Laboratorium 6

Wstęp do Analizy Danych | Politechnika Krakowska

Jakub Kapała

Numer albumu: 151885

Data: 31.05.2025

Testowanie hipotez statystycznych dla średniej populacji w przypadku dużych prób

Testowanie hipotez statystycznych jest jednym z kluczowych narzędzi analizy danych, umożliwiającym podejmowanie decyzji w oparciu o dane próbne. W praktyce często mamy do czynienia z przypadkami, gdzie interesuje nas porównanie średniej próby z wartością teoretyczną (hipotetyczną średnią populacji). Takie testy wykorzystuje się w różnych dziedzinach, takich jak przemysł, medycyna, socjologia czy finanse, aby odpowiedzieć na pytanie: czy obserwowane dane dostarczają wystarczających dowodów na różnicę w porównaniu do założonej wartości?

W tym laboratorium skupimy się na testach hipotez dla średniej populacji, zakładając, że:

- Odchylenie standardowe populacji jest znane lub nieznane,
- Wielkość próby jest duża ($n > 30$), co pozwala korzystać z rozkładu normalnego (rozkładu z).

Cel ćwiczeń:

1. Zrozumienie, jak formułować i testować hipotezy statystyczne dotyczące średniej populacji.
2. Przeprowadzenie obliczeń statystyki z i jej interpretacja w kontekście odrzucenia lub przyjęcia hipotezy zerowej.
3. Praktyczne zastosowanie wiedzy do rozwiązywania rzeczywistych problemów statystycznych za pomocą środowiska R.

Interpretacja wyników: Jeśli wartość statystyki z jest większa od wartości krytycznej (lub p-wartość jest mniejsza od α), odrzucamy hipotezę zerową.

Zadanie 1 - Test dla średniej populacji - Znane odchylenie standardowe

Treść Firma twierdzi, że średnia waga ich paczek wynosi 5 kg. Aby zweryfikować tę informację, wybrano losowo 100 paczek, których średnia waga wyniosła 5.1 kg. Odchylenie standardowe populacji wynosi 0.3 kg. Czy na poziomie istotności $\alpha = 0.05$ można stwierdzić, że średnia waga paczek różni się od 5 kg?

Rozwiązanie Deklaracja danych - liczebność próby (n), średnia próby (mean_sample), odchylenie standardowe populacji (sd), poziom istotności (alpha) oraz średnia z hipotezy zerowej (mean_0):

```
n <- 100
mean_sample <- 5.1
sd <- 0.3
alpha <- 0.05
mean_0 <- 5
```

Statystyka testowa z jest obliczana według wzoru:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Obliczenie statystyki z:

```
z <- (mean_sample - mean_0) / (sd / sqrt(n))
```

$$z = 3.3333$$

Teraz obliczmy wartość p-value dla tego testu. Ponieważ jest to test jednostronny, użyjemy funkcji `pnorm`:

```
p_value <- 2 * (1 - pnorm(abs(z)))
```

$$P = 0.0008581$$

Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ w teście dwustronnym wynosi:

```
z_critical <- qnorm(1 - alpha / 2)
```

$$z_{kr} = 1.96$$

Z racji na to, że wartość statystyki z jest większa niż wartość krytyczna ($|z| > z_{kr}$), możemy odrzucić hipotezę zerową. Ostatecznie, ponieważ p-value jest mniejsze niż α, również odrzucamy hipotezę zerową.

Zadanie 2 - Test dla średniej populacji – Brak istotnej różnicy w średniej

Treść Badacz uważa, że przeciętna długość życia w mieście wynosi 75 lat. W badaniu przeprowadzonym na próbie 400 mieszkańców średnia długość życia wyniosła 74.6 lat, a odchylenie standardowe populacji wynosi 3 lata. Czy na poziomie istotności $\alpha = 0.01$ można stwierdzić, że przeciętna długość życia różni się od 75 lat?

