

Wartości zaokrąglone i powtarzające się w empirycznym rozkładzie wynagrodzeń obserwowanych w Badaniu Aktywności Ekonomicznej Ludności (BAEL)

Konsekwencje dla imputacji braków odpowiedzi
oraz innych aspektów przetwarzania i analizy danych

Tomasz Piasecki

Urząd Statystyczny w Łodzi
Ośrodek Statystyki Matematycznej



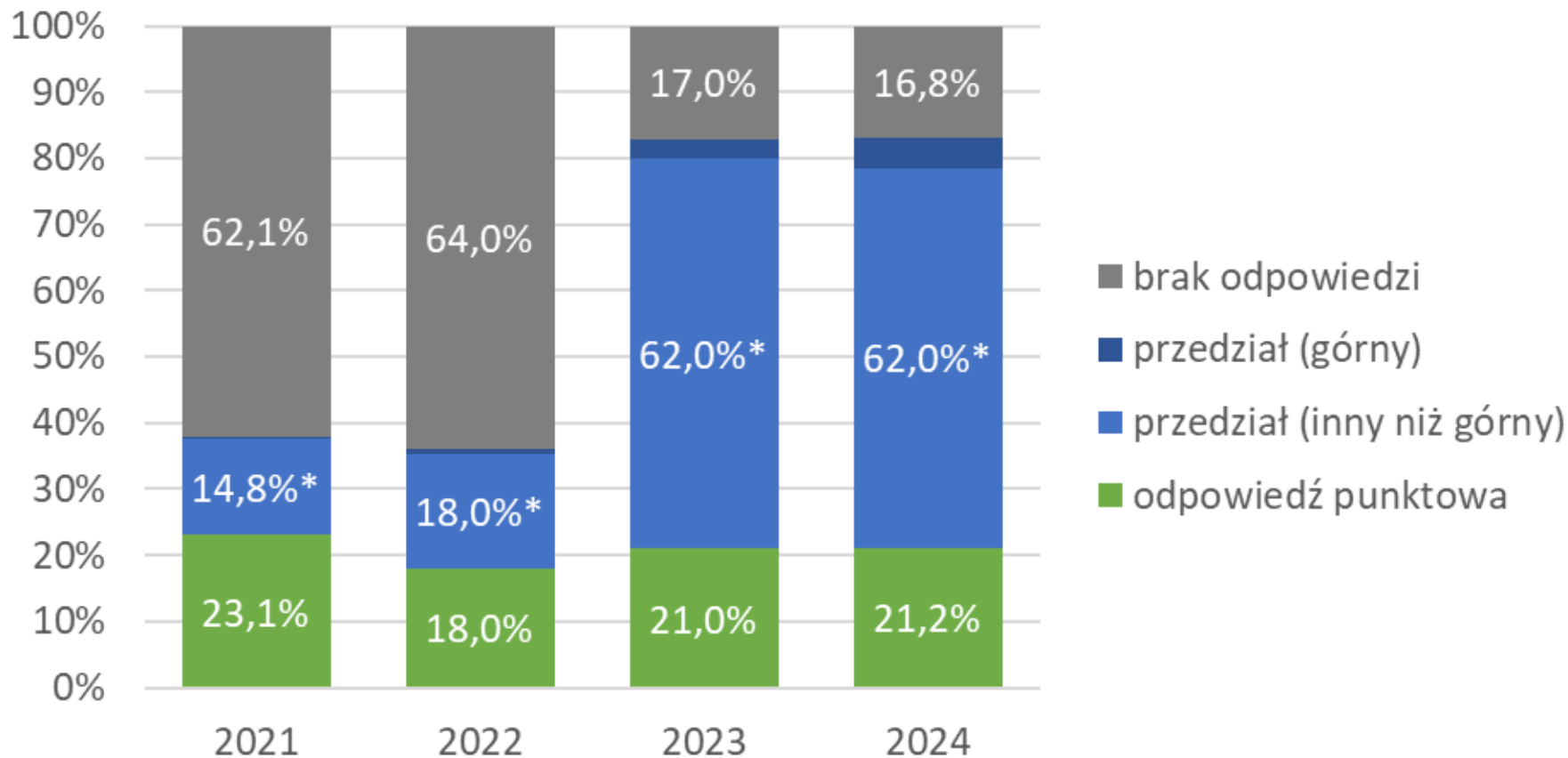
- Zagadnienia omawiane w prezentacji powiązane są z analizami wykonanymi przy realizacji projektu pn. **Obszar statystyki pracy - Infrastruktura statystyczna LFS w ramach zintegrowanej europejskiej statystyki społecznej (IESS) – Moduł 2023; Imputacja zmiennej „miesięczne wynagrodzenie z pracy głównej”** współfinansowanego ze środków Unii Europejskiej na podstawie umowy o udzielenie dotacji nr **101101153 – 2022-PL-LFS-LMB** w ramach programu *Single Market Programme* (SMP ESS). Projekt realizowany w latach 2022-2024 przez **Departament Rynku pracy GUS** z udziałem **Urzędu Statystycznego w Poznaniu** oraz **Urzędu Statystycznego w Łodzi**
- **Metodyka imputacji** danych brakujących dotyczących wynagrodzenia z pracy głównej w BAEL została opracowana w ramach ww. projekt przez jednostki go realizujące z wiodącym udziałem zespołu **Ośrodka Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu**

