paramteru θ, a zatem w tym przypadku zróżnicowanie wyników, jak i ich odległości od 1 są znaczące.

Zadanie 6. W tym zadaniu będziemy się zajmować rozkładem Cauchy'ego $C(\theta, \sigma)$ z parametrem przesunięcia θ i skali σ. Naszym celem będzie oszacowanie wartości estymatora największej wiarygodności parametru θ na podstawie generowanych przez nas prób. Podobnie jak poprzednio wykorzystywać będziemy metodę Newtona z wartością wejściową równą medianie obserwacji. Zakładać będziemy, że zatrzymanie algorytmu następować będzie wtedy i tylko wtedy, gdy odległość pomiędzy kolejnymi przybliżeniami będzie mniejsza niż 10^{-4}.

```
MLE_for_cauchy <- function(x, sigma, theta0 = median(x),
                           epsilon = 0.0001, stop = FALSE) {
  while(stop == FALSE) {
    first_deriv <- 2*sum(((x-theta0)/sigma)/(1+((x-theta0)/sigma)^2))
    sec_deriv <- 2*sum((sigma*(x - theta0^2)-sigma^3)/
    (            sigma^2 + (x - theta0)^2))
    theta0 = theta0 - first_deriv/sec_deriv
    stop = abs(first_deriv/sec_deriv) < epsilon
  }
  return(theta0)
}
```

Zaczniemy od doświadczenia, w którym wygenerujemy 50 obserwacji pochodzących kolejno z $C(1,1)$, $C(4,1)$, $C(1,2)$. Estymatory $\hat{\theta}$ obliczone w przedstawiony powyżej sposób są wówczas równe

$C(1,1)$	$C(4,1)$	$C(1,2)$
1.171	3.978	1.226

Możemy zauważyć, że w przypadku rozkładu Cauchy'ego przybliżanie estymatora największej wiarygodności metodą Newtona bardziej odbiega od wartości parametru θ, niż w przypadku rozkładu logistycznego. Dla rozkładu Cauchy'ego różnica ta jest względnie duża, przykładowo dla $C(1,2)$ przekracza 0.2.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów.

```
stats_for_cauchy <- function(n, theta, sigma) {
  trials <- rep(NA, 10000)
  for (i in 1:10000) {
```

```
    trials[i] = MLE_for_cauchy(rcauchy(n, theta, sigma), sigma)
  }
  return(c(var(trials), sum((trials - theta)^2)/10000,
           mean(trials) - theta))
}
```

W wyniku otrzymujemy następujące statystyki

	$L(1,1)$	$L(4,1)$	$L(1,2)$
wariancja	0.042561	0.042202	0.170103
błąd średniokwadratowy	0.042557	0.042198	0.170086
obciążenie	0.000474	-0.000794	0.000186

Możemy zatem zauważyć, że w przypadku każdego z estymtorów, wartość oczekiwana jest bliska wartości odpowiedniego parametru θ, różnica nie przekracza 0.0008. W przypadku rozkładów $L(1,1)$ oraz $L(4,1)$ wartości przyjmowane przez estymatory są słabo rozproszone oraz ich średnia odległość od rzeczywistej wartości paramteru θ jest niewielka. Dla rozkładu $L(1,2)$ wspomniane wcześniej wartości są ponad trzykrotnie wyższe, co stanowi już sporą różnicę.

Zadanie 7. Przejdziemy teraz do badania wpływu wielkości próby na wyniki estymacji parametrów. Zaczniemy od powtórzenia części Zadania 1., tj. zbadamy dokładoność estymacji parametru θ przy pomocy omawianych wcześniej estymatorów $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ oraz $\hat{\theta}_4$ wykorzystując kolejno 20- oraz 100-elementową próbę z rozkładu $N(1,1)$. W poniższej tabeli prezentujemy otrzymane wyniki.

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
20 obserwacji	0.719	1.289	0.605	0.257
100 obserwacji	1.096	0.962	0.956	0.123

Na podstawie powyższych wyników obserwujemy, że dla większej próby estymatory $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ są bliższe 1, niż dla próby 20-elementowej. Możemy zatem przypuszczać, że poprawność estymacji wzrasta wraz ze wzrostem liczebności próby.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, co pozwoli nam na oszacowanie wariancji, błędu średniokwadratowego oraz obciążenia każdego z estymatorów. Dla próby 20-elementowej otrzymujemy następujące statystyki

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
wariancja	0.050432	0.074365	0.066382	0.048502
błąd średniokwadratowy	0.050432	0.074358	0.066386	1.047384
obciążenie	0.002317	0.000391	0.003402	-0.999443

Zaś dla próby 100-elementowej otrzymane statystyki wynoszą

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
wariancja	0.009848	0.015635	0.013581	0.010066
błąd średniokwadratowy	0.009849	0.015633	0.013581	1.008790
obciążenie	-0.001500	-0.000164	0.001172	-0.999362

Możemy zatem zauważyć, że dla większej próby wartości estymatorów są mniej rozproszone i bardziej zbliżone do rzeczywistej wartości estymowanego parametru θ. Dodatkowo obciążenie estymatorów w przypadku próby 100-elementowej jest mniejsze niż dla mniejszej próby. Możemy więc przypuszczać asymptotyczną nieobciążoność estymatorów $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$. Widzimy zatem, że wraz ze wzrostem liczby obserwacji wzrasta poprawność estymacji.

Przejdziemy teraz do powtórzenia powyższej procedury dla rozkładu logistycznego $L(1, 1,)$. W poniższej tabeli prezentujemy otrzymane wyniki.

20 obserwacji	100 obserwacji
1.325	1.028

Powtarzamy doświadczenie 10 000 razy. Otrzymane statystyki wynoszą

	20 obserwacji	100 obserwacji
wariancja	0.148939	0.030154
błąd średniokwadratowy	0.148926	0.030152
obciążenie	-0.001497	-0.000616

Widzimy, że w przypadku rozkładu logistycznego, podobnie jak dla wyżej omawianego rozkładu normalnego, wraz ze zwiększeniem liczebności próby idzie zwiększenie poprawności estymacji.

W kolejnym kroku rozpatrywać będziemy rozkład Cauchy'ego $C(1, 1)$. Znów zaczniemy od oszacowania wartości estymatora największej wiarygodności parametru θ na podstawie wygenerowanej próby 20- oraz 100-elementowej. Uzyskane wyniki prezentujemy w poniższej tabeli

20 obserwacji	100 obserwacji
0.703	1.213

Możemy zauważyć, że podobnie jak we wcześniejszych przykładach, dla większej liczebności próby, wartość szacowanego estymatora parametru największej wiarygodności jest bliższa rzeczywistej wartości parametru θ.

W kolejnym kroku powtarzamy powyższe doświadczenie 10 000 razy, aby obliczyć wariancję, błąd średniokwadratowy oraz obciążenie dla wykonywanego przez nas oszacowania. W poniższej tabeli przedstawiamy uzyskane wyniki

	20 obserwacji	100 obserwacji
wariancja	0.116850	0.020771
błąd średniokwadratowy	0.116839	0.020769
obciążenie	0.000372	0.000852

Widzimy zatem, że anlogicznie do wcześniejszych wyników, dla większej liczebności próby, wartość oszacowania estymatora największej wiarygodności parametru θ jest bliższa rzeczywistej wartości parametru θ.