Analiza danych z badań międzynarodowych
za pomocą pakietu intsvy w R

Mateusz Kleczaj (m.kleczaj@ibe.edu.pl)

Spis treści

[1 Wprowadzenie 2](#_Toc200986863)

[1.1 Dlaczego warto używać pakietu intsvy? 2](#_Toc200986864)

[1.2 Ograniczenia pakietu 2](#_Toc200986865)

[2 Instalacja pakietu intsvy 3](#_Toc200986866)

[3 Pobieranie danych 3](#_Toc200986867)

[3.1 Struktura plików danych 4](#_Toc200986868)

[3.1.1 Struktura danych PISA 5](#_Toc200986869)

[3.2 Wczytywanie i łączenie danych 5](#_Toc200986870)

[3.2.1 Dane z badań IEA 6](#_Toc200986871)

[3.2.2 Dane z badań OECD 7](#_Toc200986872)

[4 Funkcje pakietu intsvy 7](#_Toc200986873)

[4.1 Funkcje analityczne pakietu: 7](#_Toc200986874)

[5 Przykładowe analizy 8](#_Toc200986875)

[5.1 Średnie 8](#_Toc200986876)

[5.2 Tabela rozkładów częstości płci 8](#_Toc200986877)

[5.3 Regresja liniowa 9](#_Toc200986878)

[5.4 Regresja logistyczna 9](#_Toc200986879)

[5.5 Percentyle 10](#_Toc200986880)

[5.6 Obliczanie odsetka uczniów na każdym z poziomów umiejętności 10](#_Toc200986881)

[6 Prezentacja graficzna 11](#_Toc200986882)

[6.1 Średnie 11](#_Toc200986883)

[6.2 Regresje 12](#_Toc200986884)

[6.3 Częstości 13](#_Toc200986885)

[7 Podsumowanie 13](#_Toc200986886)

## 1 Wprowadzenie

Pakiet intsvy w R to narzędzie stworzone m.in. dla analityków, naukowców i studentów zajmujących się międzynarodowymi badaniami edukacyjnymi: PISA, PIAAC, TIMSS, PIRLS oraz ICILS. Upraszcza ono analizę dużych zbiorów danych, automatycznie uwzględniając metodologię badań OECD i IEA, taką jak wagi statystyczne, wagi replikacyjne czy wartości prawdopodobne (*plausible values*, PVs). Dzięki intsvy można szybko obliczać statystyki, takie jak średnie, procenty, percentyle czy współczynniki regresji, oraz przedstawiać wyniki
w formie graficznej.

|  |
| --- |
|  Dodatkowe zasoby |
| * [Pełna dokumentacja](https://github.com/eldafani/intsvy)
* [Dokumentacja CRAN](https://cran.r-project.org/package%3Dintsvy)
* [Przykłady i film instruktażowy](https://daniel-caro.com/r-intsvy/)
 |

###

### 1.1 Dlaczego warto używać pakietu intsvy?

* **Automatyzacja**: upraszcza przetwarzanie złożonych danych, eliminując konieczność ręcznego uwzględniania wag i replikacji
* **Dostępność**: ma łatwo dostępne funkcje do analizy i wizualizacji danych
* **Elastyczność**: obsługuje dane z wielu badań edukacyjnych

###

### 1.2 Ograniczenia pakietu

|  |
| --- |
|  Ważne ograniczenia |
| Pakiet intsvy obsługuje tylko część międzynarodowych badań edukacyjnych. Ponadto nie wspiera modeli wielopoziomowych (np. HLM) ani zaawansowanej analizy korelacji. W celu wykonania takich analiz warto rozważyć użycie innych pakietów, takich jak:* lme4 do modeli wielopoziomowych
* psych do analizy korelacji
* EdSurvey oferuje wsparcie dla szerszego zakresu badań międzynarodowych

Niektóre funkcje pakietu intsvy jak np. select.merge działają tylko na plikacho rozszerzeniu .sav |

##

## 2 Instalacja pakietu intsvy

Jak w przypadku innych pakietów R, intsvy instaluje się i ładuje do środowiska R za pomocą komend:

install.packages("intsvy")
library(intsvy)

W razie potrzeby zainstaluj dodatkowe pakiety do wczytywania danych, np. haven dla obsługi plików .sav (SPSS):

install.packages("haven")
library(haven)

##

## 3 Pobieranie danych

Dane i dokumentacja (np. opisy nazw zbiorów i zmiennych) z badań dostępne są na poniższych stronach:

* **PISA**: [www.oecd.org/pisa/data](https://www.oecd.org/pisa/data)
* **TIMSS**: [www.iea.nl/data-tools/repository/timss](https://www.iea.nl/data-tools/repository/timss)
* **PIRLS**: [www.iea.nl/data-tools/repository/pirls](https://www.iea.nl/data-tools/repository/pirls)
* **ICILS**: [www.iea.nl/data-tools/repository/icils](https://www.iea.nl/data-tools/repository/icils)
* **PIAAC**: [www.oecd.org/skills/piaac/data](https://www.oecd.org/skills/piaac/data)

|  |
| --- |
|  Wskazówka |
| Aby korzystać z wszystkich funkcjonalności pakietu intsvy, najlepiej pobrać danew formacie .sav (SPSS) ze stron odpowiednich badań. |

### 3.1 Struktura plików danych

Przed rozpoczęciem analizy dane z badań międzynarodowych muszą zostać zaimportowane do środowiska R, co wymaga zrozumienia ich złożonej struktury plików.

#### 3.1.1 Struktura danych IEA

Zbiory IEA (TIMSS, PIRLS i ICILS) są zazwyczaj podzielone na dużą liczbę plików, pogrupowanych według kraju, poziomu klasy oraz zastosowanego narzędzia badawczego. Po ściągnięciu danych z badania TIMSS 2023 dla klasy 4 w folderze zobaczymy ponad 500 plików.

|  |
| --- |
|  Przykład nazewnictwa IEA |
| **Przykład**: Plik asapolm8.sav zawiera dane uczniów z Polski z badania TIMSS 2023 dla klasy 4 w formacie .savGdzie:- asa: odpowiedzi na zadania i wyniki uczniów klasy 4- pol: kod kraju (Polska)- m8: cykl badania (TIMSS 2023)**Pierwsza litera w nazwie pliku wskazuje poziom klasy:**- a – klasa 4- b – klasa 8**Dalsze litery określają typ danych:**- asa/bsa – wyniki uczniów oraz wartości prawdopodobne (PV)- asp/bsp – dane procesowe (np. czasy odpowiedzi)- ash – dane z kwestionariusza rodzica- asg/bsg – dane z kwestionariusza ucznia- acg/bcg – dane z kwestionariusza szkoły- atg/btg – dane z kwestionariusza nauczyciela |

#### 3.1.2 Struktura danych PISA

W przypadku badania PISA dane z różnych krajów są połączone w jeden zbiorczy plik (często bardzo duży), bez podziału na osobne pliki krajowe. Pliki są podzielone według cyklu badania oraz typu danych.

|  |
| --- |
|  Przykład nazewnictwa PISA |
| **Przykład**: Plik CY08MSP\_STU\_QQQ.sav zawiera dane z kwestionariuszy uczniówz wszystkich krajów z badania realizowanego w roku 2022.Gdzie:- CY08 – cykl badania (tu: 2022)- MSP – Main Study (badanie główne)- STU\_QQQ – dane uczniów (student)**Inne oznaczenia:**- SCH\_QQQ – kwestionariusze szkół- TCH\_QQQ – kwestionariusze nauczycieli- STU\_COG – wyniki testów kognitywnych uczniów (czytanie, matematyka itp.)- STU\_FLT – wyniki testów z edukacji finansowej- STU\_ICT – kwestionariusz dotyczący technologii informacyjno-komunikacyjnych- STU\_WBQ – kwestionariusz dobrostanu uczniów |

###

### 3.2 Wczytywanie i łączenie danych

Pakiet intsvy pozwala na szybkie wczytanie i połączenie danych z różnych krajów i źródeł za pomocą funkcji \*.select.merge(), gdzie \* to prefiks badania (timss4g, timssg8, pisa, pirls lub intsvy dla ICILS).

|  |
| --- |
|  Ważne |
| Funkcja ta obsługuje tylko pliki w formacie .sav (SPSS). Pakiet wymaga wskazania folderu, w którym znajdują się dane oraz zdefiniowania zmiennych, które chcemy połączyć. Wagi replikacyjne i końcowe oraz wartości prawdopodobne (*plausible values*, PV) są dodawane automatycznie. Dla badań IEA tworzona jest także zmienna IDCNTRYL z pełną nazwą kraju. |

####

#### 3.2.1 Dane z badań IEA

timss23 <- timssg4.select.merge(
 folder = "C:/ILSA/TIMSS/T23\_Data\_SPSS\_G4/SPSS Data",
 countries = c("AUS", "BHR", "ARM", "POL"),
 student = c("ITSEX", "ASDAGE"),
 home = c("ASBH01A"),
 school = c("ACBGDAS")
)

W powyższym przykładzie łączymy dane z badania TIMSS 2023. Upewnij się, że ścieżka do folderu z danymi (folder) prowadzi do miejsca, w którym zostały zapisane dane pobrane ze strony IEA. Należy pamiętać, że w R ukośnik odwrotny \ (stosowany w Windows) należy zastąpić ukośnikiem prostym /.

