W03. Dane psychologiczne i ich reprezentacja

Psychoinformatyka | Wykład 03

Cele i zakres

- Rozróżnić obserwację, jednostkę badania i konstrukt
- Rozpoznać typy danych oraz formaty ich zapisu
- Skontrolować kompletność 14-dniowego EMA

Od zachowania do rekordu

- Pomiar wymaga operacjonalizacji konstruktu
- Rekord przechowuje wynik i metadane zdarzenia
- Agregacja może usuwać informację o przebiegu zdarzeń

Źródła danych psychologicznych

- Samoopis: kwestionariusz i EMA
- Zachowanie: RT, poprawność, sekwencje wyborów
- Sygnały fizjologiczne i ślady aktywności cyfrowej

Ilościowe, jakościowe, mieszane

- Ilościowe: czas reakcji w milisekundach
- Jakościowe: wypowiedź uczestnika
- Mieszane: skala stresu i komentarz tekstowy

Skale pomiarowe

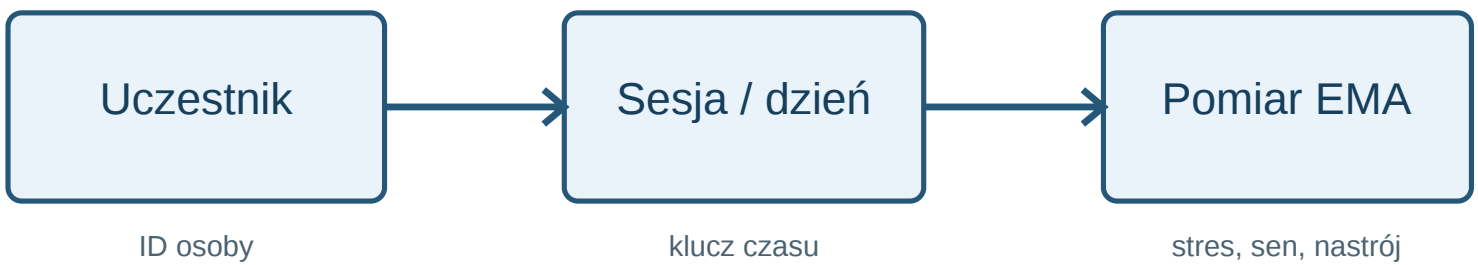
- Nominalna: etykieta grupy; porządkowa: odpowiedź Likerta
- Przedziałowa: równe różnice bez zera absolutnego
- Ilorazowa: czas w sekundach lub ms

Typ danych a semantyka

- Identyfikator 0007 zachowujemy jako tekst
- Zero kroków jest wynikiem, pusty rekord jest brakiem
- Liczba zakodowana jako int nie staje się automatycznie skalą ilorazową

Jednostka obserwacji

- Uczestnik, sesja, próba i zdarzenie to różne poziomy
- 504 obserwacje 36 osób nie oznaczają próby N=504
- Powtarzane pomiary wymagają modelu zależności



Rysunek 1: Zagnieżdżenie danych psychologicznych

Projekt przekrojowy i podłużny

- Przekrój: pomiar w jednym punkcie czasu
- Dane podłużne: obserwacje wielu dni tych samych osób
- Trend populacyjny może różnić się od trendów indywidualnych

Szereg czasowy

- Czasy pomiarów EMA często są nieregularne
- Znacznik zdarzenia ≠ czas zapisu na serwerze
- Próbkowanie i agregacja determinują rozdzielczość analizy

Kwestionariusze jako macierze

- Wiersz uczestnika, kolumny P1–Pk
- Odwrócenie punktacji przed agregacją
- Kontrola zakresów i braków poprzedza ocenę rzetelności

Dane behawioralne

- Wiersz oznacza próbę eksperymentalną
- Kolumny: warunek, odpowiedź, poprawność i RT [ms]
- Brak reakcji nie jest czasem reakcji równym zeru