Problem w skrócie

- Informacja o **wynagrodzeniu miesięcznym z pracy głównej** jest pozyskiwana w BAEL **od respondenta w wywiadzie** jako wartość netto i konwertowana na brutto
- Wartość netto **może być sprawozdana** przez respondenta jako:
 - **wartość punktowa** (konkretna kwota)
 - **przedział** (wskazywany przez respondenta z listy)
 - **nie sprawozdana** (brak odpowiedzi)
- W przypadku braków odpowiedzi oraz odpowiedzi przedziałowych wartości punktowe są uzyskiwane przy użyciu imputacji
- **W rozkładzie odpowiedzi punktowych z wywiadu występuje wiele wartości okrągłych** (zaokrąglonych), co m.in.
 - **zaburza proces imputacji danych brakujących**, co w szczególności dotyczy imputacji wartości punktowej dla danego przedziału
 - powoduje problemy z analizą rozkładu zmiennej, dotyczące np. określania kwantyli (w tym decyli) i podziału zbiorowości na grupy kwantylowe
- Rozwiązania wdrożone w związku z imputacją skutecznie znoszą rozpoznane zagrożenia dotyczące tego etapu przetwarzania. Ich szersze zastosowanie mogłoby przynieść dalej idące korzyści, jednak wiąże się z zagadnieniami dotyczącymi zintegrowanej metodologii europejskiego badania siły roboczej (EU-LFS)

