Analiza wariancji

Weronika Przysiężna

Marzec 2025

1 Teoretyczne wprowadzenie do analizy wariancji

1.1 Wstęp

Analiza wariancji jest analizą statystyczną, która wykrywa różnice między dwiema lub więcej grupami określonymi dla pojedynczego czynnika lub zmiennej niezależnej. Identyfikuje ona zmienność lub wariancję pomiędzy obserwacjami przypisując ją różnym źródłom, które (po odpowiednim przetestowaniu) wskazują, czy zaobserwowane różnice między średnimi grupowymi są prawdopobnie rzeczywiste, czy jedynie wynikiem przypadku. [5]

W naszym badaniu skupimy się na wykrywaniu różnic przy pomiarach na różnych osobach.

1.2 Założenia dotyczące jednoczynnikowej analizy wariancji (ANOVA) z jednym czynnikiem międzyobiektowym [3]

- **Zmienna zależna mierzona na skali ilościowej**: zmienna zależna powinna być zmienną ilościową (na poziom interwałowym lub ilorazowym).
- **Losowość i niezależność obserwacji**: nie ma związku między obserwacjami w każdej grupie lub między samymi grupami, a w każdej grupie są różni uczestnicy badania i żaden uczestnik nie należy do więcej niż jednej grupy; uczestnicy badania są dobierani losowo.
- **Równoliczność obserwacji w grupach**: poszczególne kategorie zmiennej niezależnej powinny być statystycznie równoliczne (aby sprawdzić, czy analizowane grupy różnią się istotnie statycznie pod względem liczebności, można zastosować test zgodności Chi-kwadrat).
- **Rozkład normalny**: rozkład wyników w analizowanych grupach jest zbliżony do rozkładu normalnego (oceny tego założenia można dokonać stosując test Kołomogorowa-Smirnova lub Shapiro-Wilka).
- **Wariancje w grupach są jednorodne (homogeniczność wariancji)**: zmienność w każdej porównywanej grupie powinna być podobna; jeśli wariancje różnią się między grupami, to można zastosować test Welcha lub Browna-Forsythe'a, które wprowadzają poprawkę na nierówne wariancje do statystyki F.

1.3 Jednoczynnikowa ANOVA

1.3.1 Suma kwadratów SS (ang. Sum of Squares)

Wariancja próby mierzy zmienność w dowolnym zbiorze obserwacji poprzez obliczenie sumy kwadratów odchyleń od ich średniej:

$$SS = \sum (X - \overline{X})^2.$$

Następnie suma kwadratów SS jest dzielona przez liczbę stopni swobody $n - 1$:

$$s^2 = \frac{SS}{df},$$

gdzie:

- $\overline{X}$ - średnia próby,
- s^2 - wariancja próby,
- $df = n - 1$ - stopnie swobody.

1.3.2 Średnia kwadratów MS (ang. Mean Square)

Średnia kwadratów to oszacowanie wariancji uzyskane przez podzielenie sumy kwadratów SS przez liczbę stopni swobody $n - 1$. Ogólny wzór na oszacowanie wariancji ma postać:

$$MS = \frac{SS}{df},$$

gdzie:

- MS - średnia kwadratów,
- SS - suma kwadratów odchyleń od średniej,
- $df = n - 1$ - liczba stopni swobody.

1.3.3 Wzory definicyjne na sumy kwadratów [5]:

1. SS_{total} - całkowita suma kwadratów odchyleń od średniej ogólnej (zmienność całkowita)
$$SS_{total} = \sum (X - \overline{X}_{grand})^2.$$
Równoważny wzór obliczeniowy:
$$SS_{total} = \sum X^2 - \frac{G^2}{N}.$$

2. $SS_{between}$ - suma kwadratów odchyleń średnich grupowych od średniej ogólnej (zmienność między grupami)
$$SS_{between} = \sum n(\overline{X}_{group} - \overline{X}_{grand})^2.$$
Równoważny wzór obliczeniowy:
$$SS_{between} = \sum \frac{T^2}{n} - \frac{G^2}{N}.$$

