Politechnika Wrocławska

Lab 3

Statystyka w zastosowaniach

04.05.2025

Mateusz Nasewicz

Spis treści

1 Zadanie 1: Test Z dla jednej próby

W tym zadaniu przeprowadzono test Z dla jednej próby, mający na celu sprawdzenie, czy średnia próbki istotnie różni się od wartości referencyjnej. Założono, że próbka pochodzi z rozkładu normalnego z określoną średnią i odchyleniem standardowym.

1.1 Założenia

Założenia do przeprowadzenia testu Z były następujące:

- Średnia populacji (wartość referencyjna) wynosi $\mu = 50$.
- Odchylenie standardowe populacji wynosi $\sigma = 5$.
- Liczba próbek wynosi $n = 30$.
- Poziom istotności testu $\alpha = 0.05$.

1.2 Metoda

Test Z został przeprowadzony na podstawie wzoru:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

gdzie:

- $\bar{x}$ to średnia próbki,
- μ to średnia populacji (wartość referencyjna),
- σ to odchylenie standardowe populacji,
- n to liczba próbek.

Na podstawie obliczonej wartości Z, wyliczono również wartość p:

$$p = 2 \cdot (1 - \Phi(|Z|))$$

gdzie Φ to funkcja dystrybuanty standardowego rozkładu normalnego.

1.3 Wyniki

Po przeprowadzeniu obliczeń uzyskano następujące wyniki:

- Średnia próbki: $\bar{x} \approx 52.21$
- Z-statystyka: $Z \approx 2.20$
- Wartość p: $p \approx 0.0275$

1.4 Interpretacja

Na podstawie obliczonej wartości p, która wynosi około 0.00275, porównujemy ją z poziomem istotności $\alpha = 0.05$. Ponieważ wartość p jest mniejsza niż poziom istotności, odrzucamy hipoteze zerową. Oznacza to, że średnia próbki różni się istotnie od wartości referencyjnej (50)

1.5 Wnioski

Na podstawie przeprowadzonego testu Z mamy podstawy do odrzucenia hipotezy zerowej. W związku z tym, można stwierdzić istotną różnicę między średnią próbki a wartością referencyjną 50.

2 Zadanie 2: Test t-Studenta dla jednej próby

W tym zadaniu przeprowadzono test t-Studenta dla jednej próby, mający na celu sprawdzenie, czy średnia próbki istotnie różni się od wartości referencyjnej. Założono, że próbka pochodzi z rozkładu normalnego, a średnia w populacji wynosi określoną wartość.

2.1 Założenia

Założenia do przeprowadzenia testu t-Studenta były następujące:

- Średnia populacji (wartość referencyjna) wynosi $\mu = 50$.
- Odchylenie standardowe próbki zostało oszacowane na podstawie danych.
- Liczba próbek wynosi $n = 30$.
- Poziom istotności testu $\alpha = 0.05$.

2.2 Metoda

Test t-Studenta dla jednej próby został przeprowadzony na podstawie wzoru:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

gdzie:

- $\bar{x}$ to średnia próbki,
- μ to średnia populacji (wartość referencyjna),
- s to odchylenie standardowe próbki,
- n to liczba próbek.

Na podstawie obliczonej t-statystyki, wyliczono również wartość p:

$$p = 2 \cdot (1 - \Phi(|t|))$$

gdzie Φ to funkcja dystrybuanty rozkładu t-Studenta.

2.3 Wyniki

Po przeprowadzeniu obliczeń uzyskano następujące wyniki:

- Średnia próbki: $\bar{x} \approx 51.21$
- Odchylenie standardowe próbki: $s \approx 5.50$
- t-statystyka: $t \approx 1.21$
- Wartość p: $p \approx 0.2365$

2.4 Interpretacja

Na podstawie obliczonej wartości p, która wynosi około 0.2365, porównujemy ją z poziomem istotności $\alpha = 0.05$. Ponieważ wartość p jest większa niż poziom istotności, nie odrzucamy hipotezy zerowej. Oznacza to, że średnia próbki nie różni się istotnie od wartości referencyjnej (50) w granicach błędu statystycznego.

2.5 Wnioski

Na podstawie przeprowadzonego testu t-Studenta nie mamy podstaw do odrzucenia hipotezy zerowej. W związku z tym, nie stwierdzamy istotnej różnicy między średnią próbki a wartością referencyjną 50.