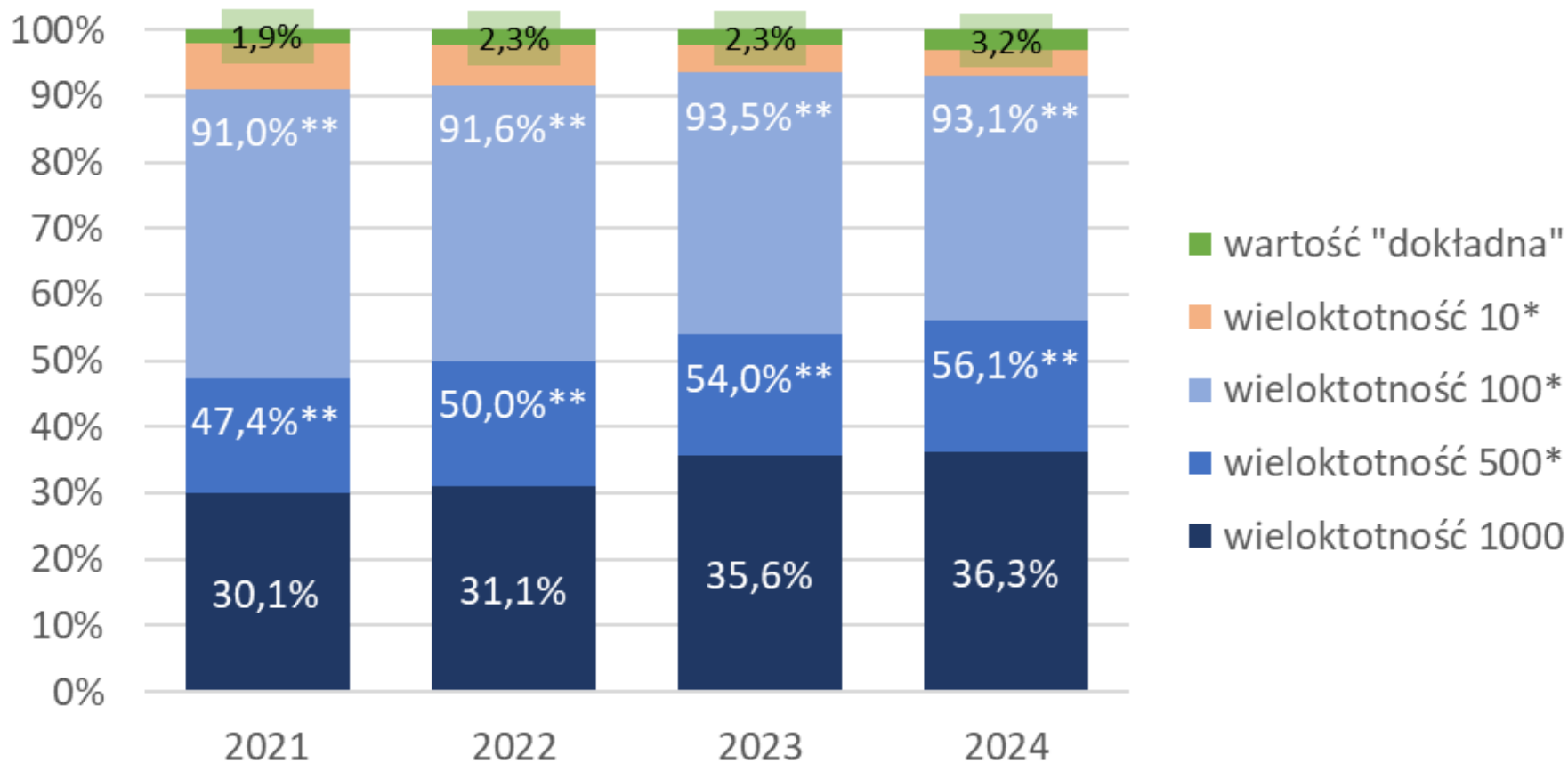
Pozycyjne wskaźniki odpowiedzi

dotyczące wynagrodzenia z pracy głównej (wg roku i typu odpowiedzi)



* wszystkie odpowiedzi przedziałowe (również górny przedział)

Wartości okrągłe (wg poziomu zaokrąglenia) w rozkładzie odpowiedzi punktowych dot. wynagrodzenia z pracy głównej