Rozwiązanie Deklaracja danych - liczebność próby (n), średnia próby (mean_sample), odchylenie standardowe populacji (sd), poziom istotności (alpha) oraz średnia z hipotezy zerowej (mean_0):

```
n <- 400
mean_sample <- 74.6
sd <- 3
alpha <- 0.01
mean_0 <- 75
```

Obliczenie statystyki z:

```
z <- (mean_sample - mean_0) / (sd / sqrt(n))
```

$$z = -2.6667$$

Teraz obliczmy wartość p-value dla tego testu. Ponieważ jest to test jednostronny, użyjemy funkcji `pnorm`:

```
p_value <- 2 * (1 - pnorm(abs(z)))
```

$$P = 0.00766$$

Wartość krytyczna dla poziomu istotności $\alpha = 0.01$ w teście dwustronnym wynosi:

```
z_critical <- qnorm(1 - alpha / 2)
```

$$z_{kr} = 2.5758$$

Z racji na to, że wartość statystyki z jest większa niż wartość krytyczna ($|z| > z_{kr}$), odrzucamy hipotezę zerową. Na poziomie istotności $\alpha = 0.01$ można stwierdzić, że przeciętna długość życia różni się od 75 lat.

Zadanie 3 - Test dla średniej populacji – Porównanie z wartością referencyjną

Treść W firmie twierdzi się, że przeciętny czas obsługi klienta wynosi 10 minut. Przeprowadzono badanie na próbie 250 klientów, w którym średni czas obsługi wyniósł 10.4 minuty. Znane odchylenie standardowe populacji to 1.2 minuty. Czy na poziomie istotności $\alpha = 0.05$ można uznać, że średni czas obsługi różni się od 10 minut?

Rozwiązanie Deklaracja danych - liczebność próby (n), średnia próby (mean_sample), odchylenie standardowe populacji (sd), poziom istotności (alpha) oraz średnia z hipotezy zerowej (mean_0):

```
n <- 250
mean_sample <- 10.4
sd <- 1.2
alpha <- 0.05
mean_0 <- 10
```

Obliczenie statystyki z:

```
z <- (mean_sample - mean_0) / (sd / sqrt(n))
```

$$z = 5.2705$$

Teraz obliczmy wartość p-value dla tego testu. Ponieważ jest to test jednostronny, użyjemy funkcji `pnorm`:

```
p_value <- 2 * (1 - pnorm(abs(z)))
```

$$P = 0.000000136$$

Wartość krytyczna dla poziomu istotności $\alpha = 0.05$ w teście dwustronnym wynosi:

```
z_critical <- qnorm(1 - alpha / 2)
```

$$z_{kr} = 1.96$$

Ponieważ $|z| > z_{kr}$ oraz p-value $< \alpha$, odrzucamy hipotezę zerową. Na poziomie istotności $\alpha = 0.05$ można uznać, że średni czas obsługi różni się od 10 minut.

Zadanie 4 - Test średniej – Brak znanego odchylenia standardowego

Treść Producent twierdzi, że średnia długość kabli w partii wynosi 10 metrów. Przeprowadzono kontrolę na losowej próbie 50 kabli, uzyskując średnią 10.2 metra i odchylenie standardowe próby wynoszące 0.4 metra. Czy na poziomie istotności $\alpha = 0.05$ można stwierdzić, że średnia długość kabli różni się od deklarowanej wartości?

Rozwiązanie Deklaracja danych - liczebność próby (n), średnia próby (mean_sample), odchylenie standardowe próbki (sd), poziom istotności (alpha) oraz średnia z hipotezy zerowej (mean_0):

```
n <- 50
mean_sample <- 10.2
sd_sample <- 0.4
alpha <- 0.05
mean_0 <- 10
```

Z racji na to, że odchylenie standardowe populacji nie jest znane (znamy jedynie odchylenie standardowe próby), użyjemy rozkładu t-Studenta. Obliczamy statystykę t:

```
t_stat <- (mean_sample - mean_0) / (sd_sample / sqrt(n))
```

$$t = 3.5355$$

Teraz obliczmy stopnie swobody (df) i p-value:

```
df <- n - 1
p_value <- 2 * (1 - pt(abs(t_stat), df))
```

$$P = 0.0009$$

Obliczmy także wartość krytyczną dla poziomu istotności $\alpha = 0.05$:

```
t_crit <- qt(1 - alpha / 2, df)
```

$$t_{kr} = 2.0096$$

Ponieważ $|t| > t_{kr}$ oraz p-value $< \alpha$, odrzucamy hipotezę zerową. Na poziomie istotności $\alpha = 0.05$ można stwierdzić, że średnia długość kabli różni się od deklarowanej wartości 10 metrów.

