Projekt 1

Wstęp do Analizy Danych | Politechnika Krakowska

Jakub Kapała

Numer albumu: 151885

Data: 27.05.2025

Treść projektu

Załóżmy, że zebraliśmy próbkę zarobków 10 studentów, którzy niedawno ukończyli studia.

Dane:

`45617 7166 18594 2236 1278 19828 4033 28151 2414 3800`

Załóżmy, że zarobki mają rozkład normalny z nieznaną średnią populacji i ze standardowym odchyleniem $15,000$. Dokonaj estymacji średniej zarobków studentów, którzy niedawno ukończyli studia. Wylicz przedziały ufności: 90% i 95% (poziom istotności $\alpha = 0.1$ i $\alpha = 0.05$).

Powtórzmy a) z jedną różnicą: standardowe odchylenie nie jest znane i musimy dokonać estymacji standardowego odchylenia używając próbki. (Użyj t-test do estymacji średniej).

Tym razem nie zakładamy, że zarobki mają rozkład normalny. Pokaż histogram oraz wykres kwantyl-kwantyl i skomentuj. Przedziały ufności wyznaczone powyżej są tylko aproksymacją. (W środowisku statystycznym R, QQ plot, funkcja `plot()`)

Użyj metody bootstrap do konstrukcji przedziałów ufności:

Bootstrap method:

1. From our sample of size 10, draw a new sample, with replacement, of size 10.
2. Compute the sample average, which we call the bootstrap estimate.
3. Record it.
4. Repeat steps 1 to 3, 1000 times.
5. For a 90% confidence, we will use the 5% sample quantile as the lower bound, and the 95% sample quantile as the upper bound (alpha = 10%, so alpha/2 = 5%). Construct 90% and 95% bootstrap intervals.

Dodatkowo:

Wybierz trzy dowolne dane próbkowe z biblioteki `MASS`. Opisz wybrane przez Ciebie dane. Zbadaj normalność próbki. Oblicz przedziały ufności dla średniej, standardowego odchylenia oraz wariancji.

1. Analiza zarobków studentów

1.1. Estymacja zarobków, wyliczenie przedziałów ufności

Zakładam, że zarobki mają rozkład normalny z nieznaną średnią populacji i ze standardowym odchyleniem $15,000$. Dokonuje zatem estymacji średniej zarobków studentów, którzy niedawno ukończyli studia.

Ustawiam ziarno losowania na swój numer albumu, aby uzyskać powtarzalne wyniki:

```
set.seed(151885)
```

Estymuje średnią zarobków studentów,

```
# Deklaracja danych - próbka, liczba obserwacji, odchylenie standardowe
data_sample <- c(45617, 7166, 18594, 2236, 1278, 19828, 4033, 28151, 2414, 3800)
sd <- 15000
n <- 10
# Estymacja średniej zarobków
mean_salary <- mean(data_sample)
```

Następnie obliczam przedziały ufności dla 90 i 95 (poziom istotności $\alpha = 0.1$ i $\alpha = 0.05$):

```
# Poziom istotności
alpha_90 <- 0.1
alpha_95 <- 0.05

# Z-scores z rozkładu normalnego
z_90 <- qnorm(1 - alpha_90 / 2)
z_95 <- qnorm(1 - alpha_95 / 2)

# Granice przedziałów ufności
lower_bound_90 <- mean_salary - z_90 * (sd / sqrt(n))
upper_bound_90 <- mean_salary + z_90 * (sd / sqrt(n))

lower_bound_95 <- mean_salary - z_95 * (sd / sqrt(n))
upper_bound_95 <- mean_salary + z_95 * (sd / sqrt(n))

# Wyliczenie przedziałów ufności
ci_90 <- c(lower_bound_90, upper_bound_90)
ci_95 <- c(lower_bound_95, upper_bound_95)
```

Wyniki:

```
## Estymowana średnia zarobków: 13311.7

## Przedział ufności 90%: [ 5509.47 ,  21113.93 ]

## Przedział ufności 95%: [ 4014.77 ,  22608.63 ]
```

1.2. Estymacja odchylenia standardowego i średniej zarobków z próbki, przybliżenie przedziałów ufności

Tym razem nie zakładam, że zarobki mają rozkład normalny. Nie znam także średniej populacji ani odchylenia standardowego. Dokonuję estymacji odchylenia standardowego z próbki:

```r
# Deklaracja danych - próbka, liczba obserwacji
data_sample <- c(45617, 7166, 18594, 2236, 1278, 19828, 4033, 28151, 2414, 3800)
n <- 10
# Estymacja odchylenia standardowego z próbki
sd <- sd(data_sample)
```

Używam testu t do estymacji średniej zarobków oraz przybliżenia przedziałów ufności 95% i 90%.

```r
t_test_95 <- t.test(data_sample, conf.level = 0.95)
t_test_90 <- t.test(data_sample, conf.level = 0.90)
```

Wyniki estymacji:

```
## Estymowane odchylenie standardowe: 14662.04
```

```
## Estymowana średnia zarobków: 13311.7
```

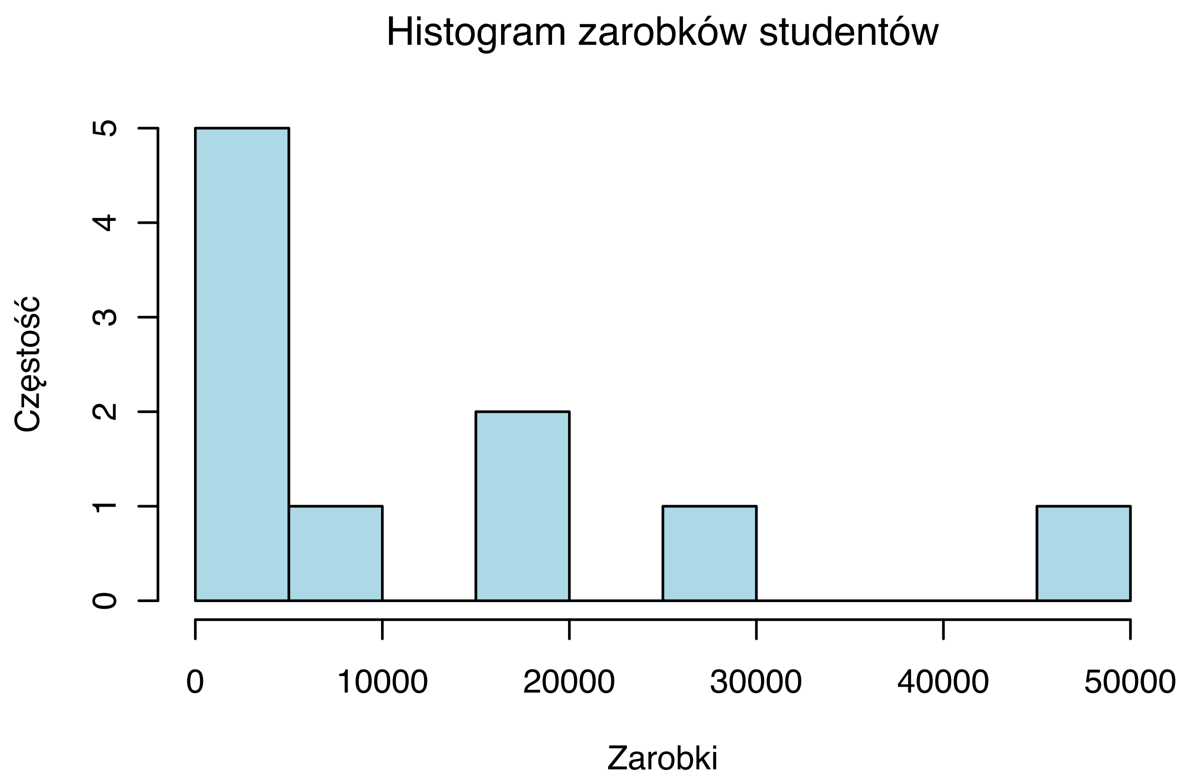
```
## Przedział ufności 90%: [ 4812.39 ,  21811.01 ]
```

```
## Przedział ufności 95%: [ 2823.11 ,  23800.29 ]
```

