Metody Statystyczne II

Projekt zaliczeniowy

Analiza statystyczna w ramach modelu Neymana-Rubina na podstawie danych Students Performance

Emilia Matrejek 100916

Studia magisterskie sobotnio-niedzielne, semestr zimowy 2024/2025

Warszawa, 10.01.2025 r.

Spis treści

Cel raportu i opis problemu badawczego

Celem tego projektu jest zaprezentowanie praktycznego zastosowania modelu Neymana-Rubina jako narzędzia umożliwiającego analizę przyczynowego wpływu określonego czynnika na zmienną zależną. W ramach tej analizy skupiamy się na danych dotyczących wyników edukacyjnych uczniów, aby odpowiedzieć na postawione pytanie badawcze: **Czy bardziej pozytywnie na średnią ocen (GPA) wpływa otrzymywanie korepetycji czy zajęcia pozalekcyjne?**

Średnia ocen (GPA) została wybrana jako wskaźnik poziomu osiągnięć edukacyjnych. Przyjęto, że uczestnictwo w zajęciach pozalekcyjnych oraz pobieranie korepetycji mogą mieć pozytywny wpływ na wyniki uczniów, ponieważ tego rodzaju aktywności rozwijają różnorodne umiejętności, sprzyjają lepszemu zarządzaniu czasem oraz zwiększają zaangażowanie uczniów w proces nauki. Dodatkowo, indywidualny kontakt ucznia z nauczycielem podczas korepetycji ułatwia przyswajanie wiedzy i umożliwia dostosowanie trybu, szybkości oraz sposobu przekazywania wiedzy w zależności od potrzeb.

W toku analizy przeprowadzona zostanie krótka eksploracja danych w zakresie koniecznym dla analizy, wybór odpowiedniej metody statystycznej w ramach modelu Neymana-Rubina, a także ocena wyników za pomocą wizualizacji i interpretacji statystycznej. Ostatecznie projekt dostarczy odpowiedzi na pytanie, która z aktywności - uczestnictwo w zajęciach pozalekcyjncyh czy korepetycje - wpływają bardziej korzystnie na średnią ocen uczniów, oraz pozwoli na dyskusję ograniczeń i potencjalnych implikacji uzyskanych wyników.

Opis zbioru danych

W projekcie skorzystano ze zbioru danych Student_Performance_Dataset, dostępnego na platformie OpenML (ID: 46255). Zbiór ten zawiera

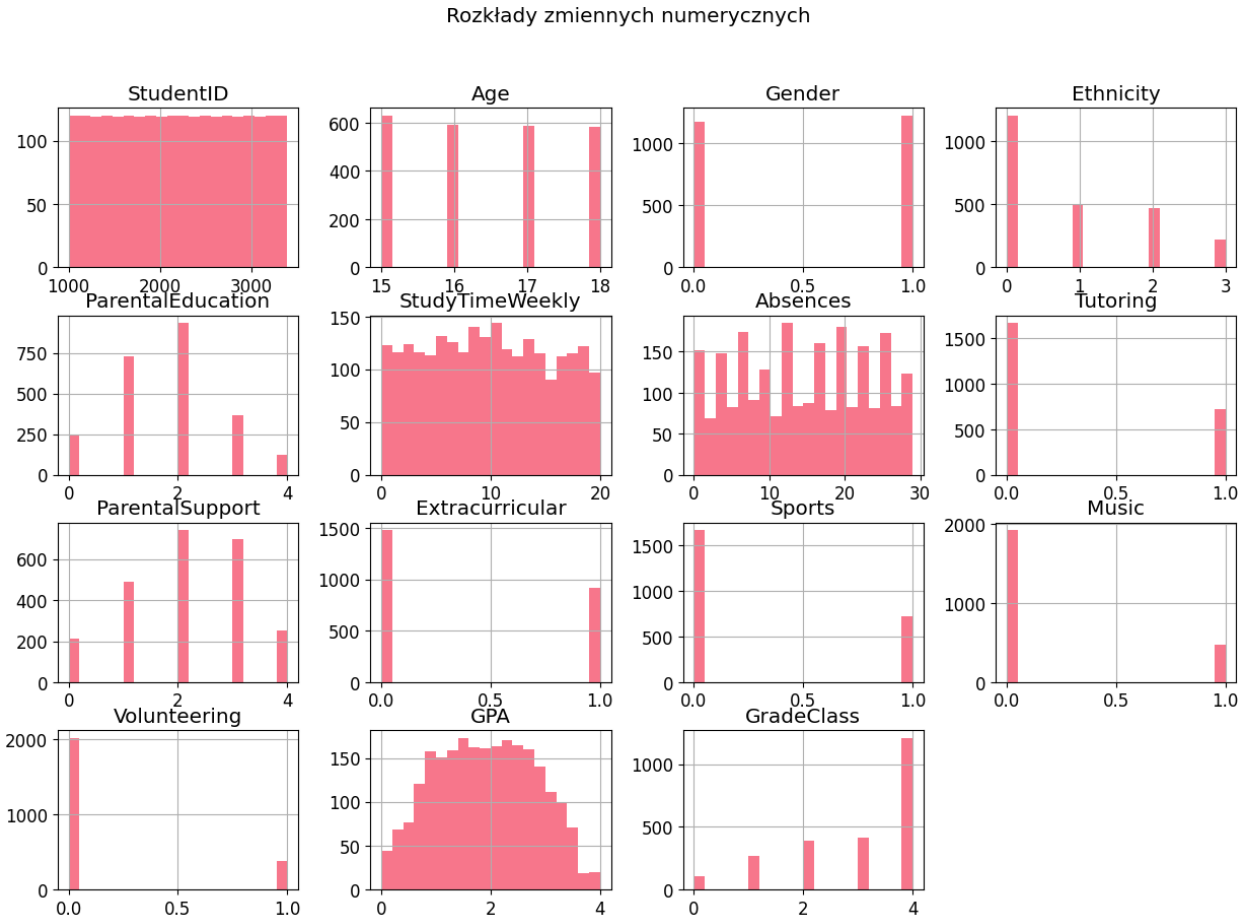
Zbiór danych Student_Performance_Data zawiera informacje o różnych czynnikach wpływających na wyniki uczniów. Zawiera dane demograficzne, nawyki związane z nauką, uczestnictwo w zajęciach dodatkowych oraz osiągnięcia edukacyjne. Dzięki temu można przeanalizować, jak różne aspekty, takie jak wsparcie rodziców, czas poświęcany na naukę czy udział w zajęciach pozalekcyjnych, wpływają na średnią ocen (GPA) uczniów.