3 Zadanie 3: Test t-Studenta dla dwóch niezależnych próbek

W tym zadaniu przeprowadzono test t-Studenta dla dwóch niezależnych próbek, mający na celu sprawdzenie, czy średnie tych próbek różnią się istotnie. Założono, że próbki pochodzą z rozkładu normalnego, a średnie w obu populacjach są różne.

3.1 Założenia

Założenia do przeprowadzenia testu t-Studenta były następujące:

- Średnia pierwszej próbki wynosi $\mu_1 = 50$.
- Średnia drugiej próbki wynosi $\mu_2 = 55$.
- Odchylenie standardowe obu próbek zostało oszacowane na podstawie danych.
- Liczba próbek w każdej grupie wynosi $n_1 = n_2 = 30$.
- Poziom istotności testu $\alpha = 0.05$.

3.2 Metoda

Test t-Studenta dla dwóch niezależnych prób został przeprowadzony na podstawie wzoru:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

gdzie:

- $\bar{x}_1$ i $\bar{x}_2$ to średnie próbki 1 i 2,
- s_1 i s_2 to odchylenia standardowe próbek,
- n_1 i n_2 to liczba próbek w każdej grupie.

Na podstawie obliczonej t-statystyki, wyliczono również wartość p:

$$p = 2 \cdot (1 - \Phi(|t|))$$

gdzie Φ to funkcja dystrybuanty rozkładu t-Studenta.

3.3 Wyniki

Po przeprowadzeniu obliczeń uzyskano następujące wyniki:

- Średnia pierwszej próbki: $\bar{x}_1 \approx 52.21$
- Średnia drugiej próbki: $\bar{x}_2 \approx 53.55$
- t-statystyka: $t \approx -1.02$
- Wartość p: $p \approx 0.3098$

3.4 Interpretacja

Na podstawie obliczonej wartości p, która wynosi około 0.3098, porównujemy ją z poziomem istotności $\alpha = 0.05$. Ponieważ wartość p jest większa niż poziom istotności, nie odrzucamy hipotezy zerowej. Oznacza to, że średnie próbki nie różnią się istotnie.

3.5 Wnioski

Na podstawie przeprowadzonego testu t-Studenta nie mamy podstaw do odrzucenia hipotezy zerowej. W związku z tym, nie stwierdzamy istotnej różnicy między średnimi obu próbek.

4 Zadanie 4: Test chi-kwadrat dla zgodności rozkładu

W tym zadaniu przeprowadzono test chi-kwadrat dla zgodności rozkładu, mający na celu sprawdzenie, czy obserwowany rozkład zmiennych kategorycznych różni się od rozkładu oczekiwanego. Zakłada się, że kategorie są rozkładające się równomiernie, co daje oczekiwaną liczbę obserwacji w każdej kategorii.

4.1 Założenia

Założenia do przeprowadzenia testu chi-kwadrat były następujące:

- Dane kategoryczne składają się z 4 kategorii: A, B, C i D.
- Liczba obserwacji w poszczególnych kategoriach wynosi:
  - Kategoria A: 50
  - Kategoria B: 30
  - Kategoria C: 10
  - Kategoria D: 10
- Oczekiwana liczba obserwacji w każdej kategorii, zakładając rozkład równomierny, wynosi 25.
- Poziom istotności testu $\alpha = 0.05$.

4.2 Metoda

Test chi-kwadrat dla zgodności rozkładu został przeprowadzony na podstawie wzoru:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

gdzie:

- O_i to liczba obserwacji w i-tej kategorii,
- E_i to liczba oczekiwana w i-tej kategorii.

Na podstawie obliczonej statystyki chi-kwadrat, wyliczono również wartość p:

$$p = P(\chi^2 \geq \text{obliczone wartości chi-kwadrat})$$

gdzie P to funkcja dystrybuanty rozkładu chi-kwadrat.

4.3 Wyniki

Po przeprowadzeniu obliczeń uzyskano następujące wyniki:

- Chi-kwadrat: $\chi^2 = 44.00$
- Wartość p: $p = 0.0000$

4.4 Interpretacja

Na podstawie obliczonej wartości p, która wynosi 0.0000, porównujemy ją z poziomem istotności $\alpha = 0.05$. Ponieważ wartość p jest znacznie mniejsza niż poziom istotności, odrzucamy hipotezę zerową. Oznacza to, że obserwowany rozkład zmiennych kategorycznych różni się istotnie od rozkładu oczekiwanego.

4.5 Wnioski

Na podstawie przeprowadzonego testu chi-kwadrat odrzucamy hipotezę zerową, co oznacza, że rozkład obserwowanych danych różni się istotnie od oczekiwanego rozkładu równomiernego.