Jak widzimy, estymowana średnia zarobków przy pomocy t-test nie różni się od estymowanej średniej zarobków w podpunkcie a, natomiast odchylenie standardowe i przedziały ufności są inne.

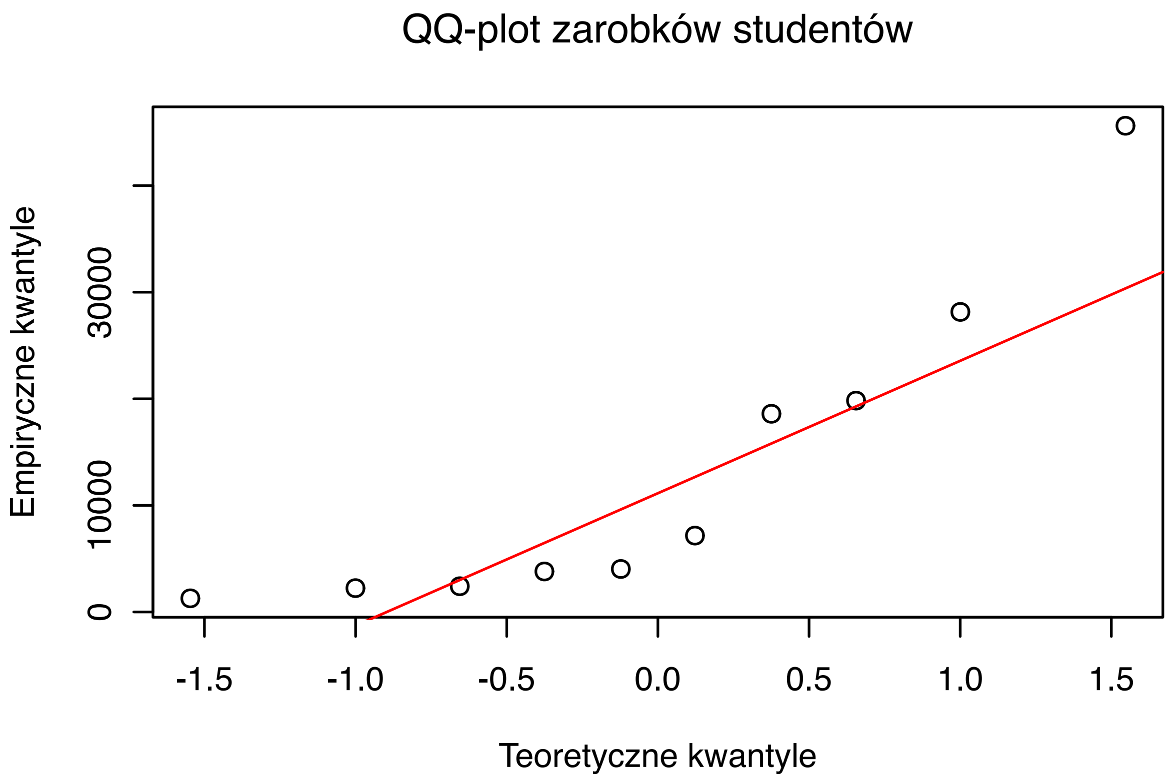
Teraz spójrzmy na histogram oraz wykres kwantyl-kwantyl (QQ-plot).

```r
# Histogram
hist(data_sample,
     main = "Histogram zarobków studentów",
     xlab = "Zarobki",
     ylab = "Częstość",
     col = "lightblue",
     border = "black",
     breaks = 7)
```



Na histogramie możemy zauważyć, że rozkład zarobków studentów jest asymetryczny prawostronny (prawostronnie skośny), co sugeruje, że zarobki są bardziej skoncentrowane w niższych wartościach - zwłaszcza w przedziale $[0, 5000]$.

```
# QQ-plot
qqnorm(data_sample,
     main = "QQ-plot zarobków studentów",
     xlab = "Teoretyczne kwantyle",
     ylab = "Empiryczne kwantyle")
qqline(data_sample, col = "red")
```



Wykres kwantyl-kwantyl (QQ-plot) pokazuje, że dane nie są zgodne z rozkładem normalnym - widać na nim odchylenia od linii prostej. Oznacza to, że przedziały ufności wyznaczone poprzez t-test są jedynie aproksymacją.

1.3. Użycie metody bootstrap do konstrukcji przedziałów ufności

Używam metody bootstrap do konstrukcji przedziałów ufności. Najpierw deklaruje zmienne:

```r
# Deklaracja danych - próbka, liczba obserwacji, liczba bootstrapów
data_sample <- c(45617, 7166, 18594, 2236, 1278, 19828, 4033, 28151, 2414, 3800)
n <- 10
n_bootstrap <- 1000
# Inicjalizacja wektora do przechowywania wyników bootstrap
bootstrap_means <- numeric(n_bootstrap)
```

Następnie wykonuję pętle bootstrap 1000 razy - losuję nową próbke z powtórzeniami o rozmiarze 10, obliczam średnią z tej próbki i zapisuje ją do wektora:

```r
for (i in 1:n_bootstrap) {
  bootstrap_sample <- sample(data_sample, size = n, replace = TRUE)
  bootstrap_means[i] <- mean(bootstrap_sample)
}
```

Wyznaczam teraz przedziały ufności 90% i 95%:

```r
ci_90_bootstrap <- quantile(bootstrap_means, probs = c(0.05, 0.95))
ci_95_bootstrap <- quantile(bootstrap_means, probs = c(0.025, 0.975))
```

Wyniki:

```
## Przedział ufności 90% (bootstrap): [ 6561.64 ,  20663.03 ]
```

```
## Przedział ufności 95% (bootstrap): [ 5475.19 ,  21592.79 ]
```