Zbiór danych składa się z 2392 obserwacji i 15 zmiennych – wszystkie z nich są numeryczne, przy czym zmienne takie jak Ethnicity czy Gender zostały uprzednio zakodowane (zawierają wartości numeryczne odpowiadające konkretnym grupom etnicznym lub płci męskiej/żeńskiej).

Odpowiedź na pytanie, czy bardziej pozytywnie na średnią ocen wpływa otrzymywanie korepetycji czy zajęcia pozalekcyjne może mieć istotne znaczenie praktyczne – jeśli okaże się, że wpływ zajęć pozalekcyjncyh jest istotnie bardziej korzystny, może to zachęcić szkoły do większych inwestycji w tego rodzaju programy, szczególnie dla uczniów potrzebujących wsparcia. Natomiast brak związku lub negatywny wpływ mogą prowadzić do refleksji nad tym, jakie inne formy wsparcia mogłyby być bardziej efektywne dla uczniów. Z drugiej strony, jeśli okaże się, że to korepetycje mają większy wpływ na poprwę średniej ocen, być może należy zadać sobie pytanie, dlaczego dopiero korzystanie z płatnej, indywidualnej formy pomocy umożliwia poprawę wyników w nauce – optymalne byłoby zrównanie efektu zajęć pozalekcyjnych, aby dać dostęp wszystkim młodym ludziom, również z biedniejszych rodzin, do rozwijania swoich umiejętności.

Prezentowany zbiór danych jest dobrym zbiorem do przeprowadzenia takiego badania, ponieważ zawiera wystarczająco dużo obserwacji, do tego nie zawiera braków danych, a zawarte informacje wydają się być wiarygodne – nie zawiera wartości odstających, a rozkład istniejących wartości jest dosyć równomierny, co pokazano na wykresie 1.