Argumenty countries, student, home i school pozwalają na wybór konkretnych krajów
i zmiennych według rodzaju narzędzia badawczego. W przykładzie wybrano Australię, Bahrajn, Armenię, Polskę oraz zmienne z kwestionariusza ucznia, rodzica i szkoły.

Obiekt timss23 zawiera wybrane dane z badania TIMSS 2023 wskazane w argumentach student, home i school oraz domyślnie dodawane zmienne, np. wartości prawdopodobne.

|  |
| --- |
|  Wskazówka |
| Funkcja \*.var.label może być użyta przed importem danych w celu przejrzenia plików źródłowych i podjęcia decyzji, które kraje i zmienne wybrać. |

timssg4.var.label(folder = "C:/ILSA/TIMSS/T23\_Data\_SPSS\_G4/SPSS Data")

#### 3.2.2 Dane z badań OECD

Sposób wczytywania danych z badania PISA jest podobny, ale wymaga dodatkowo podania nazw plików:

pisa22 <- pisa.select.merge(
 folder = "C:/ILSA/PISA",
 school.file = "CY08MSP\_SCH\_QQQ.SAV",
 student.file = "CY08MSP\_STU\_QQQ.SAV",
 student = c("ST260Q01JA", "ST313Q09JA", "ESCS", "ST004D01T"),
 school = c("MCLSIZE", "CLSIZE", "SC001Q01TA")
)

|  |
| --- |
|  Adnotacja |
| Dane PIAAC są dostępne w osobnych plikach dla każdego kraju (np. PRGPOLPUF.sav dla Polski) i wymagają łączenia w przypadku analiz międzynarodowych.Można to zrealizować za pomocą dowolnej funkcji w R, np. rbind. |

##

## 4 Funkcje pakietu intsvy

Funkcje analityczne pakietu mają postać \*.funkcja(), gdzie \* to prefiks odpowiadający badaniu (intsvy dla ICCS, pisa, timss, pirls, piaac).

### 4.1 Funkcje analityczne pakietu:

|  |
| --- |
|  Adnotacja |
| Funkcje zwracają obiekty klas (np. intsvy.mean, intsvy.reg, intsvy.table), które można wizualizować za pomocą plot(). |

##

## 5 Przykładowe analizy

Poniższe przykłady pokazano na podstawie wcześniej utworzonych danych z badania TIMSS (timss23) i PISA (pisa22).

### 5.1 Średnie

# Średni wynik z matematyki według kraju i płci w badaniu TIMSS
timss.mean.pv(
 pvlabel = paste0("ASMMAT0", 1:5),
 by = c("IDCNTRYL", "ITSEX"),
 data = timss23
)

# Średni wynik z matematyki według kraju i płci w badaniu PISA
pisa.mean.pv(
 pvlabel = paste0("PV", 1:10, "MATH"),
 by = c("CNT", "ST004D01T"),
 data = pisa22
)

###

### 5.2 Tabela rozkładów częstości płci

# Rozkład częstości płci uczniów w poszczególnych krajach (TIMSS)
timss.table(variable = "ITSEX", by = "IDCNTRYL", data = timss23)

# Rozkład częstości płci uczniów w poszczególnych krajach (PISA)
pisa.table(variable = "ST004D01T", by = "CNT", data = pisa22)

### 5.3 Regresja liniowa

# Wpływ płci ucznia i wczesnych aktywności związanych z liczeniem (ASBH01K)
# na wyniki z matematyki w TIMSS
timss.reg.pv(
 pvlabel = paste0("ASMMAT0", 1:5),
 by = "IDCNTRYL",
 x = c("ITSEX", "ASBH01K"),
 data = timss23
)

# Wpływ statusu społeczno-ekonomicznego (ESCS) na wyniki z matematyki w PISA
pisa.reg.pv(
 pvlabel = paste0("PV", 1:10, "MATH"),
 by = "CNT",
 x = c("ESCS"),
 data = pisa22
)