Zadanie 5 - Test średniej – Ocena przeciętnej temperatury

Treść Meteorolog twierdzi, że przeciętna temperatura w danym regionie w maju wynosi 18°C. W badaniu przeprowadzonym na podstawie 60 losowo wybranych dni majowych w różnych latach uzyskano średnią 17.8°C i odchylenie standardowe próby 1.5°C. Czy na poziomie istotności $\alpha = 0.05$ można stwierdzić, że przeciętna temperatura różni się od 18°C?

Rozwiązanie Deklaracja danych - liczebność próby (n), średnia próby (mean_sample), odchylenie standardowe próbki (sd), poziom istotności (alpha) oraz średnia z hipotezy zerowej (mean_0):

```
n <- 60
mean_sample <- 17.8
sd <- 1.5
alpha <- 0.05
mean_0 <- 18
```

Odchylenie standardowe populacji nie jest znane, używam więc rozkładu t-Studenta:

```
t_stat <- (mean_sample - mean_0) / (sd / sqrt(n))
```

$$t = -1.0328$$

Teraz obliczmy stopnie swobody (df) i p-value:

```
df <- n - 1
p_value <- 2 * (1 - pt(abs(t_stat), df))
```

$$P = 0.31$$

Obliczmy także wartość krytyczną dla poziomu istotności $\alpha = 0.05$:

```
t_crit <- qt(1 - alpha / 2, df)
```

$$t_{kr} = 2.001$$

Ponieważ $|t| < t_{kr}$ oraz p-value $> \alpha$, nie ma podstaw do odrzucenia hipotezy zerowej. Na poziomie istotności $\alpha = 0.05$ nie można stwierdzić, że przeciętna temperatura różni się od 18°C.

Zadanie 6 - Test jednostronny – Podejrzenie niższych wyników

Treść Średni wynik z egzaminu końcowego na danym kursie wynosi 75 punktów. Nauczyciel podejrzewa, że w tym roku wyniki są niższe. W próbie 40 studentów średni wynik wyniósł 73 punkty, a odchylenie standardowe próby wyniosło 5 punktów. Czy na poziomie istotności $\alpha = 0.01$ można uznać, że wyniki egzaminu są niższe niż 75 punktów?

Rozwiązanie Deklaracja danych - liczebność próby (n), średnia próby (mean_sample), odchylenie standardowe próbki (sd), poziom istotności (alpha) oraz średnia z hipotezy zerowej (mean_0):

```
n <- 40
mean_sample <- 73
sd <- 5
alpha <- 0.01
mean_0 <- 75
```

Statystyka t (test jednostronny: czy średnia jest mniejsza niż 75):

```
t_stat <- (mean_sample - mean_0) / (sd / sqrt(n))
df <- n - 1
p_value <- pt(t_stat, df) # p-value jednostronne (lewy ogon)
t_crit <- qt(alpha, df) # jednostronny test, lewy ogon
```

$$t = -2.5298$$

$$P = 0.01$$

$$t_{kr} = -2.4258$$

Ponieważ $t < t_{kr}$ oraz p-value $< \alpha$, odrzucamy hipotezę zerową. Na poziomie istotności $\alpha = 0.01$ można uznać, że wyniki egzaminu są niższe niż 75 punktów.

Zadanie 7 - Test średniej – Czas działania baterii

Treść Firma produkująca baterie twierdzi, że średni czas działania ich baterii wynosi 100 godzin. Losowo wybrano próbę 50 baterii, a ich średni czas działania wyniósł 98 godzin. Odchylenie standardowe próby wynosi 4 godziny. Czy na poziomie istotności $\alpha = 0.05$ można stwierdzić, że średni czas działania baterii różni się od deklarowanych 100 godzin?