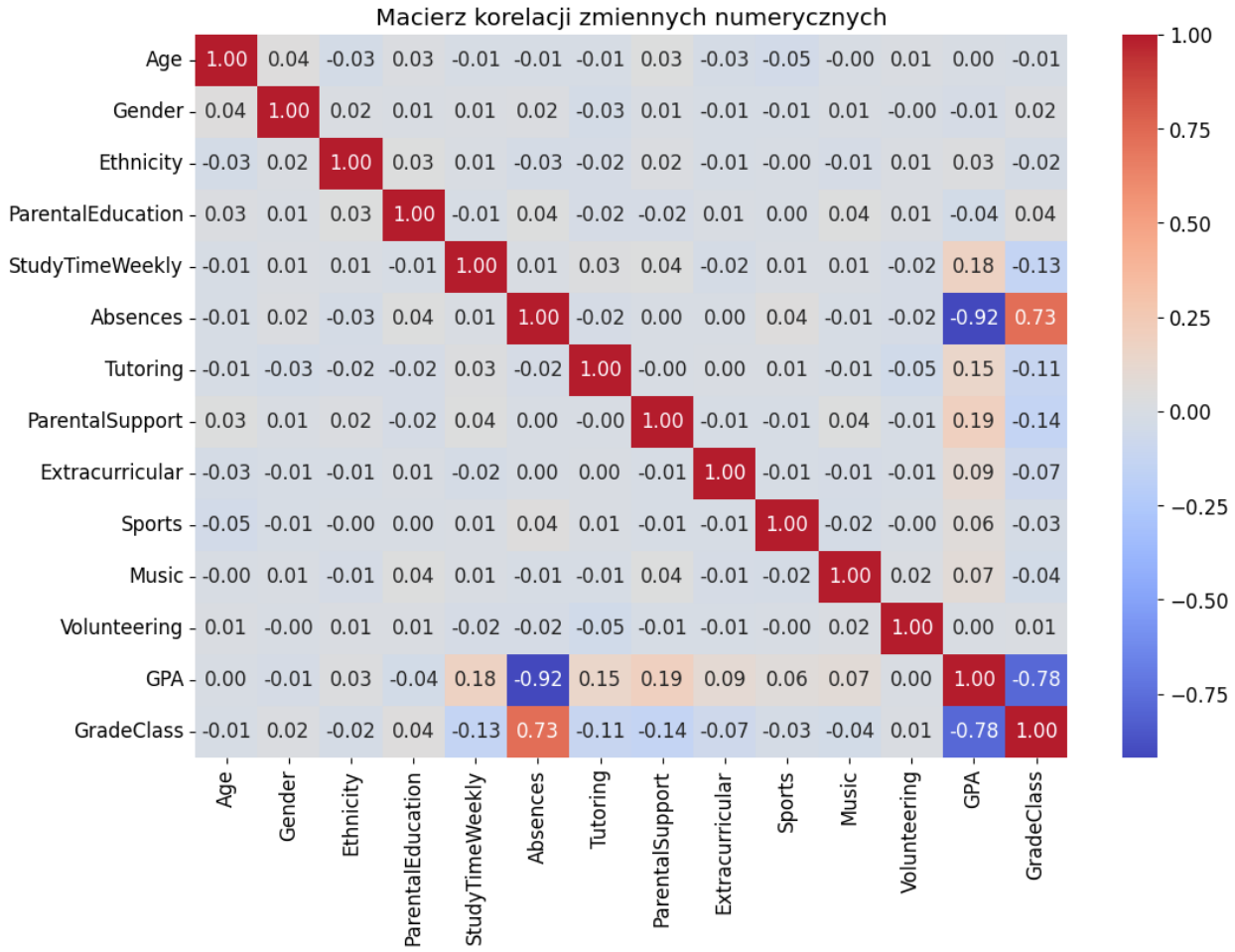
Wykres 1. Histogramy zmiennych występujących w zbiorze danych Students_Performance_Data.



Jak widać na Wykresie 1, rozkład średniej ocen (GPA) jest zbliżony do rozkładu normalnego. Jeśli natomiast chodzi o podział na grupę „treatment" oraz grupę kontrolną („control") w obrębie zmiennej Extracurricular (uczestnictwo w zajęciach pozalekcyjnych(, to w zbiorze znajduje się 917 uczestników zajęć pozalekcyjnych oraz 1475 osób nieuczestniczących w zajęciach. W przypadku zmiennej Tutoring (pobieranie korepetycji) – 721 uczniów uczęszczało na korepetycje, a 1671 nie. Pod tym względem próba jest niezbilansowana – liczba osób nieuczestniczących w zajęciach jest o około 60% wyższa od liczby uczestników zajęć pozalekcyjnych, a liczba osób niepobierających korepetycji jest o około 132% wyższa niż osób z nich korzystających.

W celu wstępnego zbadania, jakiego można spodziewać się wyniku analizy, wygenerowano macierz korelacji (Wykres 2.).

Wykres 2. Macierz korelacji.



Z powyżej macierzy można odczytać, że zmienna Tutoring jest znacznie silniej pozytywnie skorelowana ze zmienną wynikową GPA (współczynnik korelacji 0.15) niż zmienna Extracurricular (współczynnik korelacji 0.09), zatem prawdopodobnie silniejszy pozytywny wpływ wykazywać będzie pobieranie korepetycji.