###

### 5.4 Regresja logistyczna

# Prawdopodobieństwo osiągnięcia ≥550 pkt z matematyki w TIMSS na podstawie
# płci i postawywobec matematyki
timss.log.pv(
 pvlabel = paste0("ASMMAT0", 1:5),
 cutoff = 550,
 x = c("ITSEX", "ASBGSLM"),
 by = "IDCNTRYL",
 data = timss23
)

# Prawdopodobieństwo posiadania komputera w domu w zależności od statusu
# społeczno-ekonomicznego (ESCS) w PISA
pisa.log(y = "ST250Q02JA", x = "ESCS", by = "CNT", data = pisa22)

### 5.5 Percentyle

# Percentyle osiągnięć matematycznych w TIMSS (np. 5, 25, 50, 75, 95)
timss.per.pv(
 pvlabel = paste0("ASMMAT0", 1:5),
 per = c(5, 25, 50, 75, 95),
 by = "IDCNTRYL",
 data = timss23
)

# Percentyle osiągnięć matematycznych w PISA (np. 10, 25, 75, 90)
pisa.per.pv(
 pvlabel = paste0("PV", 1:10, "MATH"),
 per = c(10, 25, 75, 90),
 by = "CNT",
 data = pisa22
)

###

### 5.6 Obliczanie odsetka uczniów na każdym z poziomów umiejętności

# Odsetek uczniów w TIMSS osiągających wyniki równe lub wyższe
# od progów: 400, 475, 550, 625 pkt
timss.ben.pv(
 pvlabel = paste0("ASMMAT0", 1:5),
 by = "IDCNTRYL",
 cutoff = c(400, 475, 550, 625),
 data = timss23
)
# Odsetek uczniów w PISA osiągających wyniki równe lub wyższe
# od progów: poziomy 1–6 (355–698 pkt)
pisa.ben.pv(
 pvlabel = paste0("PV", 1:10, "MATH"),
 by = "CNT",
 cutoff = c(355, 407, 480, 553, 626, 698),
 data = pisa22
)

## 6 Prezentacja graficzna

Wyniki analiz przeprowadzonych za pomocą funkcji pakietu intsvy (średnich, regresji, rozkładów częstości) można wizualizować za pomocą dedykowanych funkcji plot().

|  |
| --- |
|  Dodatkowe opcje eksportu |
| Dodatkowo wyniki można:- Wyświetlić w konsoli za pomocą funkcji summary() z pakietu podstawowego R- Eksportować do pliku .csv, używając opcji export = TRUE w funkcjach analitycznychFunkcja na.omit() stosowana w przykładach usuwa braki danych z wyników analizy przed ich wizualizacją, co zapobiega błędom w generowaniu wykresów. |

###

### 6.1 Średnie

# Wykres punktowy z przedziałami ufności pokazujący średnie wyniki z matematyki
# w każdym kraju podziale na płeć (TIMSS)
plot.intsvy.mean(
 na.omit(
 timss.mean.pv(
 pvlabel = paste0("ASMMAT0", 1:5),
 by = c("IDCNTRYL", "ITSEX"),
 data = timss23)))



### 6.2 Regresje

# Wykres pokazujący związek statusu społeczno-ekonomicznego (ESCS)
# z osiągnięciami uczniów w matematyce w podziale na kraje (PISA)
plot.intsvy.reg(
 na.omit(
 pisa.reg.pv(
 pvlabel = paste0("PV", 1:10, "MATH"),
 x = "ESCS",
 by = "CNT",
 data = pisa22
 )
 )
)



### 6.3 Częstości

# Skumulowany wykres słupkowy pokazujący rozkład częstości płci uczniów
# w każdym z krajów (TIMSS)

plot.intsvy.table(
 na.omit(
 timss.table(
 variable = "ITSEX",
 by = "IDCNTRYL",
 data = timss23
 )
 ),
 stacked = TRUE
)



## 7 Podsumowanie

Pakiet intsvy automatyzuje analizy danych edukacyjnych zawierających wartości prawdopodobne (PVs) oraz wagi replikacyjne, automatycznie dobierając schemat ważenia i obsługując specyfikę danych z międzynarodowych badań edukacyjnych.

|  |
| --- |
|  Zalecenia końcowe |
| 1. **Dokumentacja**: Kluczowe jest dokładne zapoznanie się z dokumentacją konkretnego badania, aby prawidłowo określić strukturę danych, nazwy zmiennych oraz odpowiednie funkcje pakietu intsvy
2. **Weryfikacja**: Zaleca się weryfikację wyników analiz z oficjalnymi raportami międzynarodowymi lub krajowymi, aby potwierdzić ich poprawność
 |