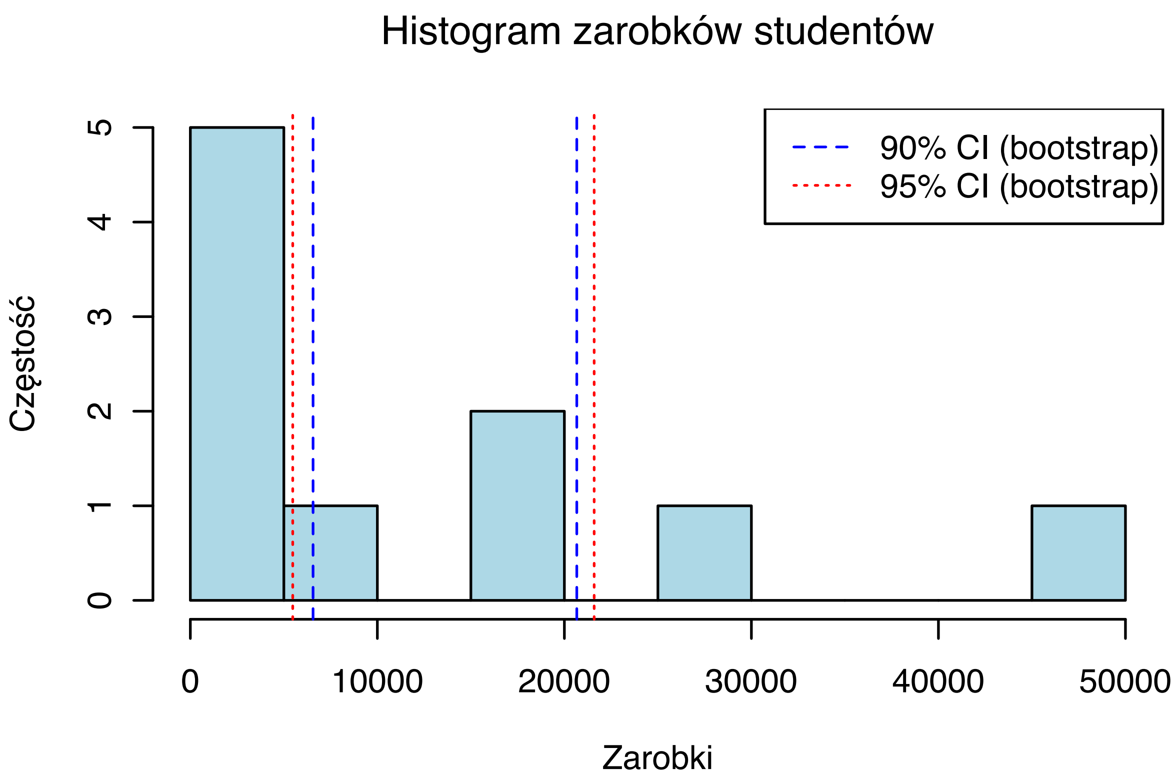
Na końcu porównam przedziały ufności uzyskane z metody bootstrap z tymi uzyskanymi z *t*-testu:

Table 1: Porównanie przedziałów ufności (bootstrap vs t-test)

Metoda	Poziom ufności	Dolna granica	Górna granica
Bootstrap	90%	6561.64	20663.03
T-test	90%	4812.39	21811.01
Bootstrap	95%	5475.19	21592.79
T-test	95%	2823.11	23800.29

1.4. Podsumowanie analizy zarobków studentów

Na podstawie przeprowadzonych analiz można zauważyć, że estymacja średniej zarobków studentów, którzy niedawno ukończyli studia, różni się w zależności od zastosowanej metody. Przedziały ufności uzyskane z metody bootstrap są węższe niż te uzyskane z t-testu, co sugeruje, że metoda bootstrap może być bardziej precyzyjna w tym przypadku. Histogram i wykres kwantyl-kwantyl pokazują, że dane nie są zgodne z rozkładem normalnym, co może wpływać na dokładność estymacji. Warto jednak zauważyć, że metoda bootstrap jest bardziej czasochłonna i wymaga większej mocy obliczeniowej, dlatego w praktyce często stosuje się t-test jako szybszą alternatywę.



2. Zbiór Cars93 z biblioteki `MASS`

Pierwszą próbką wybraną do badania normalności próbki, obliczania przedziały ufności dla średniej, standardowego odchylenia oraz wariancji, jest zbiór `Cars93` z biblioteki `MASS`. Zbiór ten zawiera dane dotyczące 93 samochodów osobowych sprzedawanych w USA w 1993 roku. Zawiera on różne cechy samochodów, takie jak cena, moc silnika, liczba miejsc, itp. W tej analizie skoncentruję się na wartości `Price` (cena samochodu).

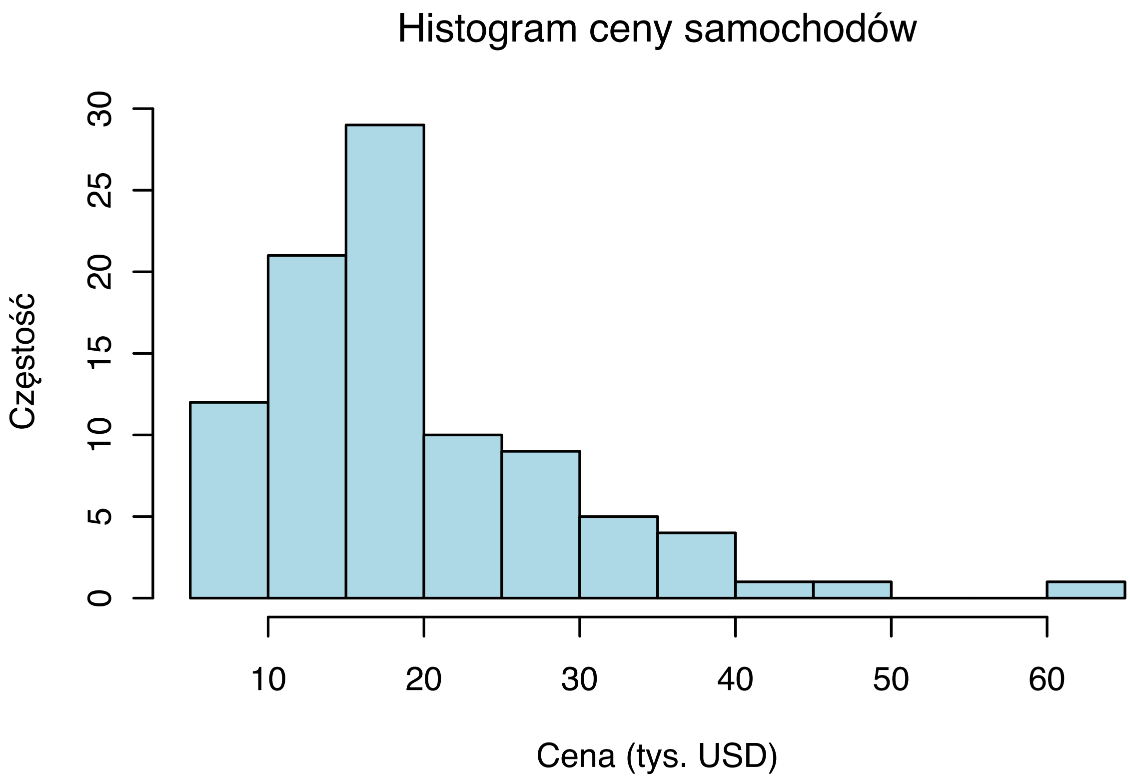
2.1. Analiza normalności próbki

Najpierw sprawdzę normalność rozkładu ceny samochodów (`Price`) w zbiorze `Cars93` przy użyciu histogramu, wykresu kwantyl-kwantyl oraz testu Shapiro-Wilka.