Dobór metody analizy

Opis modelu

W ramach analizy porównawczej badano, który z dwóch czynników – uczestnictwo w zajęciach pozalekcyjnych (zmienna Extracurricular) czy pobieranie korepetycji (zmienna Tutoring) – wywiera większy pozytywny wpływ na średnią ocen uczniów (GPA). Analiza została przeprowadzona z zastosowaniem modelu Neymana-Rubina, który pozwala na oszacowanie efektów przyczynowych (ATE, ATT, ATC) w przypadku danych obserwacyjnych. Przyjęto, że zmienne Extracurricular i Tutoring są zmiennymi „traktowania”, a zmienna GPA – zmienną

wynikową. Do oszacowania efektów przyczynowych wykorzystano metodę dopasowania najbliższego sąsiada (nearest neighbor matching), która pozwala na porównanie wyników uczniów uczestniczących w zajęciach/pobierających korepetycje z wynikami ich najbardziej podobnych rówieśników, którzy w takich zajęciach nie uczestniczyli.

Założenia modelu

1. Losowy przydział wartości T – zakładamy, że dane pochodzą z eksperymentu, a więc przydział do poszczególnych grup ($T = 0, T = 1$) jest losowy. Jeśli korzystamy z danych empirycznych, możliwa jest tzw. emulacja eksperymentu – można spróbować znaleźć takie zmienne charakteryzujące jednostki, które tłumaczą proces rozdysponowania wartości T w populacji. Następnie można próbować modelować proces przydziału T.
2. Overlap – każda jednostka ma niezerowe prawdopodobieństwo przypisania do obu grup
3. SUTVA – stabilność i niezależność efektów interwencji:
    a. Brak interferencji między jednostkami – wynik jednej jednostki nie zależy od przypisania interwencji do innych jednostek
    b. Jednoznaczność interwencji – dla każdej jednostki efekt przypisania jest jednoznacznie określony

Weryfikacja spełnienia założeń w kontekście przeprowadzanego badania

Ad. 1.

W przypadku zbioru danych Students_Performance_Data, z pewnością uczniowie nie byli „przydzielani" do grup pobierających korepetycje lub korzystających z zajęć pozalekcyjnych, więc nie ma tutaj mowy o danych eksperymentalnych. Ponadto, nie są znane charakterystyki uczniów korzystających z tych świadczeń w populacji, zatem emulacja eksperymentu byłaby trudna.

Ad. 2.

Teoretycznie, każdy z uczniów ma możliwość uczestniczenia w konkretnych zajęciach, zatem prawdopodobieństwo przypisania ich do jednej z grup można określić jako niezerowe.

Ad. 3.

Założenie to jest w pełni spełnione – uczestnictwo jednego ucznia w zajęciach nie ogranicza w żadnej sposób możliwości uczestnictwa innych uczniów; dodatkowo uczestnictwo jest jednoznacznie określone.

Etapy analizy

1. Zdefiniowanie zmiennych „traktowania" i wynikowych:
    a. Zmienne „traktowania"
        i. Extracurricular: uczestnictwo w zajęciach pozalekcyjncyh ($T = 1$) lub brak uczestnictwa ($T = 0$).
        ii. Tutoring: pobieranie korepetycji ($T = 1$) lub brak korepetycji ($T = 0$)
    b. Zmienna wynikowa:
        i. Średnia ocen uczniów (GPA) – przedstawiona na skali ciągłej
2. Zdefiniowanie zmiennych kontrolnych – aby umożliwić dopasowanie najbliższego sąsiada, włączono zmienne kontrolne takie jak wiek (Age), wykształcenie rodziców (ParentalEducation), tygodniowy czas nauki (StudyTimeWeekly), liczba nieobecności (Absences). Dane zostały przeskalowane za pomocą standaryzacji.
3. Dopasowanie najbliższego sąsiada – za pomocą metody Nearest Neighbor
    a. Obliczenie efektów ATT, ATC, ATE
        i. ATE (Average Treatment Effect) - średni wpływ zajęć pozalekcyjnych na średnią ocen w całej populacji.
        ii. ATT (Average Treatment Effect on the Treated) - średni wpływ zajęć na średnią ocen w grupie uczniów, którzy uczestniczyli w zajęciach.
        iii. ATC (Average Treatment Effect on the Controls) – średni wpływ zajęć, gdyby uczniowie z grupy kontrolnej uczestniczyli w zajęciach.
4. Obliczenie błędów standardowych i przedziałów ufności – co pozwala na ocenę precyzji oszacowań.
5. Wizualizacja efektów – wykresy gęstości, rozkłady efektów w obu grupach, różnice pomiędzy efektami ATT, ATC, ATE dla obu zmiennych

Realizacja analizy w języku Python

Analiza została przeprowadzona w języku Python z wykorzystaniem odpowiednich bibliotek statystycznych i narzędzi wizualizacyjnych.