5 Zadanie 5: Analiza wariancji (ANOVA) dla wielu grup

W zadaniu przeprowadzono analizę wariancji (ANOVA) dla trzech grup, aby sprawdzić, czy istnieje istotna różnica między średnimi grup. Zakłada się, że mamy dane z trzech grup, każda o innej średniej.

5.1 Założenia

Założenia do przeprowadzenia analizy wariancji były następujące:

- Grupa 1: średnia = 50, odchylenie standardowe = 5, liczba próbek = 30
- Grupa 2: średnia = 55, odchylenie standardowe = 5, liczba próbek = 30
- Grupa 3: średnia = 60, odchylenie standardowe = 5, liczba próbek = 30
- Poziom istotności testu $\alpha = 0.05$

5.2 Metoda

Analiza wariancji (ANOVA) została przeprowadzona przy użyciu testu F. Test ten porównuje średnie między więcej niż dwiema grupami i ocenia, czy istnieje istotna różnica między tymi średnimi.

$$F = \frac{\text{międzygrupowa wariancja}}{\text{wewnątrzgrupowa wariancja}}$$

Na podstawie wyniku testu ANOVA, obliczana jest wartość p, która pozwala na ocenę, czy różnice między grupami są statystycznie istotne.

5.3 Wyniki

Wyniki testu ANOVA dla trzech grup są następujące:

- Statystyka F: $F = 17.30$
- Wartość p: $p = 0.0000$

Ponieważ wartość p jest mniejsza od poziomu istotności $\alpha = 0.05$, odrzucamy hipotezę zerową. Oznacza to, że istnieje istotna różnica między średnimi grup.

5.4 Testy post hoc (Tukey HSD)

Po przeprowadzeniu testu ANOVA, w celu dokładniejszego określenia, które grupy różnią się między sobą, przeprowadzono test Tukeya (test post hoc). Wyniki testu Tukeya przedstawiają się następująco:

Grupa 1	Grupa 2	Różnica średnich	p skorygowane	Dolna granica	Górna granica	Odrzucenie
Grupa 1	Grupa 2	1.34	0.5536	-1.73	4.41	False
Grupa 1	Grupa 3	7.12	0.0000	4.05	10.18	True
Grupa 2	Grupa 3	5.78	0.0001	2.71	8.85	True

Tabelka 1: Wyniki testu Tukeya dla porównań par grup

5.5 Interpretacja wyników

Na podstawie wyników testu Tukeya:

- Nie stwierdzono istotnej różnicy między Grupa 1 a Grupa 2 (wartość p = 0.5536).
- Stwierdzono istotną różnicę między Grupa 1 a Grupa 3 (wartość p = 0.0000).
- Stwierdzono istotną różnicę między Grupa 2 a Grupa 3 (wartość p = 0.0001).

5.6 Wnioski

Na podstawie wyników testu ANOVA odrzucamy hipotezę zerową, co oznacza, że istnieje istotna różnica między średnimi przynajmniej jednej z grup. Testy post hoc (Tukey HSD) wykazały, że różnice są istotne pomiędzy Grupą 1 a Grupą 3, oraz Grupą 2 a Grupą 3, podczas gdy różnice między Grupą 1 a Grupą 2 nie były istotne.

6 Test normalności rozkładu (Shapiro-Wilka, Kołmogorowa-Smirnowa)

Dla analizy normalności rozkładu danych dotyczących długości działki kielicha kwiatów irysa (`sepal_length`) w zbiorze danych `iris`, wykorzystano dwa testy normalności: Shapiro-Wilka oraz Kołmogorowa-Smirnowa. Oto pierwsze kilka wartości próby danych:

```
Dane:
0      5.1
1      4.9
2      4.7
3      4.6
4      5.0
        ...
```

```
145    6.7
146    6.3
147    6.5
148    6.2
149    5.9
Name: sepal_length, Length: 150, dtype: float64
```

6.1 Test Shapiro-Wilka

Test Shapiro-Wilka sprawdza, czy dane pochodzą z rozkładu normalnego. Wynik testu:

$$\text{Statystyka testu} = 0.9761, \quad p\text{-value} = 0.0102$$

Interpretacja: - Hipoteza zerowa testu Shapiro-Wilka brzmi, że dane pochodzą z rozkładu normalnego. - Ponieważ wartość p ($p = 0.0102$) jest mniejsza niż poziom istotności $\alpha = 0.05$, odrzucamy hipotezę zerową i stwierdzamy, że dane nie pochodzą z rozkładu normalnego.