Wczytanie danych:

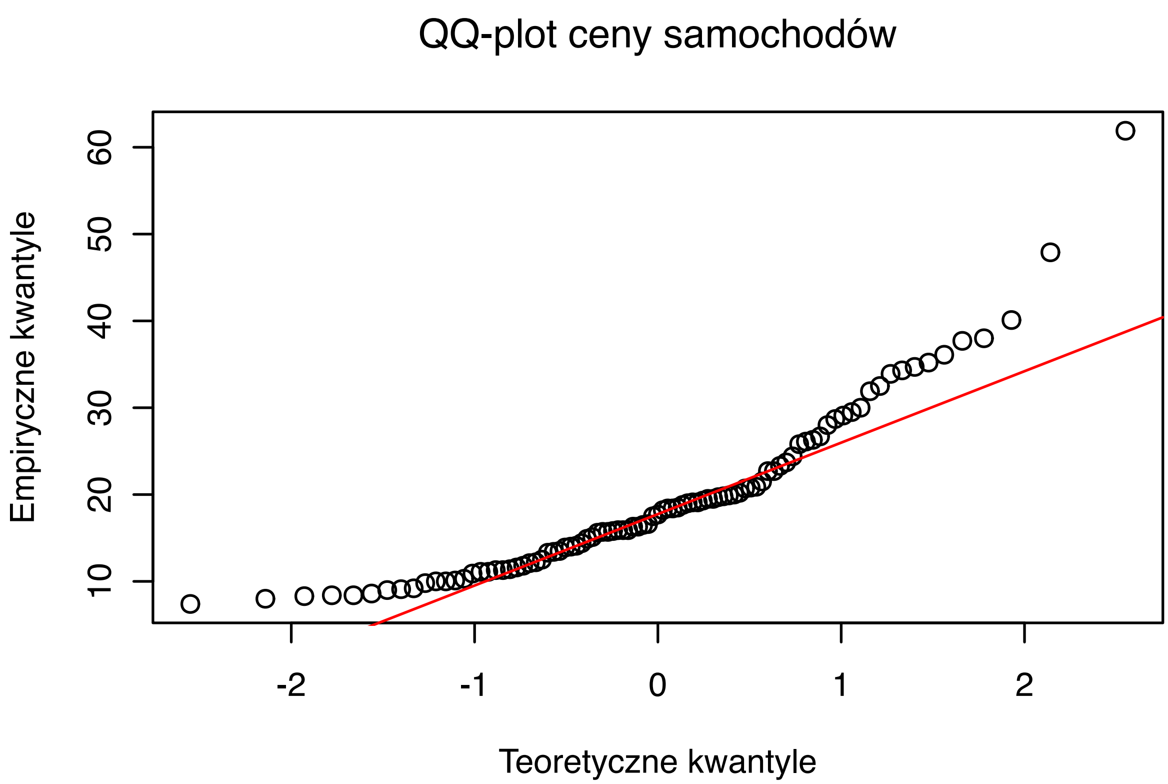
```r
library(MASS)
data(Cars93)
```

```r
# Histogram
hist(Cars93$Price,
     main = "Histogram ceny samochodów",
     xlab = "Cena (tys. USD)",
     ylab = "Częstość",
     col = "lightblue",
     border = "black")
```



Na histogramie możemy zauważyć, że rozkład ceny samochodów jest asymetryczny prawostronny (prawostronnie skośny), co sugeruje, że ceny są bardziej skoncentrowane w niższych wartościach - zwłaszcza w przedziale $[0, 20000]$ USD.

```r
# QQ-plot
qqnorm(Cars93$Price,
       main = "QQ-plot ceny samochodów",
       xlab = "Teoretyczne kwantyle",
       ylab = "Empiryczne kwantyle")
qqline(Cars93$Price, col = "red")
```



Wykres kwantyl-kwantyl (*QQ*-plot) pokazuje, że dane nie są zgodne z rozkładem normalnym - widać na nim znaczne odchylenia od linii prostej, zwłaszcza na końcach (ogony rozkładu).

Dodatkowo przeprowadzę *test Shapiro-Wilka*, aby ostatecznie ocenić normalność rozkładu ceny samochodów. Analiza *Monte Carlo* pokazała, że *test Shapiro-Wilka* ma największą moc spośród testów badających normalność rozkładu. Szczególnie polecany jest dla małych próbek - przy większych próbkach może być zbyt czuły i wykrywać nieistotne odchylenia od normalności. W naszym przypadku jest idealny, ponieważ nasz dataset ma 93 rekordy.

```r
cars_shapiro_test <- shapiro.test(Cars93$Price)
cars_shapiro_test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cars93$Price
## W = 0.88051, p-value = 4.235e-07
```

Wynik testu Shapiro-Wilka pokazuje, że wartość p jest znacznie mniejsza niż 0.05, co sugeruje, że rozkład ceny samochodów nie jest normalny. Wraz z histogramem i wykresem kwantyl-kwantyl jest to ostateczne potwierdzenie na niezgodność cen samochodów z rozkładem normalnym.

2.2. Obliczanie przedziałów ufności dla średniej, standardowego odchylenia oraz wariancji

Najpierw obliczę średnią i odchylenie standardowe, a następnie obliczę przedziały ufności korzystając z metody bootstrap, z uwagi na to, że rozkład ceny samochodów nie jest normalny.

```
cars_sd = sd(Cars93$Price)
cars_mean = mean(Cars93$Price)
# Ustawienie ziarna losowania
set.seed(151885)
# Deklaracja danych - próbka, liczba obserwacji, liczba pętli
data_sample <- Cars93$Price
n <- length(data_sample)
n_bootstrap <- 1000
# Inicjalizacja wektora do przechowywania wyników bootstrap
bootstrap_means <- numeric(n_bootstrap)
# Pętla bootstrap
for (i in 1:n_bootstrap) {
  bootstrap_sample <- sample(data_sample, size = n, replace = TRUE)
  bootstrap_means[i] <- mean(bootstrap_sample)
}
# Wyznaczanie przedziałów ufności
ci_90_bootstrap <- quantile(bootstrap_means, probs = c(0.05, 0.95), names = FALSE)
ci_95_bootstrap <- quantile(bootstrap_means, probs = c(0.025, 0.975), names = FALSE)
```