1. Przygotowanie danych – biblioteka pandas

   Dane wczytano za pomocą biblioteki pandas. Przed przystąpieniem do analizy, przeprowadzono wstępną eksplorację w celu weryfikacji integralności danych oraz zidentyfikowania brakujących wartości i potencjalnych outlierów. Funkcje w ramach biblioteki pandas pozwalają na zrozumienie struktury danych. Dane są kompletne i zgodne z założeniem istnienia grup kontrolnej i „traktowania".
2. Przygotowanie zmiennych kontrolnych – biblioteka sklearn.preprocessing

   W celu kontroli zmiennych zakłócających, wybrano atrybuty takie jak wiek, poziom wykształcenia rodziców, tygodniowy czas nauki i liczba nieobecności. Dane zostały przeskalowane za pomocą funkcji StandardScaler z biblioteki sklearn, co zapewnia porównywalność między zmiennymi o różnych skalach.
3. Dopasowanie grup za pomocą Nearest Neighbor Matching – biblioteka sklearn.neighbors

   Zastosowano metodę dopasowania najbliższego sąsiada (NearestNeighbors) w celu dopasowania uczniów z grup traktowania i kontrolnej. Dopasowanie przeprowadzono osobno dla dwóch zmiennych: uczestnictwa w zajęciach pozalekcyjnych (Extracurricular) i korepetycji (Tutoring). Funkcja kneighbors umożliwia identyfikację odpowiedników w grupie kontrolnej i traktowania.
4. Oszacowanie efektów ATE, ATT, ATC

   Za pomocą prostych różnic średnich i wartości dopasowanych obliczono trzy główne wskaźniki: ATE, ATT, ATC. Skorzystano z funkcji obliczających średnie (mean) oraz przedziały ufności (var).
5. Wizualizacja danych – biblioteki matplotlib, seaborn

   Skorzystano z funkcji sns.kdeplot oraz plt.bar aby odzworować rozkłady i różnice między grupami. Ułatwia to interpretację efektów.

Ocena zgodności z założeniami analizy

Cała analiza została zrealizowana zgodnie z założeniami modelu Neymana-Rubina. Zastosowane metody zapewniają:

- Minimalizację wpływu zmiennych zakłócających dzięki dopasowaniu najbliższego sąsiada.
- Precyzyjne oszacowanie efektów przyczynowych (ATE, ATT, ATC).
- Wizualną prezentację wyników, co ułatwia ich interpretację i komunikację.

Wyniki analizy

Wyniki główne

W Tabeli 1 przedstawiono wartości wskaźników dla badanych zmienncyh: Extracurriculum oraz Tutoring.

Table 1. Porównanie wyników głównych - efekty przyczynowe.

Korepetycje:		**Zajęcia pozalekcyjne:**	
ATE	0.289 ± 0.04	ATE	0.177 ± 0.038
ATT	0.248 ± 0.048	ATT	0.172 ± 0.043
ATC	0.241 ± 0.031	ATC	0.186 ± 0.034
Przedziały ufności:		Przedziały ufności:	
ATE	$(0.2103, 0.3684)$	ATE	$(0.1017, 0.2523)$
ATT	$(0.1545, 0.3415)$	ATT	$(0.0885, 0.2556)$
ATC	$(0.1794, 0.3022)$	ATC	$(0.1206, 0.2523)$

Według powyższej tabeli, pobieranie korepetycji powiązane było w badanej próbie ze średnią ocen wyższą przeciętnie o 0.289, natomiast uczesnictwo w zajęciach pozalekcyjnych – wyższą o 0.177. U osób pobierających korepetycje.

Wśród osób, które faktycznie wzięły korepetycje, średnia ocen podnosi się przeciętnie o 0.248, a u osób uczestniczących w zajęciach pozalekcyjnych – o 0.172.