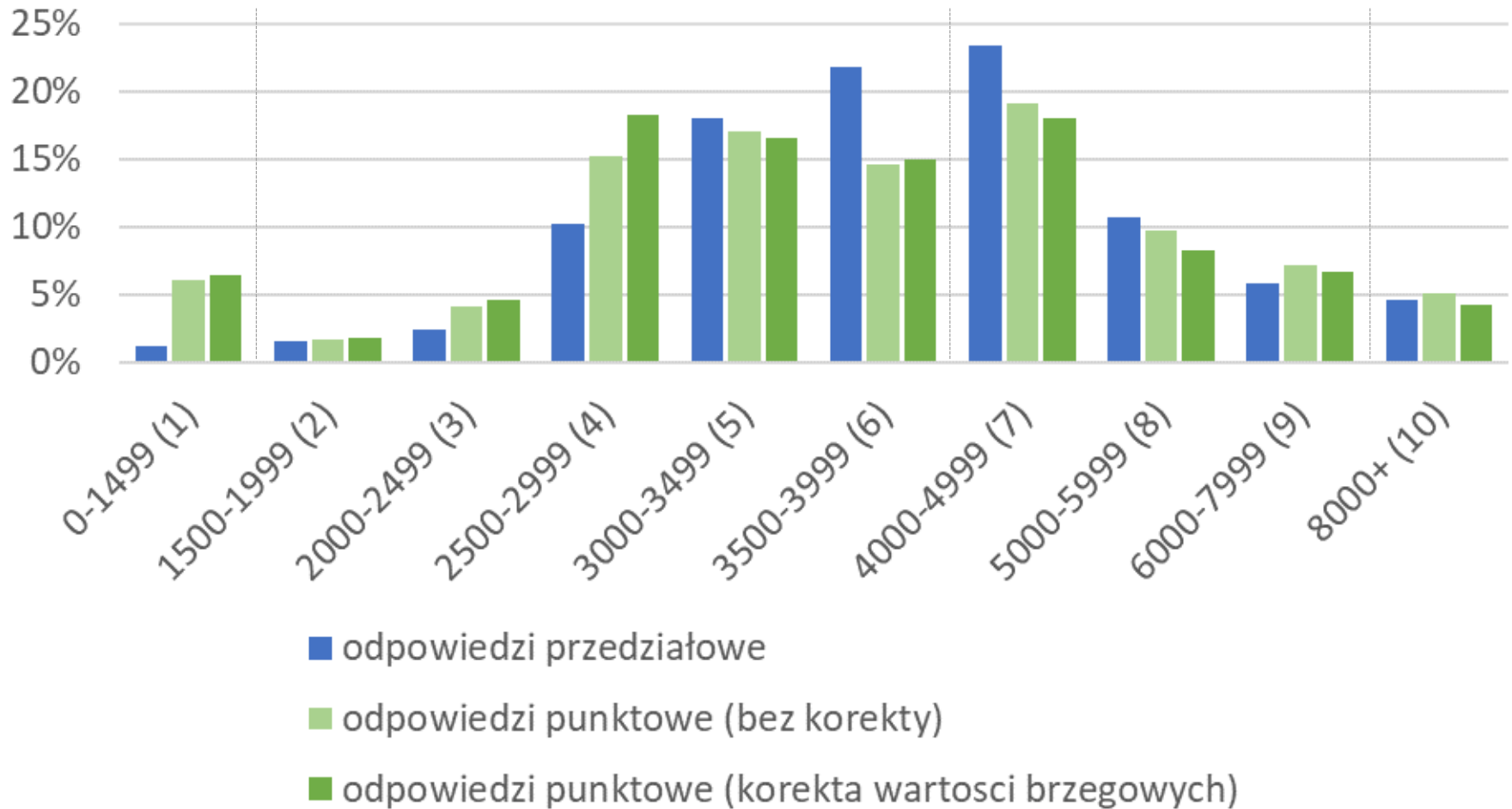
* i nie jest wielokrotnością wyższej wartości okrągłej

** łącznie z wielokrotnościami wyższych wartości okrągłych

Imputacja danych brakujących

- **Stan docelowy:** wymagane **wartości punktowe dla wszystkich obserwacji**
- Wartość punktowa jest imputowana dla wszystkich braków odpowiedzi oraz odpowiedzi przedziałowych
- Używana **metoda PMM** (*predictive mean matching*)
 - mimo wykorzystania modelu predykcyjnego, finalnie metoda **oparta na dawcach**, która przenosi wartość z innej obserwacji – **wartości imputacyjne pochodzą z rozkładu empirycznego wartości z wywiadu**
 - zachowuje specyficzne cechy rozkładu empirycznego (w tym zaokrąglenia)
- Imputacja **w 2 etapach** (metoda PMM używana w obydwu)
 - **imputacja przedziału** (tylko dla braków odpowiedzi)
 - **imputacja wartości punktowej z danego przedziału** (określonego przez respondenta dla odpowiedzi interwałowych, imputowanego w 1. etapu dla braków odpowiedzi)
- Procedura dwuetapowa (przedział -> wartość punktowa) stosowana dla braków odpowiedzi w celu minimalizacji potencjalnych obciążeń – rozkład odpowiedzi przedziałowych wydaje się bardziej wiarygodny, a przy tym różni się od rozkładu odpowiedzi punktowych