Table 2: Przedziały ufności cen samochodów ze zbioru Cars93

Metoda	Poziom ufności	Dolna granica (tys. USD)	Górna granica (tys. USD)
Bootstrap	90%	17.85	21.30
Bootstrap	95%	17.59	21.58

2.3. Podsumowanie

Średnia cena samochodu:

$$\bar{x} = 19509.68 \text{ USD}$$

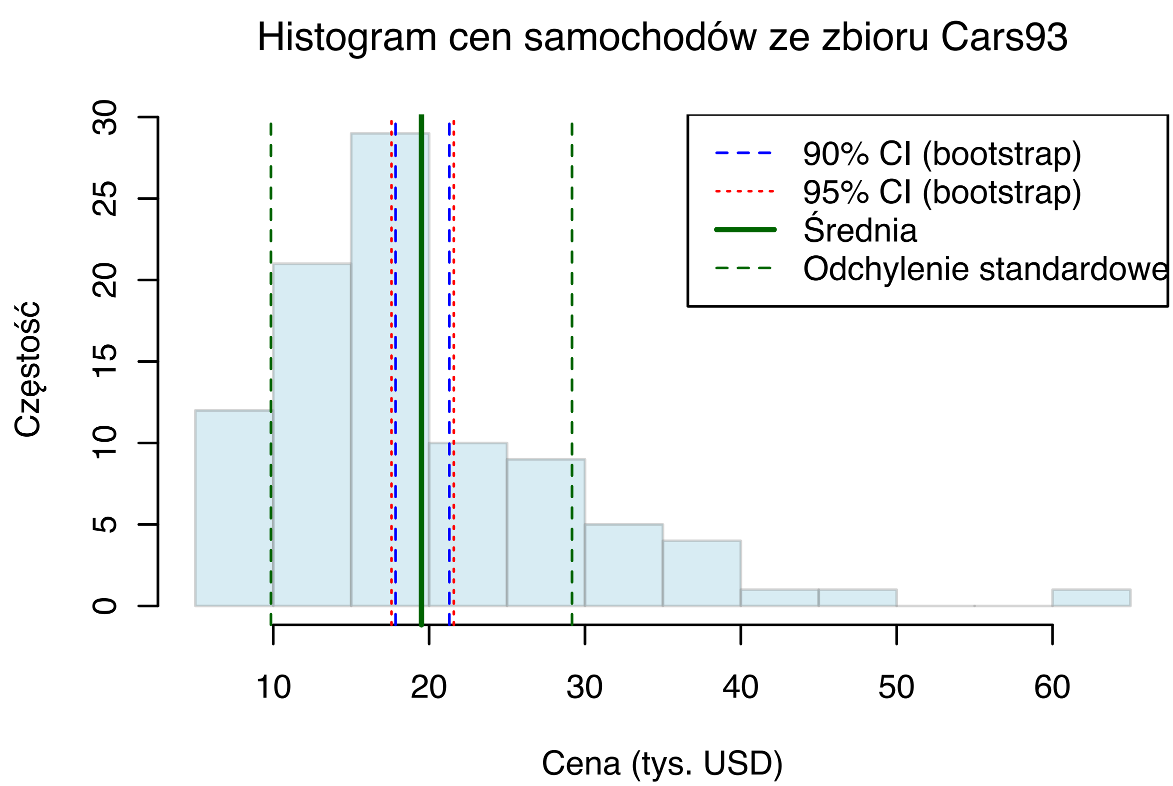
Odchylenie standardowe ceny samochodów:

$$\sigma = 9659.43 \text{ USD}$$

Przedziały ufności:

Metoda	Poziom ufności	Dolna granica	Górna granica
Bootstrap	90%	17846.77	21298.76
Bootstrap	95%	17586.02	21584.19

Przedstawienie wartości na wykresie:



3. Zbiór UScrime z biblioteki `MASS`

Drugą próbkę, jaką wybrałem, stanowi zbiór `UScrime` z biblioteki `MASS`. Dane te dotyczą przestępczości w 47 stanach USA i zawierają wiele zmiennych społeczno-ekonomicznych, takich jak nierówność dochodów, ilość przestępstw na osobę, wydatki na policję czy też średnią liczbę lat nauki. W tej analizie skoncentruję się na badaniu wydatków na policję per capita w roku 1960 (`Po1`).

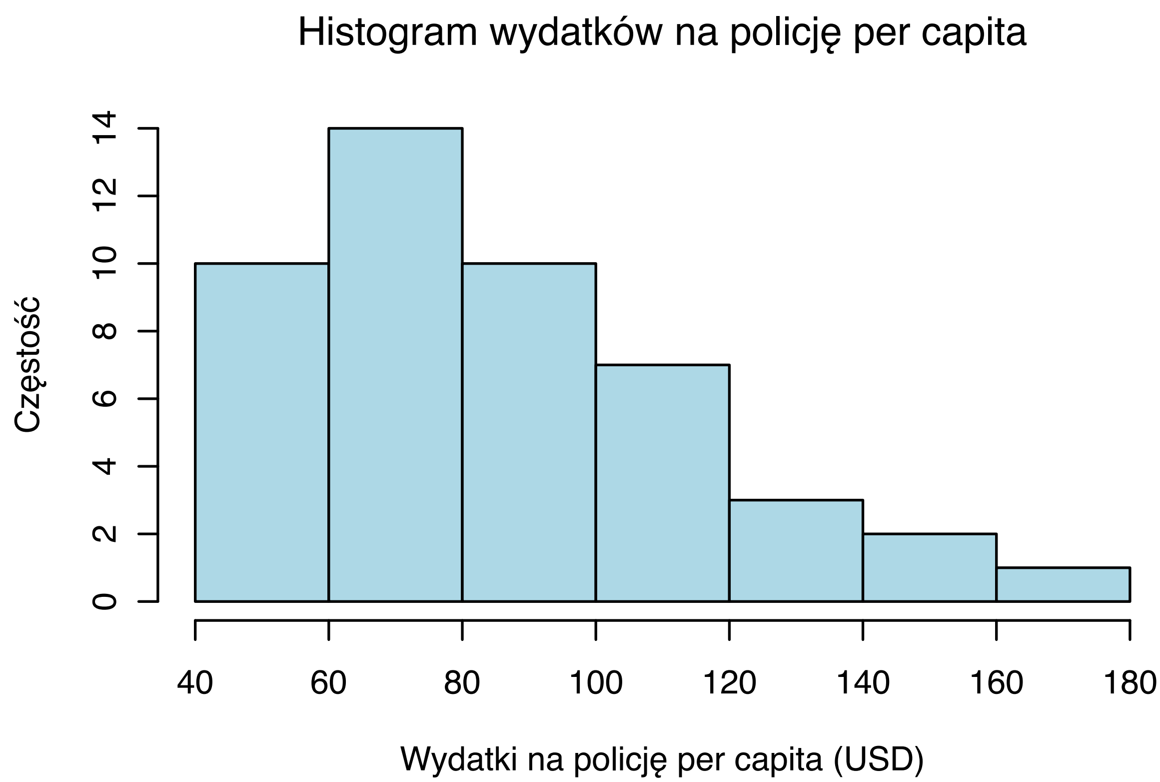
3.1. Analiza normalności próbki

Najpierw sprawdzę normalność rozkładu wydatków na policję per capita w roku 1960 (`Po1`) w zbiorze `UScrime` przy użyciu histogramu, wykresu kwantyl-kwantyl oraz testu Shapiro-Wilka.