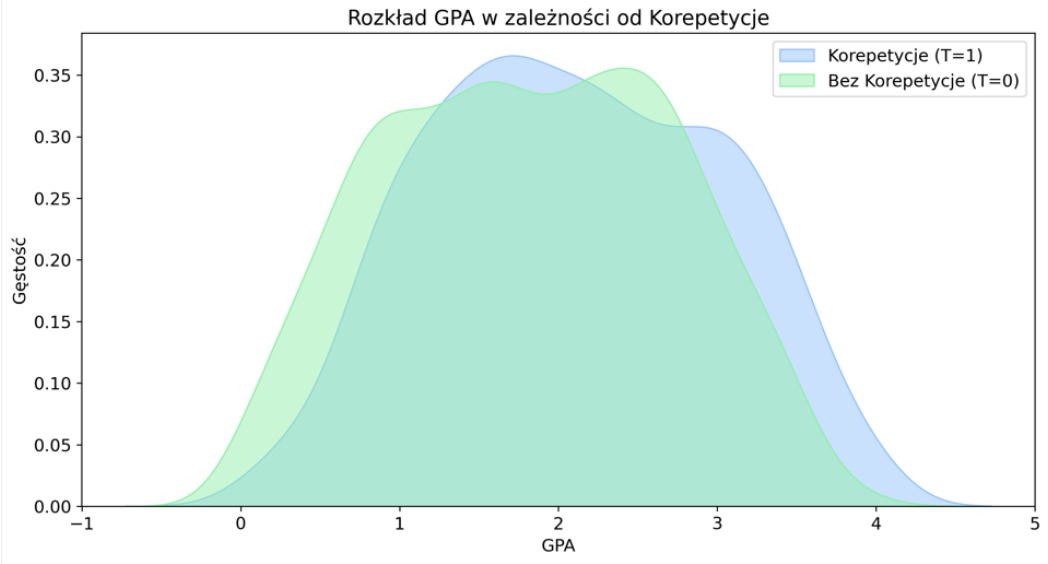
Wśród osób, które nie otrzymały korepetycji, przewidywany wzrost średniej wyniósłby przeciętnie 0.241, natomiast w przypadku uczestnictwa w zajęciach pozalekcyjnych byłoby to przeciętnie 0.186.

Można zatem wywnioskować, ze korepetycje mają silniejszy wpływ na poprawę średniej ocen, niż uczestnictwo w zajęciach pozalekcyjnych.

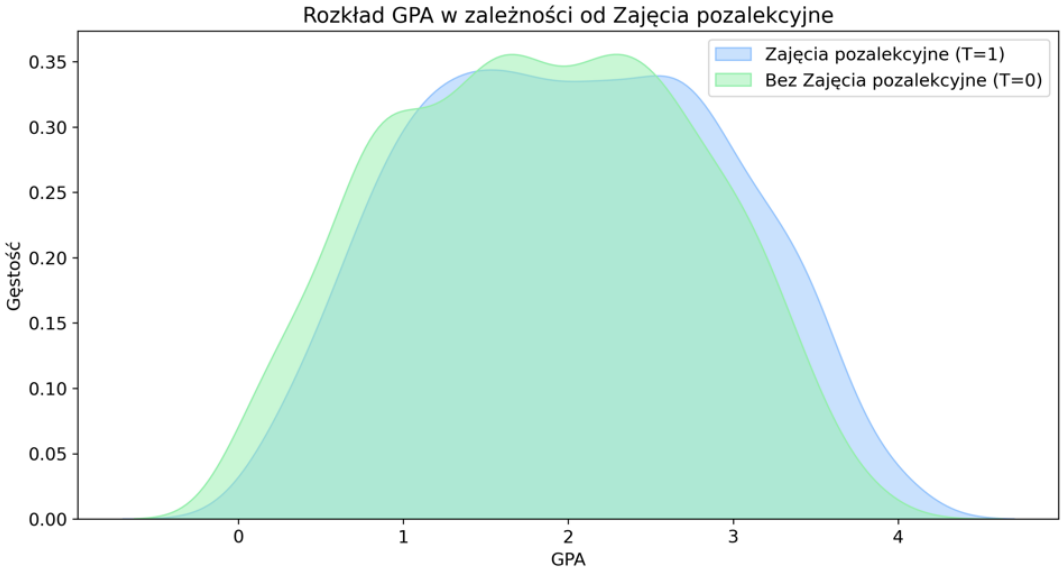
Wizualizacje

Na wykresach 3 i 4 przedstawiono wykresy gęstości dla grupy „treatment” oraz kontrolnej dla poszczególnych zmiennych.

Wykres 3. Rozkład zmiennej GPA w zależności od wartości zmiennej Tutoring.