Porównanie rozkładu odpowiedzi przedziałowych i punktowych (2023)



Zaokrąglenia a imputacja

- Problem dotyczy **imputacji wartości punktowych w zadanym przedziale**
- Przy zbieraniu informacji przedziałowych z wywiadu stosowana jest lista **przedziałów lewostronnie domkniętych**; okrągłe kwoty stanowią wartości graniczne (np. 5 000 – 5 999):
 - **przedział zawiera okrągłą wartość na lewym krańcu**
 - **nie ma takiej wartości należącej do przedziału na prawym krańcu**
- W przypadku występowania dużej liczby wartości okrągłych (zaokrąglonych) **rozkład empiryczny wartości punktowych z wywiadu należących do każdego z przedziałów jest silnie asymetryczny** – mamy dużą nadreprezentację wartości na lewym krańcu. Bezpośrednie użycie takiego rozkładu skutkowałoby prawie zawsze **imputacją wartości granicznej z lewego krańca przedziału**
- Tymczasem: można założyć, że rzeczywiste wynagrodzenia respondentów deklarujących przedział są rozłożone względnie równomiernie
- Skutkowałoby to **systematycznym niedoszacowaniem imputowanych wartości punktowych w każdym przedziale**
- Potencjalne **obciążenie jest ściśle uzależnione od sposobu zdefiniowania przedziałów**: przedziały prawostronnie domknięte skutkowałyby systematycznym przeszacowaniem wartości punktowych

Problem zaokrągleń – imputacja wartości punktowych w przedziale

Przedziały lewostronnie domknięte, np. 5 000 – 5 999

- Rzeczywiste wynagrodzenia raportujących przedział:



- Rozkład wartości punktowych z wywiadu należących do przedziału:



Przedziały prawostronnie domknięte, np. 5 001 – 6 000

- Rzeczywiste wynagrodzenia raportujących przedział:
(w większości ci sami respondenci)