3. SS_{within} - suma kwadratów odchyleń indywidualnych wyników w grupie od średnich grupowych (zmienność wewnątrz grup)
$$SS_{within} = \sum (X - \overline{X}_{group})^2.$$
Równoważny wzór obliczeniowy:
$$SS_{within} = \sum X^2 - \sum \frac{T^2}{n}.$$

4. Sprawdzamy dokładność obliczeniową weryfikując równość:
$$SS_{total} = SS_{between} + SS_{within}$$

Oznaczenia:

- X - pojedyncza wartość obserwowana,
- $\overline{X}_{group}$ - średnia dla danej grupy,
- $\overline{X}_{grand}$ - średnia ogólna dla całej próby,
- T - suma wartości w grupie,
- n - liczba obserwacji w grupie,
- G - suma wartości dla wszystkich grup (suma ogólna),
- N - całkowita liczba obserwacji (sumaryczna wielkość próby).

1.3.4 Stopnie swobody (df):

1. $df_{total} = N - 1$,
2. $df_{between} = k - 1$,
3. $df_{within} = N - k$,

gdzie:

- N - całkowita liczba obserwacji (sumaryczna wielkość próby),
- k - liczba grup.

Sprawdzamy dokładność obliczeń weryfikując równość:

$$df_{total} = df_{between} + df_{within}.$$

1.3.5 Wzory na średnie kwadratów [5]:

1. $MS_{between}$ - średni kwadrat odchyleń między grupami (zmienność między średnimi dla grup)
$$MS_{between} = \frac{SS_{between}}{df_{between}}$$
2. MS_{within} - średni kwadrat odchyleń wewnątrz grupy (zmienność wyników wewnątrz grupy; mierzy jedynie błąd losowy)
$$MS_{within} = \frac{SS_{within}}{df_{within}} = MS_{error}$$

1.3.6 Rozkład F-Snedecora

Liczymy statystykę testową F jako:

$$F = \frac{MS_{between}}{MS_{within}}$$

Określamy obszar krytyczny jako:

$$Q = \{F : F \geq F_\alpha\},$$

gdzie F_α jest wartością krytyczną odczytaną z tablic rozkładu F-Snedecora dla $(df_{between}, df_{within})$ stopni swobody, czyli $F(df_{between}, df_{within})$.

1. Jeżeli $F \in Q$ $(F \geq F_\alpha)$, to odrzucamy hipotezę zerową H_0 na korzyść hipotezy alternatywnej H_A i wnioskujemy, że badane średnie nie są sobie równe.
2. Jeżeli $F \notin Q$ $(F < F_\alpha)$, to nie ma podstaw do odrzucenia hipotezy zerowej H_0 i wnioskujemy, że badane średnie są równe.

1.3.7 Testy post-hoc (test HSD Tukey'a)

1.3.8 Obliczenie siły efektu (d Cohen'a)

2 Praktyczne zastosowanie analizy wariancji

W celu zilustrowania praktycznego zastosowania analizy wariancji przeprowadziłam badania na dwóch zestawach danych:

1. Dane uzyskane za pośrednictwem ankiet, których celem jest analiza hipotetycznych zależności między czasem poświęcanym tygodniowo na aktywność fizyczną a nawykami oraz jakością życia.

2. Dane udostępnione na platformie Kaggle.com, umożliwiające zbadanie potencjalnych zależności między [...] [1]

W dalszej części omówię proces pozyskania obu zbiorów danych oraz przedstawię w jaki sposób przeprowadziłam analizę statystyczną danych.