EMA w codziennym środowisku

- 36 uczestników × 14 dni = 504 planowane pomiary
- Stres 0–10 jest dziennym samoopisem
- Brak odpowiedzi nie może oznaczać braku stresu

```
osoba,dzien,sen_h,stres
01,1,7.2,4
01,2,6.5,6
02,1,8.0,
```

ECG, EEG, EDA i eye tracking

- Sygnał jest szeregiem próbek z jednostką i zegarem
- Szum i artefakty wymagają odrębnej kontroli jakości
- Porównanie urządzeń wymaga synchronizacji i kalibracji

Wide i long

- Long: uczestnik, dzień, zmienna, wartość
- Wide: kolumny stres_d01, stres_d02 itd.
- Transformacja pivot wymaga jednoznacznego klucza

osoba	dzien	stres	nastroj
01	1	4	7
01	2	6	5

Identyfikatory i relacje

- Klucz EMA: osoba + dzień + ewentualny epizod
- Duplikat klucza może oznaczać błąd lub korektę
- Pseudonimowe ID pozostaje daną potencjalnie identyfikującą

5 minut przerwy

Pytania do samodzielnego przemyślenia

1. Czym różnią się dane podłużne od przekrojowych?
2. Dlaczego pojedyncza pozycja Likerta ma charakter porządkowy?
3. Dlaczego 504 wiersze od 36 osób nie oznaczają 504 niezależnych osób?
4. Jak identyfikować pojedynczy pomiar EMA?

Słownik danych

- Nazwa, jednostka, typ, zakres, kodowanie braków
- Wersja procedury i informacja o aparaturze
- Definicja wyniku jest częścią protokołu badania

Kontrola wartości i jednostek

- Sprawdź sen_h w (0,24], stres w [0,10]
- Sekundy i milisekundy różnią się czynnikiem 1000
- Reguły odrzucania zapisuj jawnie, bez cichego poprawiania danych

Czas UTC i data lokalna

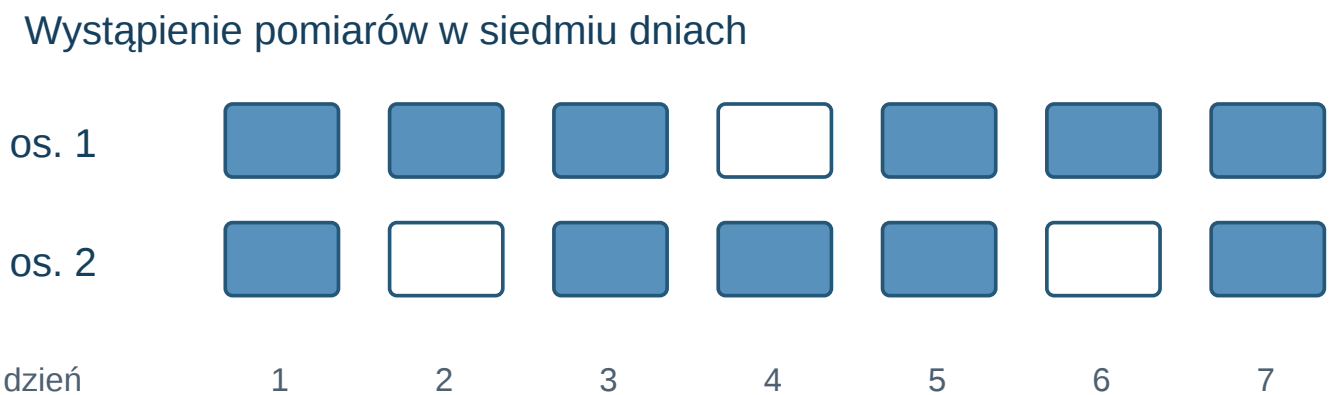
- Przechowuj timestamp UTC oraz strefę uczestnika
- Zmiana czasu może powodować powtórzone godziny lokalne
- Doba badawcza wymaga jednoznacznej definicji