Rozwiązanie Deklaracja danych - liczebność próby (n), średnia próby (mean_sample), odchylenie standardowe próbki (sd), poziom istotności (alpha) oraz średnia z hipotezy zerowej (mean_0):

```
n <- 50
mean_sample <- 98
sd <- 4
alpha <- 0.05
mean_0 <- 100
```

Statystyka t (test dwustronny):

```
t_stat <- (mean_sample - mean_0) / (sd / sqrt(n))
df <- n - 1
p_value <- 2 * (1 - pt(abs(t_stat), df))
t_crit <- qt(1 - alpha / 2, df)
```

$$t = -3.5355$$

$$P = 0.0009$$

$$t_{kr} = 2.0096$$

Ponieważ $|t| > t_{kr}$ oraz p-value $< \alpha$, odrzucamy hipotezę zerową. Na poziomie istotności $\alpha = 0.05$ można stwierdzić, że średni czas działania baterii różni się od deklarowanych 100 godzin.

Zadanie 9 - Test średniej jednej populacji (test t)

Założenia

- Dane pochodzą z populacji o **rozkładzie normalnym**.
- Próba jest losowa i mała ($n < 30$).

Treść

9.1 Dla próby o wielkości $n = 12$, średniej $\bar{x} = 14.3$ i odchyleniu standardowym $s = 2.1$, sprawdź hipotezę:

$$H_0 : \mu = 15 \quad \text{vs} \quad H_1 : \mu \neq 15$$

9.2 Przyjmij poziom istotności $\alpha = 0.05$.

9.3 Sprawdź założenie normalności rozkładu przy użyciu testu Shapiro-Wilka.

9.4 Wykonaj test t dla jednej średniej.

Rozwiązanie

Najpierw ustawiam seed na swój numer albumu dla powtarzalności wyników:

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

Następnie deklaruję dane:

```
n <- 12
mean_sample <- 14.3
sd_sample <- 2.1
alpha <- 0.05
mean_0 <- 15
```

Generowanie danych z rozkładu normalnego:

```
data <- rnorm(n, mean = mean_sample, sd = sd_sample)
```

Sprawdzam normalność rozkładu przy użyciu testu Shapiro-Wilka:

```
shapiro_test <- shapiro.test(data)
shapiro_test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data
## W = 0.95086, p-value = 0.6496
```

Wynik testu Shapiro-Wilka: Statystyka W wynosi 0.9509, a p-value to 0.64962. Ponieważ p-value jest większe niż $\alpha = 0.05$, nie ma podstaw do odrzucenia hipotezy o normalności rozkładu.

Test dwustronny - $H_0 : \mu = 15$ vs $H_1 : \mu \neq 15$:

```
t_test <- t.test(data, mu = mean_0, alternative = "two.sided")
t_test
```

```
##
##  One Sample t-test
##
## data:  data
## t = -1.8382, df = 11, p-value = 0.09316
## alternative hypothesis: true mean is not equal to 15
## 95 percent confidence interval:
##  12.88710 15.18976
## sample estimates:
## mean of x
##  14.03843
```

Z testu wynika, że statystyka t wynosi -1.8382, a p-value to 0.09316. Z racji na to, że p-value jest większe niż $\alpha = 0.05$, nie ma podstaw do odrzucenia hipotezy zerowej. Średnia z próby nie jest istotnie różna od 15.

Zadanie 10 - Test dwóch średnich dla prób niezależnych (równe wariancje)

Założenia

- Obie próbki pochodzą z populacji o **rozkładzie normalnym**.
- Próby są **niezależne**.
- Wariancje populacji są **równe** (sprawdzone przed testem).