Wczytanie danych:

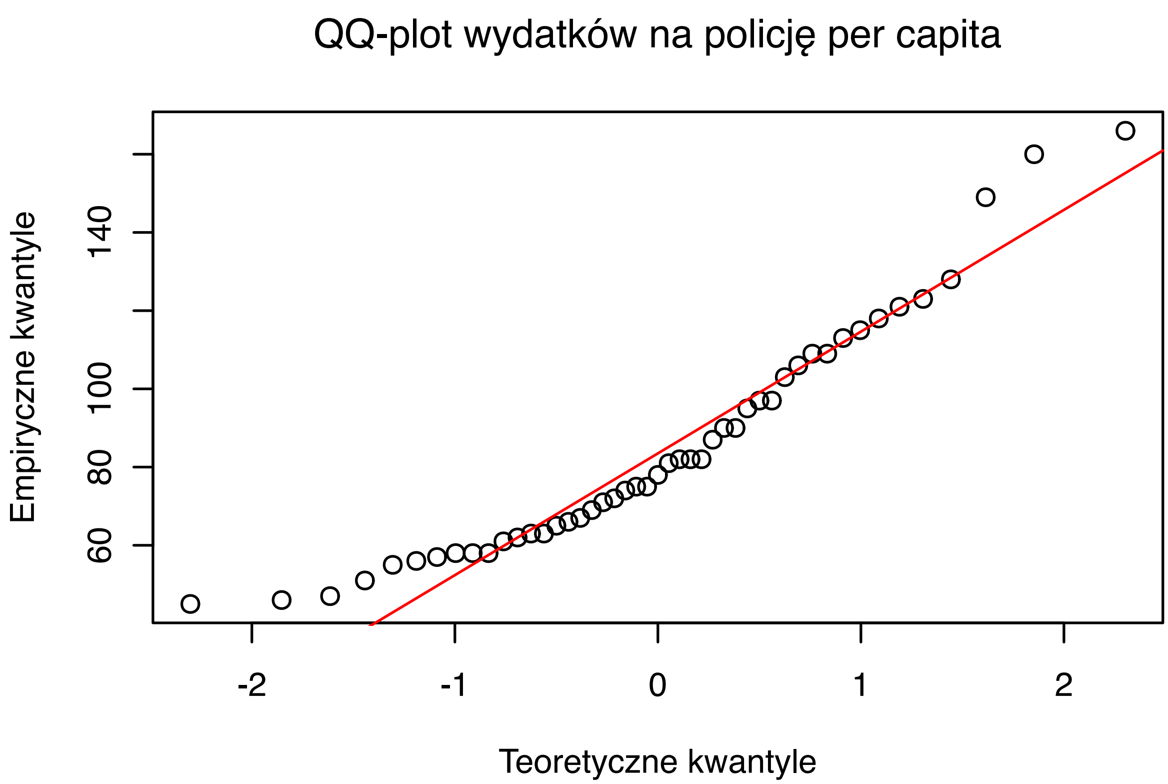
```
library(MASS)
data(UScrime)
```

```r
# Histogram
hist(UScrime$Po1,
     main = "Histogram wydatków na policję per capita",
     xlab = "Wydatki na policję per capita (USD)",
     ylab = "Częstość",
     col = "lightblue",
     border = "black",
     breaks = 7)
```



Na histogramie możemy zauważyć, że rozkład wydatków na policję per capita jest asymetryczny prawostronny (prawostronnie skośny), co sugeruje, że wydatki są bardziej skoncentrowane w niższych wartościach - zwłaszcza w przedziale $[0, 100]$ USD.

```r
# QQ-plot
qqnorm(UScrime$Po1,
       main = "QQ-plot wydatków na policję per capita",
       xlab = "Teoretyczne kwantyle",
       ylab = "Empiryczne kwantyle")
qqline(UScrime$Po1, col = "red")
```



Wykres kwantyl-kwantyl (*QQ*-plot) pokazuje, że dane nie są zgodne z rozkładem normalnym - widać na nim odchylenia od linii prostej, zwłaszcza na końcach (ogony rozkładu).

Podobnie jak w poprzedniej analizie na zbiorze `Cars93`, przeprowadzę test Shapiro-Wilka w celu ostatecznej oceny normalności rozkładu wydatków na policję per capita.

```r
crime_shapiro_test <- shapiro.test(UScrime$Po1)
crime_shapiro_test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  UScrime$Po1
## W = 0.92307, p-value = 0.004281
```

Wynik testu Shapiro-Wilka pokazuje, że wartość p jest znacznie mniejsza niż 0.05, co sugeruje, że rozkład nie jest normalny. Wraz z histogramem i wykresem kwantyl-kwantyl jest to ostateczne potwierdzenie na niezgodność wydatków na policję per capita w roku 1960 w 47 stanach USA z rozkładem normalnym.

3.2. Obliczanie przedziałów ufności dla średniej, standardowego odchylenia oraz wariancji

Najpierw obliczę średnią i odchylenie standardowe, a następnie obliczę przedziały ufności korzystając z metody bootstrap, z uwagi na to, że rozkład wydatków na policję per capita nie jest normalny.

```
uscrime_sd = sd(UScrime$Po1)
uscrime_mean = mean(UScrime$Po1)
# Ustawienie ziarna losowania
set.seed(151885)
# Deklaracja danych - próbka, liczba obserwacji, liczba pętli
data_sample <- UScrime$Po1
n <- length(data_sample)
n_bootstrap <- 1000
# Inicjalizacja wektora do przechowywania wyników bootstrap
bootstrap_means <- numeric(n_bootstrap)
# Pętla bootstrap
for (i in 1:n_bootstrap) {
  bootstrap_sample <- sample(data_sample, size = n, replace = TRUE)
  bootstrap_means[i] <- mean(bootstrap_sample)
}
# Wyznaczanie przedziałów ufności
ci_90_bootstrap <- quantile(bootstrap_means, probs = c(0.05, 0.95), names = FALSE)
ci_95_bootstrap <- quantile(bootstrap_means, probs = c(0.025, 0.975), names = FALSE)
```

Table 4: Przedziały ufności wydatków na policję per capita w roku 1960 ze zbioru UScrime

Metoda	Poziom ufności	Dolna granica	Górna granica
Bootstrap	90%	77.68	92.45
Bootstrap	95%	76.62	93.47