6.2 Test Kołmogorowa-Smirnowa

Test Kołmogorowa-Smirnowa sprawdza, czy dane są zgodne z rozkładem normalnym. W tym przypadku, porównujemy dane z rozkładem normalnym o średniej i odchyleniu standardowym odpowiadającym danym.

$$\text{Statystyka testu} = 0.0895, \quad p\text{-value} = 0.1706$$

Interpretacja: - Hipoteza zerowa testu Kołmogorowa-Smirnowa brzmi, że dane pochodzą z rozkładu normalnego. - Ponieważ wartość p ($p = 0.1706$) jest większa niż poziom istotności $\alpha = 0.05$, nie odrzucamy hipotezy zerowej i przyjmujemy, że dane pochodzą z rozkładu normalnego.

6.3 Wnioski

Wyniki testów wskazują na sprzeczność w wynikach testów normalności: - Test Shapiro-Wilka sugeruje, że dane nie pochodzą z rozkładu normalnego (odrzucamy hipotezę zerową). - Test Kołmogorowa-Smirnowa sugeruje, że dane są zgodne z rozkładem normalnym (nie odrzucamy hipotezy zerowej).

Warto zauważyć, że różnice te mogą wynikać z różnych założeń testów oraz charakterystyki samych danych. Test Shapiro-Wilka jest bardziej wrażliwy na odchylenia od normalności w małych próbkach, podczas gdy test Kołmogorowa-Smirnowa jest bardziej ogólny i może być mniej wrażliwy na specyficzne odchylenia w rozkładzie.

7 Test istotności dla współczynnika korelacji

Celem zadania było wygenerowanie dwóch zestawów danych: skorelowanych i nieskorelowanych, a następnie przeprowadzenie testu istotności dla współczynników korelacji Pearsona i Spearmana.

7.1 Dane

- **Zestaw skorelowany:** $y = 2x + \varepsilon$, gdzie $\varepsilon \sim \mathcal{N}(0, 0{,}5)$.
- **Zestaw nieskorelowany:** $x \sim \mathcal{N}(0, 1)$ i $y \sim \mathcal{N}(0, 1)$ niezależnie.

7.2 Wyniki

Dane skorelowane:

- Współczynnik korelacji Pearsona: $r = 0{,}9655$, p-value $< 0{,}0001$
- Współczynnik korelacji Spearmana: $\rho = 0{,}9640$, p-value $< 0{,}0001$

Interpretacja: W obu przypadkach odrzucamy hipotezę zerową o braku korelacji – istnieje istotna statystycznie korelacja między zmiennymi.

Dane nieskorelowane:

- Współczynnik korelacji Pearsona: $r = -0{,}0003$, p-value $= 0{,}9980$
- Współczynnik korelacji Spearmana: $\rho = 0{,}0266$, p-value $= 0{,}7928$

Interpretacja: W obu przypadkach nie ma podstaw do odrzucenia hipotezy zerowej – brak istotnej statystycznie korelacji.

Wizualizacja danych

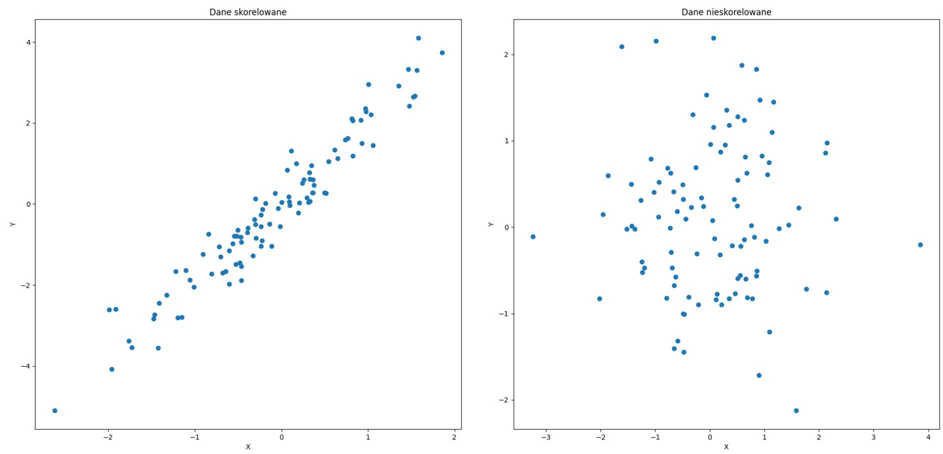

Fig... 1: Porównanie danych skorelowanych (lewy wykres) i nieskorelowanych (prawy wykres).