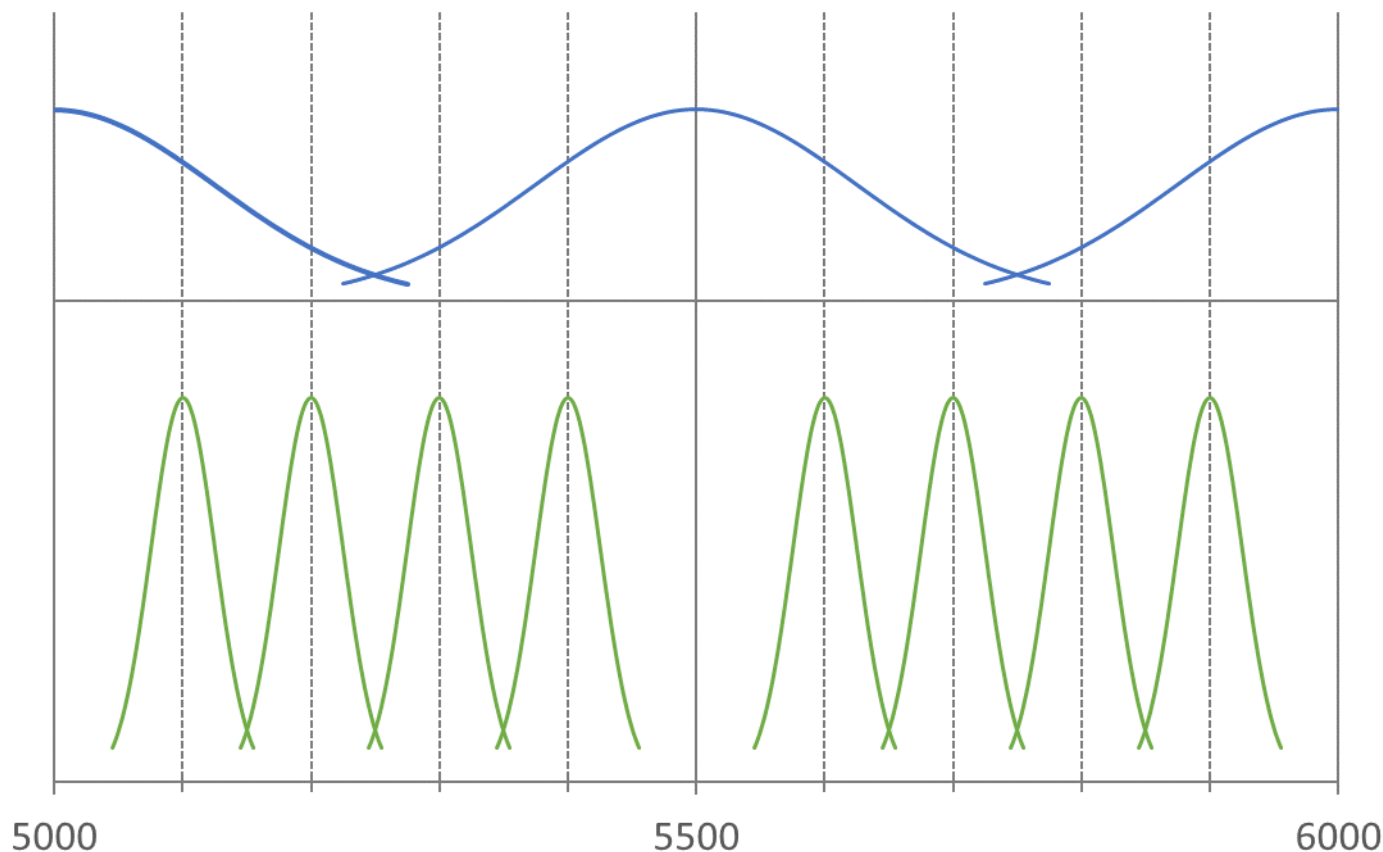
- Rozkład wartości punktowych z wywiadu należących do przedziału:



Zaokrąglenia a imputacja – rozwiązanie

- **Losowe zaburzenie** okrągłych wartości punktowych **przed imputacją** (modyfikuje zbiór „dawców”)
 - zastosowane do wszystkich wielokrotności 10 PLN
 - **dodanie losowego szumu** o średniej (wartości oczekiwanej) 0
 - odchylenie standardowe zależy od „poziomu zaokrąglenia”:
 - wielokrotności 500: 125 PLN
 - pozostałe wielokrotności 100: 25 PLN
 - pozostałe wielokrotności of 50: 12,50 PLN
 - pozostałe wielokrotności 10: 2,50 PLN
- **Zaburzenie wartości punktowych z wywiadu** (sprawozdanych przez respondentów) **nie jest zachowywane** w ostatecznych danych wynikowych – zaburzenie dotyczy **tylko** zbioru dawców danych imputacyjnych, **oddziałuje wyłącznie na wartości imputowane**

Zaburzanie wartości okrągłych – ilustracja



Zaokrąglenia a imputacja – definicja przedziałów

- Dodatkowy środek zaradczy: w wyniku doświadczeń związanych z analizą opisywanego problemu zdecydowano od 2024 r. **zmienić sposób definiowania predefiniowanych przedziałów wynagrodzenia** stosowanych w kwestionariuszu wywiadu
 - unikanie stosowania okrągłych wartości jako granic przedziałów (przesunięcie ich na wartości „mniej okrągłe”)
 - preferowanie sytuacji, gdy **wartość okrągła znajduje się w środku przedziału, nie na jego brzegu**