2.1 Badanie zależności między czasem poświęcanym na aktywność fizyczną a jakością życia

2.1.1 Metodologia pozyskiwania danych ankietowych

Dane zostały zebrane za pomocą ankiety utworzonej w Google Forms i udostępnionej w mediach społecznościowych, w tym na grupach na Facebooku zrzeszających osoby aktywne fizycznie oraz studentów, a także na prywatnych profilach na Facebooku i Instagramie, skierowanych do znajomych.

Grupą badaną są młodzi dorośli w wieku 18-35 lat, posługujący się językiem polskim, posiadający dostęp do internetu oraz do mediów społecznościowych.

[1] Dane wykorzystane w analizie pochodzą z platformy Kaggle.com i zostały syntetycznie wygenerowane. Oznacza to, że nie są wynikiem rzeczywistych pomiarów, lecz zostały stworzone na podstawie określonych założeń i symulacji.

Pytaliśmy uczestników badania o ich wiek, płeć oraz odpowiedzi na poniższe pytania:

1. Ile czasu średnio tygodniowo poświęcasz na aktywność fizyczną? (np. siłownia/rower/bieganie/taniec/joga)
2. Ile czasu średnio tygodniowo spędzasz na pozasportowych spotkaniach towarzyskich? Chodzi o spotkania ze znajomymi poza szkołą/miejscem pracy.
3. Ile czasu średnio dziennie spędzasz przed ekranem poza pracą/szkołą? (TV, komputer, telefon)
4. Ile czasu średnio śpisz w ciągu doby?
5. Ile razy chorowałeś/chorowałaś w ciągu ostatnich 12 miesięcy? (przeziębienie, grypa, choroby inne niż przewlekłe)
6. Na ile (w skali 1 - 10) oceniasz swoje ogóle poczucie szczęścia i zadowolenia z życia?

Możliwe do wyboru odpowiedzi dla poszczególnych pytań prezentuje tabela:

Pytania 1, 2, 3, 4	**Pytanie 5**	**Pytanie 6**
mniej niż 1 godzinę	0 - wcale	
1 - 2 godziny	1 raz	1 - bardzo źle
2 - 3 godziny	2 razy	2
3 - 4 godziny	3 razy	3
4 - 5 godzin	4 razy	4
5 - 6 godzin	5 razy	5
6 - 7 godzin	6 razy	6
7 - 8 godzin	7 razy	7
8 - 9 godzin	8 razy	8
więcej niż 9 godzin	9 razy	9
	10 - 10 lub więcej niż 10 razy	10 - bardzo dobrze

Tabela 1

2.1.2 Wyniki ankiety

W wyniku ankiety uzyskaliśmy 597 odpowiedzi.

Uzyskane dane wymagały przetworzenia w formę umożliwiającą ich dalszą analizę, tak aby odpowiedzi na pytania 1-6 były zapisane jako dane numeryczne. W tym celu opracowałam program w języku Python, który automatycznie wykonuje tę operację:

```
import pandas as pd

dane = pd.read_excel('dane.xlsx')        # wczytanie pliku

dane = dane.iloc[:, 1:]                  # usunięcie 1. kolumny

dane.columns = ['wiek', 'płeć',
                'akt_fiz', 'spot_tow',
                'ekran', 'sen',
                'choroby', 'szczescie'] # zmiana nazw kolumn

dane = dane[dane['wiek'] <= 35]          # usunięcie wierszy,
                                         # w których wiek > 35

dane = dane.reset_index(drop=True)       # resetowanie indeksów

# funkcja do konwersji wartosci w kolumnach
def konwertuj(wartosc):
    if 'mniej niż' in wartosc:
        return 0
    elif 'więcej niż' in wartosc:
        return 9
    elif '-' in wartosc:
        return int(wartosc.split('-')[0].strip())
    elif 'Kobieta' in wartosc:
        return 'K'
    elif 'Mężczyzna' in wartosc:
        return 'M'
    elif 'Inne' in wartosc:
        return 'I'

dane['płeć'] = dane['płeć'].astype(str).apply(konwertuj)
dane['akt_fiz'] = dane['akt_fiz'].astype(str).apply(konwertuj)
dane['spot_tow'] = dane['spot_tow'].astype(str).apply(konwertuj)
dane['ekran'] = dane['ekran'].astype(str).apply(konwertuj)
dane['sen'] = dane['sen'].astype(str).apply(konwertuj)

print(dane)
```

