Projekt 1

Wstęp do Analizy Danych | Politechnika Krakowska

Jakub Kapała

Numer albumu: 151885

Data: 15.05.2025

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.

a) 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 ]
```

b) 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:

```
# 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%.

```
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
```

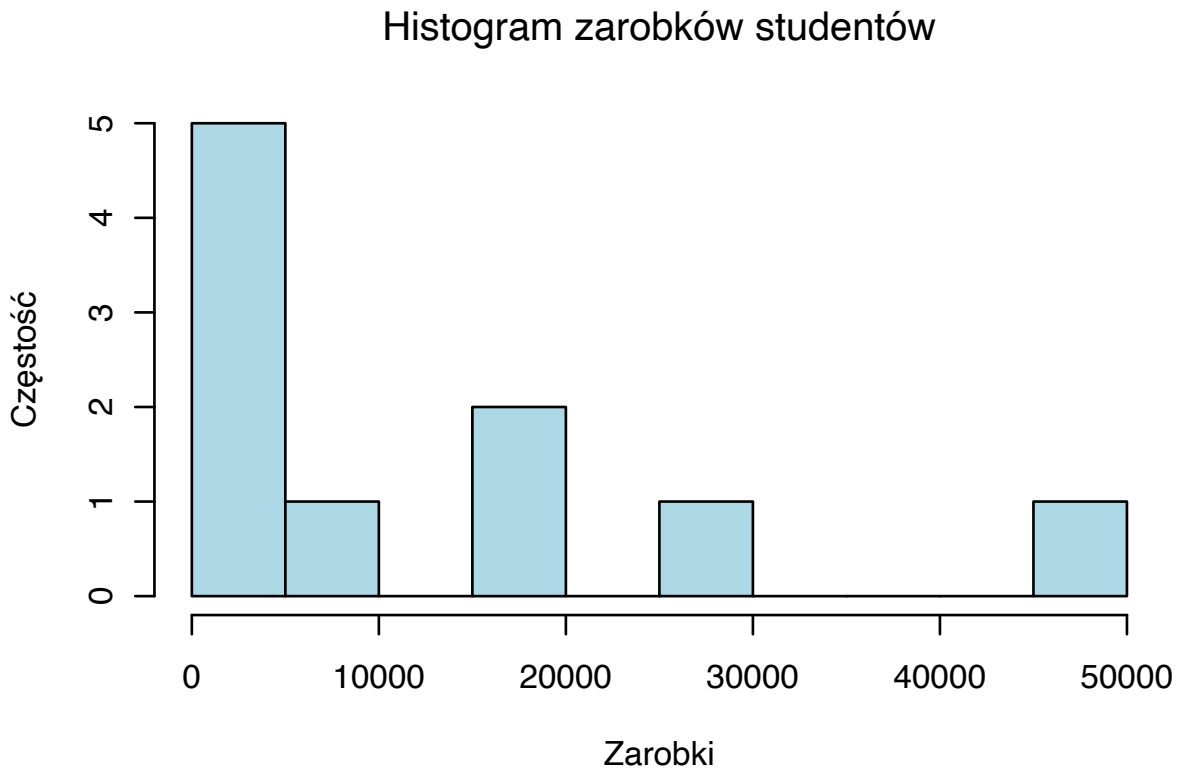
```
## Przedział ufności 90%: [ 4812.39 ,  21811.01 ]
```

```
## 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.

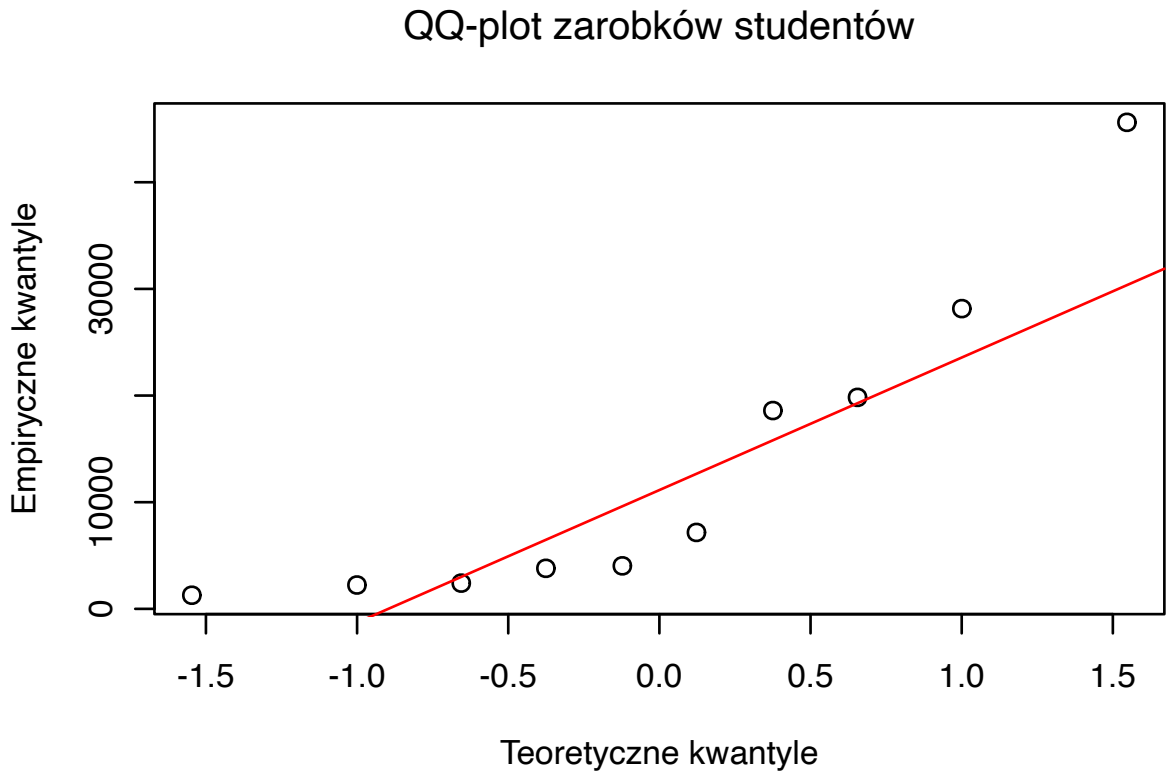
Teraz spójrzmy na histogram oraz wykres kwantyl-kwantyl (QQ-plot).

```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 $[0, 5000]$.

```r
# 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.

c) Użycie metody bootstrap do konstrukcji przedziałów ufności

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

```
# 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:

```
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%:

```
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

Podsumowanie

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.