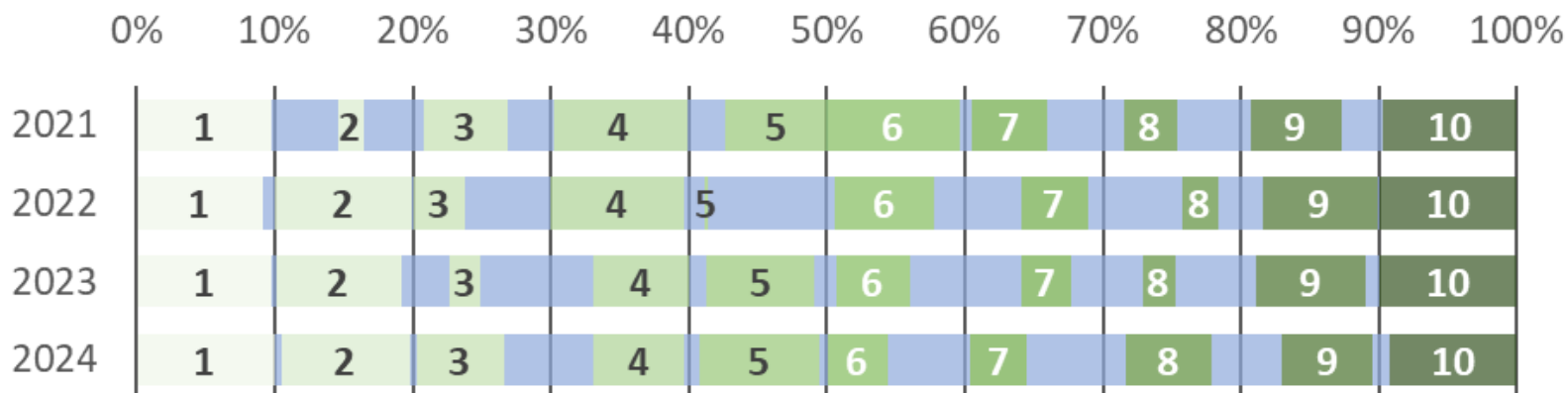
Do 2023 r.	Od 2024 r.
...	...
3 000 – 3 499	3 250 – 3 749
3 500 – 3 999	3 750 – 4 249
4 000 – 4 999	4 250 – 4 749
...	...

Zaokrąglenia a oznaczanie grup decylowych

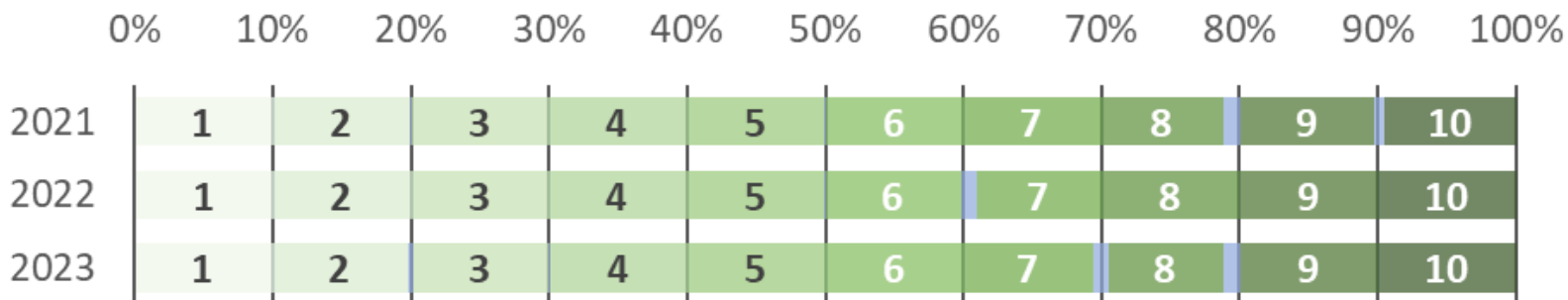
- Opracowując dane z badania EU-LFS (którego polską częścią jest BAEL) Eurostat oznacza grupy decylowe na podstawie wynagrodzenia z pracy głównej (zmienna INCDECIL)
- Zaokrąglenia zaburzają także proces oznaczania grup decylowych – są one jedną z przyczyn (najważniejszą) występowania powtarzających się wartości na granicach grup. Pojawia się problem **przypisania identycznych wartości do różnych grup decylowych**.
- Eurostat zaproponował **deterministyczny algorytm randomizacyjny**, który przypisuje te wartości w sposób „wyglądający na losowy” (na podstawie pewnej permutacji znaków wchodzących w skład identyfikatorów rekordu)
- Logika postępowania względem źródła problemu – zaokrągleń – jest podobna, jak przypadku rozwiązania dotyczącego imputacji w polskim BAEL
 - sztuczne zaburzenia pozwalają „losowo” rozdzielić powtarzające się wartości na granicy grup
 - same wartości zmiennej reprezentującej wynagrodzenie pozostają jednak niezmiennie
- Trwałe zaburzenie lub zastąpienie wartościami imputowanymi wartości okrągłych (przy potraktowaniu ich np. jako odpowiedzi przedziałowych) mogłoby rozwiązać problemy w sposób ogólny i bardziej uniwersalny