W rezultacie pozostały 522 wiersze z odpowiedziami.

	wiek	płeć	akt_fiz	spot_tow	ekran	sen	choroby	szczescie
0	31	M	9	7	4	6	2	7
1	25	M	2	3	2	6	5	7
2	29	M	2	5	5	7	4	3
3	25	M	6	0	9	6	3	4
4	29	M	3	2	7	7	1	10
..	...	...	...	...	...	...	...	...
517	25	M	6	7	3	8	1	9
518	32	M	0	3	3	5	1	7
519	32	M	1	9	1	7	5	8
520	25	M	0	2	3	8	1	5
521	25	M	0	0	9	6	2	4

2.1.3 Wstępna analiza zebranych danych

Wywołanie `describe` dla wybranych kolumn pozwoliło na uzyskanie zbiorczego opisu statystycznego tych zmiennych.

```
print(dane[['wiek', 'płeć', 'akt_fiz', 'spot_tow', 'ekran',
'sen', 'choroby', 'szczescie']].describe(include='all'))
```

W wyniku wywołania powyższej linijki kodu otrzymujemy:

	wiek	płeć	akt_fiz	spot_tow	ekran	sen	choroby	szczescie
count	522.000000	522	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000
unique	NaN	3	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	M	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	373	NaN	NaN	NaN	NaN	NaN	NaN
mean	26.168582	NaN	3.340996	3.105364	3.894636	6.358238	2.055556	6.281609
std	3.857631	NaN	2.825475	2.753189	2.415245	1.185331	1.784659	2.155920
min	18.000000	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	24.000000	NaN	1.000000	1.000000	2.000000	6.000000	1.000000	5.000000
50%	26.000000	NaN	3.000000	2.000000	3.000000	6.000000	2.000000	7.000000
75%	29.000000	NaN	5.000000	5.000000	5.000000	7.000000	3.000000	8.000000
max	35.000000	NaN	9.000000	9.000000	9.000000	9.000000	10.000000	10.000000

Funkcja zwróciła m.in.:

- liczność (`count`) – liczbę dostępnych wartości,
- średnią (`mean`) – wartość przeciętną,
- odchylenie standardowe (`std`) – miarę rozproszenia danych,
- minimalną i maksymalną wartość (`min`, `max`),
- percentyle (`25%`, `50%` (mediana), `75%`), które pokazują, jak dane są rozłożone.
- liczność (`count`),
- liczba unikalnych wartości (`unique`),
- najczęściej występująca wartość (`top`),
- częstość występowania tej wartości (`freq`).

Dzięki tym informacjom możliwe było uzyskanie pierwszego wglądu w rozkład danych.

2.1.4 Wstępna analiza zebranych danych

2.2 Badanie zależności między [...]

2.2.1 Opis danych

2.2.2 Analiza statystyczna zebranych danych

3 Bibliografia

Pozycje [1], [2], [4] służyły jako pomoc przy pisaniu kodów w LaTeX i Python.

Literatura

[1] Kamil Kozłowski. Latexpolishleters.tex, 2016.

[2] Inc NumFOCUS. pandas, 2024.

[3] Norm O'Rourke, Larry Hatcher, and Edward J. Stepanski. *A Step-by-Step Approach to Using SAS for Univariate and Multivariate Statistics*. SAS Institute, Cary, NC, USA, 2th edition, 2005.

[4] statsmodels. 0.15.0 user guide, 2025.

[5] Robert S. Witte and John S. Witte. *Statistics*. John Wiley & Sons, Hoboken, NJ, USA, 9th edition, 2010.