Treść

10.1 Dane dla dwóch prób:

- Próba 1: $n_1 = 10$, $\bar{x}_1 = 13.5$, $s_1 = 2.0$
- Próba 2: $n_2 = 12$, $\bar{x}_2 = 12.0$, $s_2 = 1.8$

10.2 Sprawdź założenie normalności dla obu prób przy użyciu testu Shapiro-Wilka.

10.3 Przeprowadź test hipotezy:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

przy założeniu równości wariancji.

10.4 Przyjmij poziom istotności $\alpha = 0.05$.

Rozwiązanie

Deklaracja danych dla obu prób:

```
n1 <- 10
mean1 <- 13.5
s1 <- 2

n2 <- 12
mean2 <- 12.0
s2 <- 1.8

alpha <- 0.05
```

Generowanie danych z rozkładu normalnego dla obu prób:

```
set.seed(151885)
data1 <- rnorm(n1, mean = mean1, sd = s1)
data2 <- rnorm(n2, mean = mean2, sd = s2)
```

Sprawdzanie normalności rozkładu dla obu prób przy użyciu testu Shapiro-Wilka:

```
shapiro_test1 <- shapiro.test(data1)
shapiro_test1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data1
## W = 0.95068, p-value = 0.6766
```

```r
shapiro_test2 <- shapiro.test(data2)
shapiro_test2
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data2
## W = 0.94122, p-value = 0.514
```

Statystyka W dla próby 1 wynosi 0.9507, a p-value to 0.67656. Dla próby 2 statystyka W wynosi 0.9412, a p-value to 0.51404. Ponieważ p-value w obu przypadkach jest większe niż $\alpha = 0.05$, nie ma podstaw do odrzucenia hipotezy o normalności rozkładu.

Test t dla dwóch średnich przy założeniu równości wariancji (test dwustronny):

```r
t_test_equal_var <- t.test(data1, data2, var.equal = TRUE, alternative = "two.sided")
t_test_equal_var
```

```
##
##  Two Sample t-test
##
## data:  data1 and data2
## t = 1.9949, df = 20, p-value = 0.05986
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07090268  3.17857062
## sample estimates:
## mean of x mean of y
##  13.23258  11.67874
```

Ponieważ p-value $> \alpha$, nie ma podstaw do odrzucenia hipotezy zerowej. Na poziomie istotności $\alpha = 0.05$ nie można stwierdzić, że średnie obu populacji są różne.

Zadanie 11 - Test dwóch średnich dla prób niezależnych (nierówne wariancje)

Założenia

- Obie próbki pochodzą z populacji o **rozkładzie normalnym**.
- Próby są **niezależne**.
- Wariancje populacji są **różne**.

Treść

11.1 Powtórz dane z zadania 10, ale tym razem nie zakładaj równości wariancji.

11.2 Sprawdź założenie normalności dla obu prób.

11.3 Przeprowadź test hipotezy:

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

Rozwiązanie

Deklaracja danych dla obu prób:

```
n1 <- 10
mean1 <- 13.5
s1 <- 2

n2 <- 12
mean2 <- 12.0
s2 <- 1.8

alpha <- 0.05
```

Generowanie danych z rozkładu normalnego dla obu prób:

```
set.seed(151885)
data1 <- rnorm(n1, mean = mean1, sd = s1)
data2 <- rnorm(n2, mean = mean2, sd = s2)
```

Sprawdzanie normalności rozkładu dla obu prób przy użyciu testu Shapiro-Wilka:

```
shapiro_test1 <- shapiro.test(data1)
shapiro_test1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data1
## W = 0.95068, p-value = 0.6766
```

```r
shapiro_test2 <- shapiro.test(data2)
shapiro_test2
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data2
## W = 0.94122, p-value = 0.514
```

Statystyka W dla próby 1 wynosi 0.9507, a p-value to 0.67656. Dla próby 2 statystyka W wynosi 0.9412, a p-value to 0.51404. Ponieważ p-value w obu przypadkach jest większe niż $\alpha = 0.05$, nie ma podstaw do odrzucenia hipotezy o normalności rozkładu.