Wartości wielokrotne na granicach decyli (2023)

Odpowiedzi punktowe z wywiadu



Finalne wartości (wartości punktowe z wywiadu i imputowane)



■ wartości wielokrotne na granicach grup decylowych

Czy modyfikować trwale dane zaokrąglone?

- Potencjalna możliwość kompleksowego rozwiązania problemu, bez uciekania się do rozwiązań szczególnych dla problemów cząstkowych (+)
- *Przypisywanie różnych wartości zmiennej finalnej respondentom, którzy udzielili takiej samej odpowiedzi w wywiadzie (pozyskano dla nich tę samą informację w badaniu) może być kontrowersyjne. Różnice między nimi, które mogą mieć znaczenie analityczne, generowane są w sposób sztuczny lub losowy, nie mający podstawy w zebranych danych (-).* Faktycznie ma to jednak miejsce również w stosowanych już rozwiązaniach, jak algorytm oznaczania grup decylowych stosowany przez Eurostat
- *Szacunek do danych sprawozdanych przez respondenta (pozyskanych z takim trudem). W niektórych przypadkach wartość okrągła może być prawdziwą wartością mierzonej cechy (-).* Ale w większości przypadków tak jednak nie jest
- **Czy wartości okrągłe są naprawdę odpowiedziami punktowymi?** Czy może raczej jest to pewien rodzaj **odpowiedzi przedziałowej** (o trudniejszych do określenia granicach przedziału)? Potraktowanie ich w ten sposób dawałoby uzasadnienie dla imputacji dokładnej wartości punktowej, tak jak to robimy w przypadku jawnych odpowiedzi przedziałowych (+)

Uwagi końcowe

- Występowanie zaokrągleń może znacząco zniekształcać rozkład. Dane pochodzące z rozkładu ciągłego przybierają *de facto* postać dyskretną.
- Zaokrąglenia mogą być źródłem problemów wymagających szczególnej uwagi i zagrożeń, tak na etapie analizy danych (np. wyznaczenie kwantyli, oznaczenie grup kwantylowych) jak i ich przetwarzania (omawiany przykład dot. imputacji)
- W przypadku wynagrodzeń badanych w BAEL i EU-LFS trwała transformacja danych zaokrąglonych do postaci ciągłej mogłaby być kompleksowym rozwiązaniem problemu. Wobec braku decyzji o trwałym modyfikowaniu danych z wywiadu faktycznie stosowane rozwiązania problemów cząstkowych mają charakter bardziej doraźny. Jednym z powodów jest brak wytycznych dających zgodę na taki rozwiązanie w zharmonizowanej metodologii EU-LFS, jednak decyzja o losowej modyfikacji danych sprawozdanych przez respondentów sama w sobie nie jest prosta

Dziękuję za uwagę

Tomasz Piasecki

Urząd Statystyczny w Łodzi

Ośrodek Statystyki Matematycznej

t.piasecki@stat.gov.pl