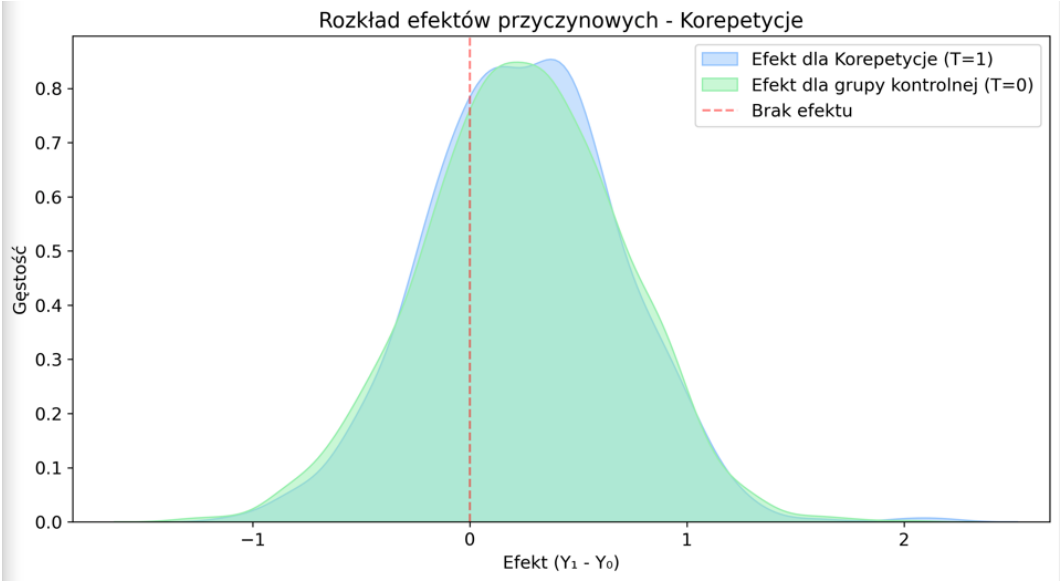
Wykres 4. Rozkład zmiennej GPA w zależności od wartości zmiennej Extracurricular.



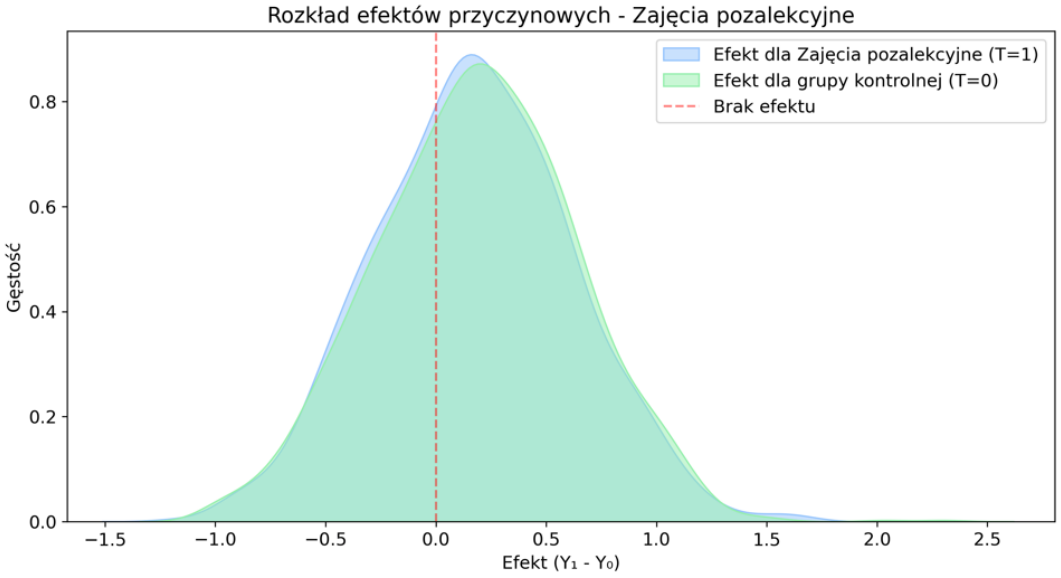
Na powyższych wykresach widać, że zmiena wartości 0 i 1 zmiennej Tutoring lepiej rozdzielają wartości zmiennej GPA, co potwierdza jej silniejszy wpływ na średnią ocen.

Na wykresach 5 oraz 6 przedstawiono rozkład efektów przyczynowych dla poszczególnych zmiennych.

Wykres 5. Rozkład efektów przyczynowych dla zmiennej Tutoring.