Test t dla dwóch średnich bez założenia równości wariancji (Welch's t-test):

```r
t_test_unequal_var <- t.test(data1, data2, var.equal = FALSE, alternative = "two.sided")
t_test_unequal_var
```

```
##
##  Welch Two Sample t-test
##
## data:  data1 and data2
## t = 1.9984, df = 19.415, p-value = 0.05987
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07119293  3.17886088
## sample estimates:
## mean of x mean of y
##  13.23258  11.67874
```

Ponieważ p-value $> \alpha$, nie ma podstaw do odrzucenia hipotezy zerowej. Na poziomie istotności $\alpha = 0.05$ nie można stwierdzić, że średnie obu populacji są różne (również przy braku założenia równości wariancji).

Zadanie 12 - Test wariancji jednej populacji (test Chi-kwadrat)

Założenia

- Dane pochodzą z populacji o **rozkładzie normalnym**.
- Próba jest losowa i mała ($n < 30$).

Treść

12.1 Dla próby o wielkości $n = 15$ i wariancji $s^2 = 4.5$, sprawdź hipotezę:

- $H_0 : \sigma^2 = 4$
- $H_1 : \sigma^2 \neq 4$

12.2 Przyjmij poziom istotności $\alpha = 0.05$.

Rozwiązanie Deklaracja danych:

```
n <- 15
sigma2_0 <- 4
sd = sqrt(4.5)
alpha <- 0.05
```

Generowanie danych z rozkładu normalnego:

```
set.seed(151885)
data <- rnorm(n, mean = 0, sd = sd)
```

Test Chi-kwadrat dla wariancji:

```
chi_sq <- (n - 1) * var(data) / sigma2_0
df <- n - 1
```

P-value dla testu Chi-kwadrat (dwustronny):

```
p_value <- 2 * min(pchisq(chi_sq, df), 1 - pchisq(chi_sq, df))
```

Wartości krytyczne:

```
chi_sq_left <- qchisq(alpha / 2, df)
chi_sq_right <- qchisq(1 - alpha / 2, df)
```

$$\chi^2 = 13.1867$$

$$P = 0.97624$$

$$\chi^2_L = 5.6287, \quad \chi^2_R = 26.1189$$

Ponieważ χ^2 mieści się w przedziale $[\chi^2_L, \chi^2_R]$ oraz p-value $> \alpha$, nie ma podstaw do odrzucenia hipotezy zerowej. Na poziomie istotności $\alpha = 0.05$ nie można stwierdzić, że wariancja różni się od 4.

Zadanie 13 - Test równości dwóch wariancji (test F)

Założenia

- Dane pochodzą z populacji o **rozkładzie normalnym**.
- Próby są **niezależne**.

Treść

13.1 Dla prób:

- Próba 1: $n_1 = 10$, $s_1 = 2.3$
- Próba 2: $n_2 = 12$, $s_2 = 3.1$

13.2 Sprawdź hipotezę:

- $H_0 : \sigma_1^2 = \sigma_2^2$
- $H_1 : \sigma_1^2 \neq \sigma_2^2$

13.3 Przyjmij poziom istotności $\alpha = 0.05$.

Rozwiązanie Deklaracja danych dla obu prób:

```
n1 <- 12
s1 <- 2.3

n2 <- 12
s2 <- 3.1

alpha <- 0.05
```

Generowanie danych z rozkładu normalnego dla obu prób:

```
set.seed(151885)
data1 <- rnorm(n1, mean = 0, sd = s1)
data2 <- rnorm(n2, mean = 0, sd = s2)
```

Test F dla równości wariancji:

```
f_test <- var.test(data1, data2, alternative = "two.sided")
f_test
```

```
##
##  F test to compare two variances
##
## data:  data1 and data2
## F = 0.39965, num df = 11, denom df = 11, p-value = 0.1436
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1150517 1.3882808
## sample estimates:
## ratio of variances
##          0.3996549
```

Ponieważ p-value $> \alpha$, nie ma podstaw do odrzucenia hipotezy zerowej. Na poziomie istotności $\alpha = 0.05$ nie można stwierdzić, że wariancje obu populacji są różne.