CSV, JSON i Parquet

- CSV: płaska tabela, schemat w osobnym słowniku
- JSON: struktury hierarchiczne i logi zdarzeń
- Parquet: typowane dane kolumnowe dla większych zbiorów

MCAR, MAR, MNAR

- MCAR: brak niezależny od danych
- MAR: warunkowo niezależny od brakującej wartości przy danych obserwowanych
- MNAR: zależność pozostaje nawet po uwzględnieniu obserwowanych danych



Rysunek 2: Przykład braku pomiarów EMA

Wartości odstające i artefakty

- Skrajny wynik może odzwierciedlać prawdziwe zachowanie
- Błąd eksportu i zanik transmisji mają techniczny mechanizm
- Automatyczne usuwanie wszystkich wartości skrajnych jest ryzykowne

Minimalizacja danych

- Ogranicz precyzję lokalizacji do potrzeb badania
- Oddziel identyfikatory od danych analitycznych
- Oceń ryzyko reidentyfikacji na podstawie czasu i wzorców aktywności

Python: generowanie EMA

- Notebook tworzy 36 × 14 dni, wyłącznie dane syntetyczne
- Ziarno RNG = 42 umożliwia powtórzenie
- Wyniki nie pochodzą z rzeczywistego badania

```
import pandas as pd
dane = pd.read_csv('pi_w03_data.csv')
assert len(dane) == 504
```

Python: walidacja schematu

- Sprawdź kolumny osoba, dzien, sen_h, stres, nastroj
- Weryfikuj unikalność (osoba,dzien)
- Sprawdź zakresy wartości bez zastępowania braków zerami

```
assert not dane.duplicated(['osoba','dzien']).any()
assert dane.stres.dropna().between(0,10).all()
```

Python: braki i reshaping

- Oblicz odsetek brakujących stresów dla każdej osoby
- Przekształć stres i nastrój poleceniem melt
- Zweryfikuj, że transformacja zachowuje klucze i braki

```
long = dane.melt(id_vars=['osoba','dzien'],
                 value_vars=['stres','nastroj'],
                 var_name='zmienna', value_name='wynik')
```

Python: wykres dynamiki

- Przedstaw stres jednej osoby na osi czasu
- Zaznacz brakujące pomiary bez nieuzasadnionej interpolacji
- Podpisz osie: dzień i stres 0–10

```
import matplotlib.pyplot as plt
osoba = dane.query('osoba == 1')
plt.plot(osoba.dzien, osoba.stres, 'o-')
plt.xlabel('Dzień'); plt.ylabel('Stres [0–10]')
plt.show()
```

Błąd agregacji

- Zależność wewnątrz osoby nie musi odpowiadać zależności między osobami
- Rekordy jednej osoby są zależne statystycznie
- Do wnioskowania mogą być potrzebne modele mieszane

Reprodukowalność i wnioski

- Zapisz schemat, jednostki, wersję danych i reguły filtracji
- Zachowaj plik surowy odrębnie od danych analitycznych
- Wniosek powinien odpowiadać poziomowi zastosowanej analizy

Bibliografia i narzędzia

- Shiffman, Stone, Hufford (2008), Ecological Momentary Assessment, Annu. Rev. Clin. Psychol. 4:1–32
- Little & Rubin (2019), Statistical Analysis with Missing Data, 3rd ed.
- pandas.pydata.org/docs; quarto.org; jupyter.org

Wnioski i samodzielna praca

- Jednostką analizy w EMA jest osoba i kolejne pomiary w czasie; 504 wiersze nie oznaczają 504 niezależnych uczestników.
- Analizę braków danych i odstępów pomiarowych trzeba wykonać przed modelowaniem.
- Raportuj schemat danych, jednostki i reguły filtracji; nie zastępuj braków zerem.
- Przećwicz kontrolę danych w `pi_w03.ipynb`.