3.3. Podsumowanie

Średnie wydatki na policję per capita w roku 1960 w 47 stanach USA:

$$\bar{x} = 85 \text{ USD}$$

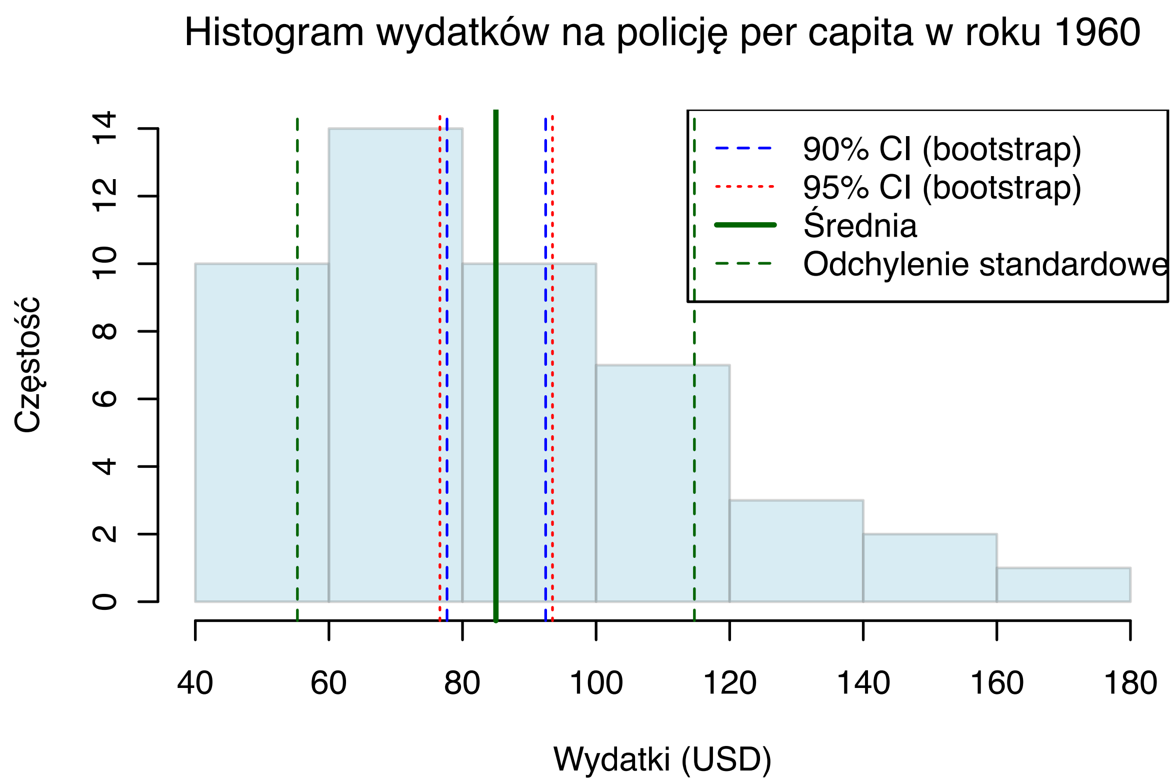
Odchylenie standardowe:

$$\sigma = 29.72 \text{ USD}$$

Przedziały ufności:

Metoda	Poziom ufności	Dolna granica	Górna granica
Bootstrap	90%	77.68	92.45
Bootstrap	95%	76.62	93.47

Przedstawienie wartości na wykresie:



4. Zbiór `road` z biblioteki `MASS`

Ostatnim datasetem który przeanalizuję w tym projekcie będzie zbiór `road` z biblioteki `MASS`. Dane te przedstawiają liczbę ofiar śmiertelnych w wypadkach drogowych w 25 stanach USA, a także liczbę kierowców, gęstość zaludnienia, długość dróg wiejskich, temperatury, zużycie paliwa oraz inne czynniki mogące mieć wpływ na wypadkowość. W tej analizie skupimy się na zbadaniu liczby zgonów w wypadkach drogowych (`deaths`).

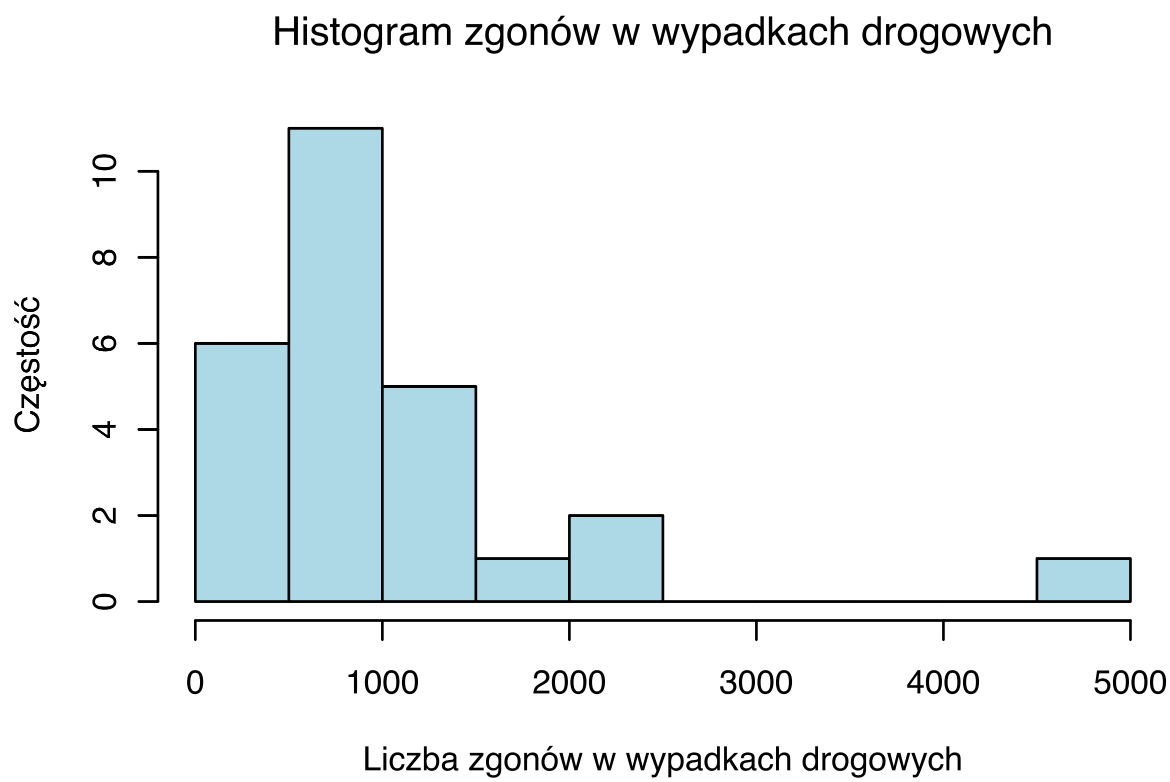
4.1. Analiza normalności próbki

Najpierw sprawdzę normalność rozkładu liczby zgonów w wypadkach drogowych (`deaths`) w zbiorze `road` przy użyciu histogramu, wykresu kwantyl-kwantyl oraz testu Shapiro-Wilka.