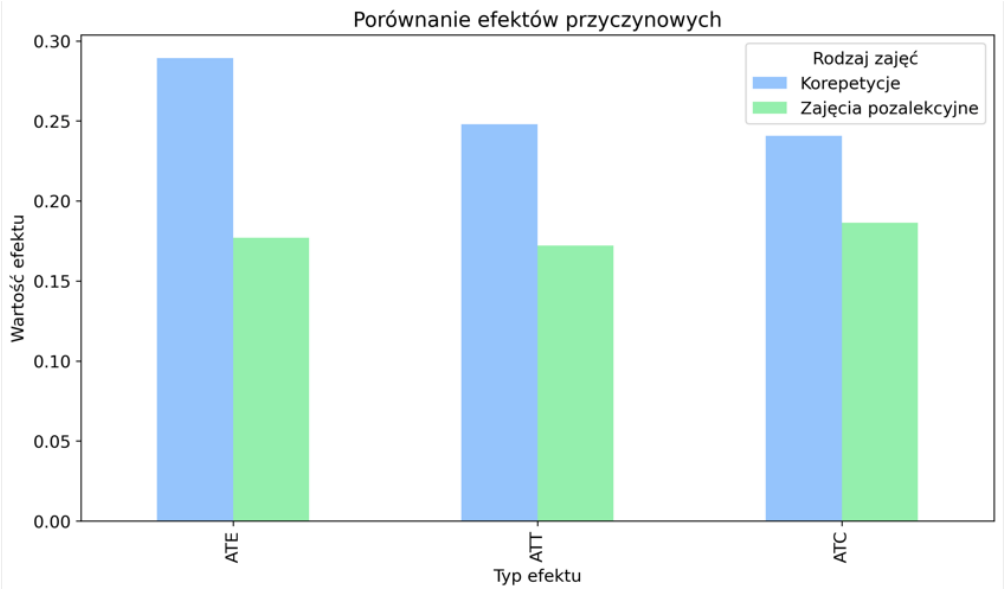
Wykres 6. Rozkład efektów przyczynowych dla zmiennej Extracurricular.



Powyższe wykresy potwierdzają, że w obydwu przypadkach efekty są dodatnie, a dla zmiennej Tutoring efekt jest silniejszy (najbardziej typowa wartość przesunięta dalej na prawo). Dodatkowo możemy powiedzieć, że efekty są dosyć homogeniczne (rozkłady dla $T = 1$ i $T = 0$ są bardzo podobne).

Na ostatnim wykresie, Wykresie 7, pokazano dodatkowo w formie graficznej różnice w efektach między zmiennymi. Jest to również potwierdzenie wniosków wyciągniętych na podstawie Tabeli 1.

Wykres 7. Porównanie efektów przyczynowych.



Dyskusja wyników

Choć analiza dostarczyła jednoznacznych wyników, należy zwrócić uwagę na kilka problematycznych kwestii:

1. Nielosowy przydział do grupy „leczenia” oraz kontrolnej – Brakuje informacji na temat sposobu zebrania danych, co podważa ich pochodzenie z kontrolowanego środowiska eksperymentalnego. Dodatkowo, brak danych o charakterystykach osób objętych świadczonymi usługami w populacji uniemożliwia pełną emulację eksperymentu. Jest to naruszenie jednego z założeń modelu Neymana-Rubina.
2. Możliwe występowanie efektów interakcji między korepetycjami a zajęciami pozalekcyjnymi – W próbie mogą znajdować się uczniowie, którzy jednocześnie korzystali

z korepetycji i uczestniczyli w zajęciach pozalekcyjnych. Taki układ sprawia, że nie było możliwe wyizolowanie efektu działania tylko jednej z tych zmiennych.

3. Brak informacji na temat typu zajęć pozalekcyjnych/korepetycji – Należałoby zastanowić się nad skupieniem się na zajęciach realizowanych wyłącznie z jednego przedmiotu, aby uniknąć potencjalnych błędów w interpretacji wyników analizy.

Jeśli na podstawie przeprowadzonej analizy należałoby udzielić konkretnych rekomendacji, to biorąc pod uwagę wyniki analizy (wskazujące na znacznie bardziej pozytywny wpływ korepetycji niż zajęć pozalekcyjnych), zaleca się skoncentrowanie większej uwagi na programie zajęć pozalekcyjnych, aby uczynić je bardziej dostępnymi i efektywnymi w kontekście podnoszenia wiedzy i kompetencji uczniów. Ponadto, wskazane jest przeprowadzenie dalszych badań nad mechanizmami wpływu obu form zajęć oraz ich kombinacji.

Spis wykresów i tabel