Wczytanie danych:

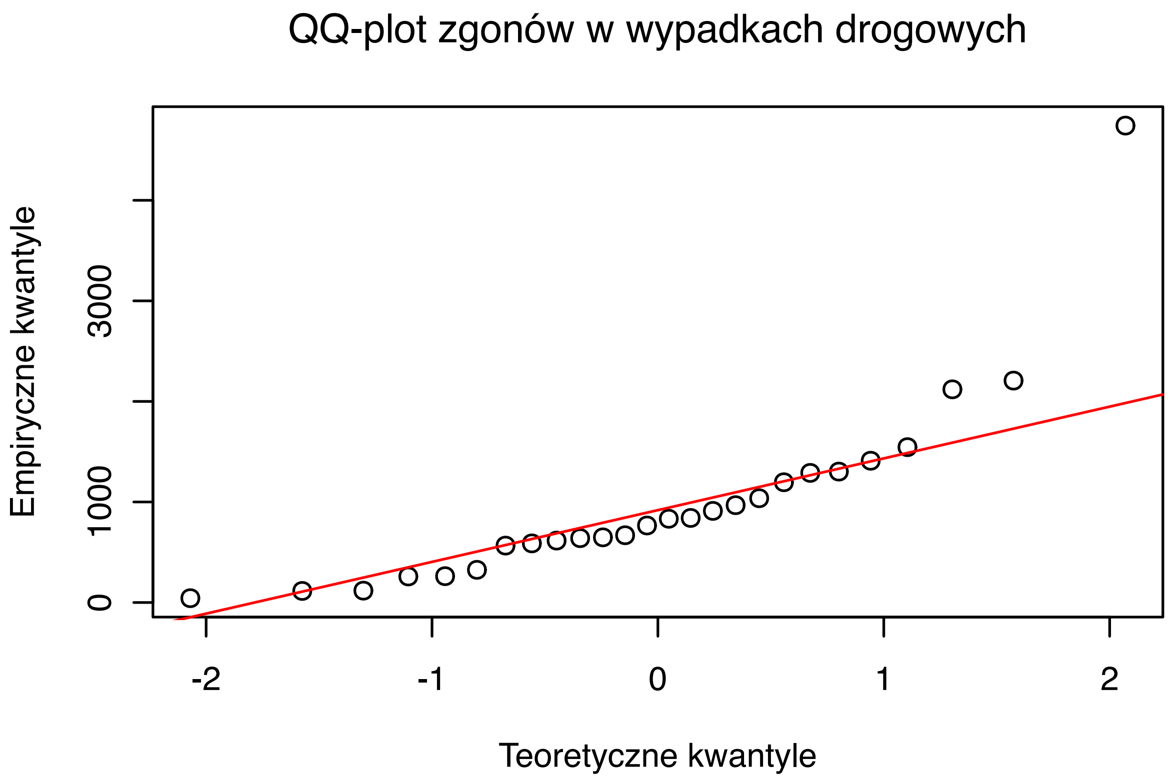
```
library(MASS)
data(road)
```

```r
# Histogram
hist(road$deaths,
     main = "Histogram zgonów w wypadkach drogowych",
     xlab = "Liczba zgonów w wypadkach drogowych",
     ylab = "Częstość",
     col = "lightblue",
     border = "black",
     breaks = 7)
```



Na histogramie możemy zauważyć, że rozkład liczby zgonów w wypadkach drogowych jest asymetryczny prawostronny (prawostronnie skośny), co sugeruje, że liczby zgonów w wypadkach w danym stanie są bardziej skoncentrowane w niższych wartościach - zwłaszcza w przedziale $[0, 1500]$ osób. Warto zaznaczyć, że większość stanów ma relatywnie niską liczbę zgonów ($[0, 2500]$), ale występują wartości odstające w przedziale $[4500, 5000]$.

```r
# QQ-plot
qqnorm(road$deaths,
       main = "QQ-plot zgonów w wypadkach drogowych",
       xlab = "Teoretyczne kwantyle",
       ylab = "Empiryczne kwantyle")
qqline(road$deaths, col = "red")
```



Wykres kwantyl-kwantyl (*QQ*-plot) pokazuje, że dane nie są zgodne z rozkładem normalnym - widać na nim odchylenia od linii prostej, zwłaszcza na prawym ogonie rozkładu. Punkty, które tam widzimy są znacznie powyżej linii, co oznacza obecność wartości odstających i skośność prawostronną.

Podobnie jak w poprzednich analizach, przeprowadzę teraz test Shapiro-Wilka. Na bazie naszych poprzednich obserwacji możemy spodziewać się, że rozkład liczby zgonów w wypadkach drogowych nie jest normalny.

```r
road_shapiro_test <- shapiro.test(road$deaths)
road_shapiro_test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  road$deaths
## W = 0.74488, p-value = 2.327e-05
```

Wynik testu Shapiro-Wilka pokazuje, że wartość p jest znacznie mniejsza niż 0.05, co sugeruje, że rozkład nie jest normalny. Wraz z histogramem i wykresem kwantyl-kwantyl jest to ostateczne potwierdzenie na niezgodność naszych danych z rozkładem normalnym.

4.2. Obliczanie przedziałów ufności dla średniej, standardowego odchylenia oraz wariancji

Najpierw obliczę średnią i odchylenie standardowe, a następnie obliczę przedziały ufności korzystając z metody bootstrap, z uwagi na to, że rozkład liczby zgonów w wypadkach drogowych nie jest normalny.

```
roads_sd = sd(road$deaths)
roads_mean = mean(road$deaths)
# Ustawienie ziarna losowania
set.seed(151885)
# Deklaracja danych - próbka, liczba obserwacji, liczba pętli
data_sample <- road$deaths
n <- length(data_sample)
n_bootstrap <- 1000
# Inicjalizacja wektora do przechowywania wyników bootstrap
bootstrap_means <- numeric(n_bootstrap)
# Pętla bootstrap
for (i in 1:n_bootstrap) {
  bootstrap_sample <- sample(data_sample, size = n, replace = TRUE)
  bootstrap_means[i] <- mean(bootstrap_sample)
}
# Wyznaczanie przedziałów ufności
ci_90_bootstrap <- quantile(bootstrap_means, probs = c(0.05, 0.95), names = FALSE)
ci_95_bootstrap <- quantile(bootstrap_means, probs = c(0.025, 0.975), names = FALSE)
```

Table 6: Przedziały ufności liczby zgonów w wypadkach drogowych ze zbioru road

Metoda	Poziom ufności	Dolna granica	Górna granica
Bootstrap	90%	733.42	1335.09
Bootstrap	95%	700.57	1406.93

4.3. Podsumowanie

Średnia liczba zgonów w wypadkach drogowych w połowie stanów USA:

$$\bar{x} = 1000.65$$

Odchylenie standardowe:

$$\sigma = 946.84$$

Przedziały ufności:

Metoda	Poziom ufności	Dolna granica	Górna granica
Bootstrap	90%	733.42	1335.09
Bootstrap	95%	700.57	1406.93

Przedstawienie wartości na wykresie: