Codebook FAQ

Spis treści

1 Co to jest codebook?

Informacje zebrane w badaniach (czy to w ankietach kwestionariuszowych, czy testach z zadaniami wypełnianymi przez uczniów) umieszczane są w zbiorach danych. Są to tabele, w których wiersze odpowiadają kolejnym obserwacjom (np. poszczególnym respondentom albo uczniom), kolumny poszczególnym pytaniom/zadaniom z kwestionariusza/testu, a wartości komórek zwierają **kody** odpowiedzi udzielonych przez danego respondenta/ucznia na dane pytanie/zadanie. Kody te są najczęściej **liczbami symbolizującymi odpowiedzi**, stąd gdzieś powinna zostać zapisana informacja o tym, jak te kody **interpretować**. Do tego właśnie służy *codebook* (po polsku *książka kodów/książka kodowa*) – jest to ***opis sposobu, w jaki zostały zakodowane dane w zbiorze danych***.

Wśród osób, które miały do czynienia z ocenianiem testów, można się spotkać z użyciem słowa codebook w znaczeniu *opis sposobu punktowania odpowiedzi na poszczególne zadania*,

w niniejszym dokumencie chodzi jednak o znaczenie przytoczone w poprzednim akapicie.

2 Po co mi codebook, przecież sposób zapisu informacji w zbiorze danych jest oczywisty?!

a) Sposób, w jaki dane zakodowane zostały w zbiorze danych może się wydawać oczywisty w trakcie realizacji badania, jednak dla innych badaczy, którzy będą chcieli skorzystać ze zbioru danych, a po pewnym czasie także dla samych twórców badania, taki nie będzie. Stąd sposób zakodowania informacji w zbiorze danych powinien zostać udokumentowany w codebooku.

b) Zbiory danych zawierają bardzo dużo informacji, na tyle dużo, że nie sposób ręcznie szukać w nich błędów (np. obecności kodów niedopuszczalnych w danym pytaniu/zadaniu), a błędy takie zdarzają się zawsze (szczególnie, gdy informacje wprowadzane są do komputera ręcznie, np. w procesie kodowania testów wykonanych przez uczniów na papierze). Odpowiednio przygotowany codebook umożliwia automatyczną weryfikację poprawności zbioru danych.

3 Dlaczego dobrze jest, gdy wszystkie codebooki mają taką samą strukturę?

Opisać, w jaki sposób dane zostały zakodowane w zbiorze danych można na wiele sposobów, warto go jednak ujednolicać, gdyż:

a) Jeśli codebooki nie będą miały takiej samej struktury, nie będzie można ich wykorzystać do automatycznej weryfikacji poprawności zbiorów danych, jak również nie będzie możliwa publikacja takich zbiorów danych za pośrednictwem programu do udostępniania zbiorów danych z badań poprzez stronę www IBE.

b) Jeśli codebooki mają różną strukturę, to wyszukiwanie w nich pożądanej informacji jest bardziej kłopotliwe, gdyż za każdym razem trzeba zacząć od zapoznania się ze strukturą danego codebooka.

4 Dlaczego to ja powinienem ułożyć codebook? Czy nie może tego zrobić wykonawca badania?

Codebook powinien ułożyć ten, kto projektował narzędzia badawcze wraz z tym, kto będzie analizował wyniki badania, a to dlatego, że to oni najlepiej wiedzą (czasem wręcz tylko oni wiedzą), w jaki sposób narzędzia badawcze powinny zostać odwzorowane w *zmienne* (patrz pytanie o to, co to jest *zmienna*). Oczywiście wykonawca badania może stworzyć codebook do narzędzi badawczych, których sam nie przygotowywał, istnieje jednak wtedy ryzyko, że zrobi to w sposób, który nie będzie umożliwiał wykonania na danych wszystkich analiz, jakie chciałby na nich wykonać zamawiający badanie. O tym, że jest to zagrożenie realne, bardzo boleśnie przekonał się w jednym ze swoich pierwszych badań jeden z zespołów dydaktycznych.

Jeśli natomiast narzędzia badawcze przygotowuje wykonawca, to on odpowiedzialny będzie za przygotowanie codebooka do zbioru danych z badania, które przeprowadza. Także wtedy warto jednak zapoznać się z przygotowanym przez niego codebookiem i upewnić się, że taki sposób przedstawienia danych w zbiorze danych, jaki zaproponował wykonawca nie pozbawia nas możliwości wykonania na tych danych tych analiz, które planowaliśmy.

5 Jaka jest struktura wzorcowego codebooka IBE?

Wzorzec codebooka IBE jest tabelą, która w kolejnych wierszach opisuje kolejne *zmienne* (patrz następne pytanie) zbioru danych, a kolumny zawierają wskazane informacje o danej zmiennej (np. jej nazwę, typ skali, zakres dopuszczalnych wartości).

6 Co to jest *zmienna* w codebooku?

Zmienna jest to informacja na zadany temat. O tym, na jaki temat informację przechowuje dana *zmienna* oraz ile *zmiennych* potrzeba, aby zawrzeć w zbiorze danych wszystkie informacje gromadzone przez narzędzia badawcze decyduje osoba projektująca zbiór danych na podstawie tego, jakie analizy zamierza później na projektowanym zbiorze danych przeprowadzać. Nie ma tu żadnej uniwersalnej zasady mówiącej, jak zapisać za pomocą *zmiennej* (lub kilku *zmiennych*) dane pytanie kwestionariusza lub zadanie testu – zawsze punktem wyjścia musi być to, jakie informacje dawane przez narzędzia badawcze chcemy zapisać w zbiorze danych. W szczególności nie istnieje coś takiego, jak *wszystkie informacje* gromadzone przez dane narzędzie badawcze, ich liczbę można bowiem mnożyć niemal w nieskończoność.

Jako przykład weźmy zadanie z testu, które składa się z dwóch podpunktów, z których każdy jest pytaniem wielokrotnego wyboru z czterech dostępnych odpowiedzi, z których tylko jedna jest poprawna. Kilka z możliwych sposobów przedstawienia tego zadania w zbiorze danych i codebooku to:

- jedna zmienna informująca o tym, czy uczeń poprawnie rozwiązał całe zadanie (wartość 1), czy nie (wartość 0);
- jedna zmienna informująca o tym, czy uczeń poprawnie rozwiązał całe zadanie (wartość 2), rozwiązał poprawnie jeden z jego podpunktów (wartość 1) lub nie udzielił poprawnej odpowiedzi na żaden z podpunktów (wartość 0);
- jedna zmienna informująca o tym, czy uczeń poprawnie rozwiązał całe zadanie (wartość 1), udzielił odpowiedzi, jednak nie były one poprawne (wartość 0) lub w ogóle nie podjął próby rozwiązania zadania (wartość 7);
- dwie zmienne, z których pierwsza opisuje pierwszy podpunkt pytania, a druga drugi i mają one wartość 1, gdy udzielono poprawnej odpowiedzi na dany podpunkt, 0, gdy udzielono niepoprawnej odpowiedzi na dany podpunkt, a 7 gdy nie podjęto próby udzielenia odpowiedzi na dany podpunkt;
- 8 zmiennych, z których każda odpowiada jednej z możliwych odpowiedzi na poszczególne podpunkty zadania (pierwsze 4 na kolejne warianty odpowiedzi podpunktu 1, kolejne 4 na kolejne warianty odpowiedzi podpunktu 2) i mają one wartość 1, gdy dany wariant odpowiedzi został zaznaczony, a 1, gdy nie został zaznaczony.

Jak widać przytoczone powyżej sposoby różnią się tym, jakie informacje na temat zadania są gromadzone oraz w jaki sposób są one kodowane, nie sposób przy tym jednoznacznie zadecydować, który z wymienionych sposobów jest lepszy. Zależy to od tego, jakie analizy mają być wykonywane na zebranych danych, a więc tego, jakie informacje będą w tych analizach potrzebne (np. czy potrzeba będzie informacji o tym, którą odpowiedź wybrano, czy wystarczy informacja o tym, czy zadanie rozwiązano poprawnie). Wbrew pozorom nie zawsze najbardziej korzystne jest gromadzenie jak największej ilości jak najbardziej szczegółowych informacji, prowadzi to bowiem do powstania zbioru danych-molocha, wydobycie z którego prostych informacji wymaga wykonania wielu złożonych operacji na zawartych w nim danych (wyobraźmy sobie np., że chcemy na podstawie zbioru danych w ostatniej z powyższych postaci dowiedzieć się,

którzy uczniowie poprawnie rozwiązali rozważane zadanie).

Jak widać, określenie tego, w jaki sposób narzędzia badawcze reprezentowane będą przez *zmienne* w zbiorze danych nie jest rzeczą oczywistą, jednocześnie będąc kluczowym dla procesu analizy danych zebranych w badaniu (tego, jakie analizy będą możliwe oraz tego, jak będą pracochłonne). Stąd:

- **o tym, w jaki sposób narzędzia badawcze powinny być reprezentowane za pomocą *zmiennych* decydować powinny wspólnie osoby, które przygotowywały narzędzia badawcze oraz osoby, które będą analizować wyniki badania przeprowadzonego za pomocą tych narzędzi;**
- **nie sposób dobrze opracować sposób reprezentacji narzędzi badawczych za pomocą *zmiennych* nie wiedząc, w jaki sposób zebrane za ich pomocą dane będą analizowane.**

7 Po co w codebooku gromadzić aż tyle informacji na temat każdej *zmiennej*? Do czego one służą? Czy są w ogóle potrzebne?

Informacje, które gromadzone są codebooku na temat każdej zmiennej podzielić można na trzy kategorie:

- służące **dokumentacji** *zmiennych*, tak by osoby korzystające ze zbioru danych wiedziały, jaka informacja znajduje się w danej *zmiennej* oraz w jaki sposób jest ona w tej *zmiennej* zakodowana (patrz odpowiedź a) na pytanie *Po co mi codebook, przecież sposób zapisu informacji w zbiorze danych jest oczywisty?!*);
- służące opisaniu tych cech *zmiennych*, które są potrzebne do **weryfikacji** poprawności zbioru danych;
- ułatwiające wyszukiwanie *zmiennych* i ich kategoryzowanie, w szczególności na potrzeby **publikacji** zbioru danych za pomocą strony www IBE[1].

Dwie pierwsze kategorie informacji są więc niezbędne dla każdej zmiennej, gdyż bez nich nie będzie możliwa analiza danych przez osoby, które nie brały bezpośrednio udziału w tworzeniu zbioru danych, ani weryfikacja poprawności zbioru danych. Trzecia kategoria jest natomiast nieobowiązkowa.

informacja	opis	dokumentacja	weryfikacja	publikacja
nazwa	Nazwa pozwalająca jednoznacznie zidentyfikować zmienną, a najlepiej także określić, do czego się ona odnosi (np. do którego pytania kwestionariusza lub którego zadania w teście).	x	x	x
krótki opis	Krótki opis wyświetlany przy danej zmiennej w			x

1 Narzędzie takie miało być gotowe przed końcem 2011 roku, jednak wyłoniony w przetargu wykonawca rozpłynął się w powietrzu po zakończeniu etapu projektowania i w tej chwili (koniec listopada 2011) trwa proces rozwiązywania podpisanej z nim umowy, po czym przetarg zostanie rozpisany ponownie. W związku z tym na chwilę obecną przewidywana data przygotowania takiego narzędzia do czerwiec 2012 r.

	narzędziu do udostępniania zbiorów danych poprzez www.			
opis	Jeśli nazwa nie wyjaśnia w sposób dostateczny tego, jaka informacja zawarta jest w zmiennej, wyjaśnienie powinno się znaleźć w opisie. W wypadku codebooków opisujących kwestionariusze ankiet opis powinien zawierać pytanie kwestionariusza.	x		x
słowa kluczowe	Słowa kluczowe używane przy wyszukiwaniu zmiennych w narzędziu do udostępniania zbiorów danych poprzez www.			x
typ	Wskazuje sposób, w jaki została zakodowana *zmienna* – tekst, liczba, data, itd.	x	x	x
skala	Wskazuje na to, jakie operacje można wykonywać na danej *zmiennej*. Jest to kluczowa informacja z punktu widzenia późniejszej analizy danych.	x		x
rozmiar	Maksymalna liczba cyfr/znaków, z których może się składać wartość zmiennej, np. gdy zmienna może przyjmować wartości od 0 do 15, to jej rozmiar wynosi 2. Dla zmiennych typu *liczba rzeczywista* jest to łączna ilość cyfr całej liczby (zarówno części całkowitej, jak i ułamkowej, przy czym liczba zawsze musi mieć przynajmniej jedną cyfrę całkowitą, nawet jeśli zawsze będzie to 0).		x	
dokładność	Liczba miejsc dziesiętnych (cyfr) po przecinku, z jaką zapisywana jest wartość zmiennej (w wypadku zmiennych typu *liczba rzeczywista*).		x	
minimum	Minimalna wartość zmiennej, jeśli takowa istnieje.		x	
maksimum	Maksymalna wartość zmiennej, jeśli takowa istnieje.		x	
etykiety	Często wartości zmiennej zawierają jedynie kody liczbowe symbolizujące rzeczywiste wartości. Aby móc zorientować się, jakim rzeczywistym wartościom (etykietom) odpowiada dany kod liczbowy niezbędny jest „słownik”. Drugą rolą tej informacji jest zweryfikowanie, czy w zbiorze danych nie znajdują się wartości, dla których nie znamy ich rzeczywistego znaczenia (które nie znajdują się w „słowniku”).	x	x	x
warunek posiadania wartości	Niekiedy dana zmienna powinna mieć wartość jedynie pod pewnymi warunkami, np. zbiór danych zawiera informacje o wszystkich zeszytach testowych wypełnionych w badaniu, lecz dane zadanie znajdowało się tylko w jednym typie zeszytów i reprezentująca je zmienna powinna mieć wartość tylko wtedy, gdy dany wiersz zbioru danych	x	x	

	odnosi się do zeszytu tego typu. Informacja ta jest ważna, gdyż pozwala rozróżnić w zbiorze danych braki danych będące błędami od braków danych „celowych” (opisanych omawianymi warunkami).			

8 Jak w Excel-u wstawić znak nowej linii wewnątrz komórki?

Edytując wartość danej komórki wystarczy skorzystać z kombinacji klawiszy *ALT+ENTER*.

9 Jakie mogą być typy *zmiennej*?

Generalnie *zmienna* może być:

- liczbą całkowitą (np. liczba punktów za zadanie, kod wybranej odpowiedzi na pytanie, liczba osób w gospodarstwie domowym, kod rodzaju zeszytu, który rozwiązywał dany uczeń, itp.);
- liczbą rzeczywistą (np. liczba punktów za zadanie [jeśli można było dostawać ułamki punktów], , itp.);
- tekstem (np. nazwa wykonywanego zawodu, rodzaj zeszytu, który rozwiązywał dany uczeń, itp.);
- datą.

Wypada zwrócić uwagę, że ta sama informacja może być zapisana na kilka sposobów, w szczególności zaś wszystkie *zmienne*, których rzeczywiste wartości to *teksty*, przy czym liczba wszystkich możliwych wartości nie jest wielka i jest z góry znana, mogą być łatwo zapisane (zakodowane) jako *liczby całkowite*. W takim wypadku każdemu tekstowi przyporządkuje się kod liczbowy i opisuje to przyporządkowanie w „słowniku” w kolumnie *etykiety* (patrz oddzielne pytanie), np. rodzaj zeszytu (dajmy na to *His1, His2, Prz1, Prz2*) można przechowywać jako *tekst* lub przydzielić poszczególnym rodzajom kody (dajmy na to 1-His1, 2-His2, 3-Prz1, 4-Prz2) i przechowywać kody – wtedy zmienna będzie typu *liczba całkowita*. O tym, jaki sposób wybrać, należy decydować ze względu na późniejszą wygodę analizy danych. Jeśli wygodniej będzie posługiwać się tekstami, zapisywać *zmienną* jako *tekst*, jeśli wygodniej kodami liczbowymi, jako *liczba całkowita*[2].

Istnieje jeszcze kilka szczególnych typów, służących do zapisu specyficznych rodzajów danych:

- terc – jest to specjalny typ zmiennej służący do zapisu 6-cyfrowego kodu gminy wg GUS (alternatywnie 4-cyfrowego kodu powiatu lub 2-cyfrowego kodu województwa), jeśli informacje gromadzone były w podziale na regiony geograficzne – jeśli nie jesteś pewien, że masz w zbiorze danych zmienną tego typu, to nie potrzebujesz używać tego typu zmiennej; więcej informacji o tym typie zmiennych znajduje się w pytaniu 20 „Czym są zmienne typu *terc*?”
- waga – jest to specjalny typ zmiennej dla wag – wagi to współczynniki służące do przeliczania wyników analiz danych zebranych w badaniu na wyniki populacyjne (np. ogólnopolskie) – jeśli nie wiesz, o co chodzi, to na pewno nie potrzebujesz używać tego typu zmiennej.

2 Kody liczbowe będą bardziej wygodne w szczególności wtedy, gdy czasem będziemy chcieli oszukać i wykonać na zmiennej nominalnej operacje, których nominalnie nie można na niej wykonać (patrz pytanie o rodzaje skal) – na tekstach na pewno nie będzie to możliwe, na kodach liczbowych natomiast tak.

10 Jakie mogą być skale *zmiennej*?

Skala opisuje, jakie operacje można wykonywać na *zmiennej*. Rozważane operacje to:

- sprawdzanie równości (czy dwie wartości są sobie równe);
- porównywanie (która z wartości jest większa);
- odejmowanie (jaka jest różnica dwóch wartości);
- dzielenie (jaki jest iloraz dwóch wartości).

Rodzaje skal natomiast to:

- dychotomiczna – skala o dwóch przeciwstawnych wartościach (np. 1 – zaznaczono odpowiedź, 0 – nie zaznaczono odpowiedzi albo 1 – respondent odpowiedział twierdząco, 0 – odpowiedział negatywnie albo 1 – kobieta, 2 - mężczyzna); na skali dychotomicznej możliwe jest jedynie sprawdzenie, czy wartości są sobie równe;
- nominalna – skala, dla której można jedynie sprawdzać, czy wartości są sobie równe, nie da się ich natomiast uporządkować od najmniejszej do największej, nie ma sensu ich odejmowanie ani dzielenie (np. kolor oczu, rodzaj zeszytu wypełnianego przez ucznia, identyfikator zeszytu, itp.);
- porządkowa – skala, dla której można sprawdzać, czy jej wartości są równe oraz uporządkować je od najmniejszej do największej, nie ma jednak sensu ich odejmowanie ani dzielenie (np. wykształcenie);
- interwałowa – skala, dla której można sprawdzać, czy jej wartości są równe, uporządkować je od najmniejszej do największej, a różnica dwóch wartości daje się sensownie interpretować, nie ma jednak sensu dzielenie przez siebie dwóch wartości (np. daty);
- ilorazowa – skala, dla której można wykonywać wszystkie wymienione na początku tego punktu operacje, a wyniki tych operacji (w szczególności odejmowania i dzielenia) mają sensowną interpretację (np. masa, temperatura, liczba punktów uzyskanych w teście).

Więcej informacji na temat skal można znaleźć choćby na Wikipedii: http://pl.wikipedia.org/wiki/Skala_pomiarowa.

11 Jak wypełnić kolumnę *etykiety*?

W kolumnie etykiety znajduje się „słownik" pozwalający przetłumaczyć zakodowane liczbowo wartości znajdujące się w zbiorze danych na wartości, które można odnieść do rzeczywistości. Np. gdy wielkość miejscowości zamieszkania została zakodowana liczbowo jako 1, 2, 3 lub 4, bez „słownika" nie będziemy w stanie zinterpretować tych wartości.

„Słownik" taki składa się z kolejnych linii, z których każda opisuje jedną parę

kod liczbowy:wartość, którą można odnieść do rzeczywistości

np. dla powyższego przykładu z wielkością miejscowości mógłby on mieć postać:

1:wieś
2:miasto do 20 tys. mieszkańców
3:miasto pomiędzy 20, a 100 tys. mieszkańców
4:miasto powyżej 100 tys. mieszkańców

Jeśli nie wiesz, jak w Excelu wpisać wiele wierszy w jedną komórkę, spójrz na pytanie 8.

12 Jak wypełnić kolumnę *warunek*?

Warunek ma postać:

nazwaZmiennej operator wartość

gdzie *operatorem* może być =, *!*= (różne), <, >, <=, >=, np. jeśli rodzaj zeszytu znajduje się w zmiennej *rodzajZeszytu* i jest kodowany tekstowo, a zmienna powinna mieć wartość tylko gdy rodzaj zeszytu ma wartość *His1*, warunek będzie wyglądał następująco:

rodzajZeszytu = His1

Należy pamiętać o **konieczności oddzielania spacją** nazwy zmiennej od operatora i operatora od wartości, np. warunek

rodzajZeszytu=His1 **nie jest poprawny**.

W wypadku bardziej złożonych warunków, gdy to, czy *zmienna* powinna mieć wartość, zależy od kilku czynników, każdy czynnik należy zapisać w nowej linii, poprzedzając go za pomocą *i* bądź *lub*, np. gdy zmienna powinna mieć wartość, gdy rodzaj zeszytu do *His1* lub *His2*, wtedy warunek będzie miał postać:

rodzajZeszytu = His1
lub rodzajZeszytu = His2

Jeśli zmienna zawsze powinna mieć wartość, pole należy pozostawić puste.

13 Dlaczego codebooki IBE są w formacie CSV?

Jako format zapisu codebooka przyjęty został CSV, gdyż jest to format bardzo uniwersalny - można go otworzyć w każdym arkuszu kalkulacyjnym, programie statystycznym, a nawet w notatniku. Również pisząc programy (np. do weryfikacji poprawności zbioru danych lub do prezentacji zbiorów danych) najłatwiej odczytać informacje z codebook-a zapisanego właśnie w tym formacie.

Jeśli jednak komuś wygodniej na co dzień pracuje się z codebookiem zapisanym w innym formacie (np. w XLS), nic nie stoi na przeszkodzie, by zapisać i na co dzień korzystać z codebooka zapisanego w innym formacie.

14 Jak poprawnie zapisać codebook w formacie CSV za pomocą Ms Excel

Należy kliknąć: okrągła ikonka MsOffice w lewym górnym rogu programu → zapisz jako → inne formaty (ostatnia pozycja na dole).

Wyświetli się okienko, w którym z listy *zapisz jako typ:* należy wybrać pozycję ***CSV (rozdzielany przecinkami)***.

Nie należy jako formatu wybierać *CSV (MS-DOS)* ani *CSV (Macintosh)*.

15 Jak mogę zweryfikować poprawność codebooka?

Pod adresem https://zai.ibe.edu.pl/zbiory można dokonać weryfikacji poprawności codebooka (jak również zbioru danych, którego strukturę dany codebook opisuje). Wystarczy wskazać w polu *codebook* plik codebooka i nacisnąć przycisk *sprawdź*.

16 Ile trwa weryfikacja zbioru danych?

Weryfikacja zbioru danych za pomocą strony https://zai.ibe.edu.pl/zbiory odbywa się z prędkością ok. 500 wartości zmiennych na sekundę. Stąd szacunkowy czas weryfikacji zbioru to *LiczbaZmiennych * LiczbaRekordów / 500* [sekund].

17 Jak można skrócić czas weryfikacji zbioru danych?

Ponieważ większość błędów wykrywanych w zbiorach danych ma charakter systematyczny, wynikający z błędów w strukturze zmiennych opisanej w codebook-u (w szczególności z błędnego używania kodu specjalnego „nie dotyczy" - patrz pytanie 18.), a nie losowy (błędy występują przypadkowo, w tylko niektórych rekordach), to najlepszym sposobem ograniczenia czasu weryfikacji wydaje się być skorzystanie z możliwości podania maksymalnej liczby wykrytych błędów, po której weryfikator zakończy sprawdzanie i ustawienie jej na małą wartość, np. 100.

Inna możliwość optymalizacji weryfikacji zbioru danych to podział zbioru danych na mniejsze części i weryfikacja każdej z nich oddzielnie (w szczególności można stronę weryfikatora otworzyć w różnych zakładkach i w każdej z nich równolegle uruchomić sprawdzanie mniejszego fragmentu zbioru danych).

18 Co oznacza komunikat *maksymalna wartość zmiennej koliduje z wartościami specjalnymi*?

W wypadku *zmiennych* typu *liczba całkowita* i *liczba rzeczywista* **trzy ostatnie** wartości całkowite *zmiennej* wykorzystywane są na kody specjalne, za pomocą których oznacza się nietypowe wartości *zmiennej*: nieudzielenie odpowiedzi, fakt, że *zmienna* nie powinna mieć wartości ze względu na *warunek* (patrz kolumna codebooka *warunek*), itp.

W związku z tym zmienna typu *liczba całkowita* o rozmiarze 1, może przyjmować de facto tylko siedem wartości „użytkowych": 0, 1, 2, 3, 4, 5, 6, pozostałe bowiem (trzy ostatnie: 7, 8, 9) są zarezerwowane na kody specjalne.

Stąd, gdy np. dana zmienna przechowywać będzie 8 „użytkowych" wartości (np. pytanie z jedną z 8 odpowiedzi do wyboru), jej rozmiar musi być zwiększony do 2 – kodami specjalnymi będą wtedy 97, 98 i 99 i maksymalna wartość zmiennej nie będzie z nimi kolidować.

W wypadku zmiennych typu *liczba rzeczywista* i *waga* zarezerwowane są wszystkie wartości większe lub równe od najmniejszego z kodów specjalnych (a więc np. dla zmiennej typu *liczba rzeczywista* o rozmiarze 2 zastrzeżone są wszystkie wartości większe lub równe od *97*).

19 Co to są kody specjalne i jakie przyjmują wartości?

Kody specjalne służą do oznaczania przyczyny, dla której dana zmienna nie ma „normalnej wartości" (tzn. takiej, która jest opisana w codebooku przez maksymalną i minimalną wartość zmiennej i/lub zestaw etykiet tej zmiennej), a więc takich sytuacji, jak:

α) zmienna nie ma wartości, gdyż danej osobie w ogóle nie zadawano tego pytania lub nie robiła zeszytu ćwiczeń z danym zadaniem, itp. (tzw. „nie dotyczy"); warto zwrócić uwagę, że jest to równoważne niespełnieniu warunku dla filtru określonego w codebooku;

β) zmienna nie ma wartości, gdyż dana osoba nie umiała udzielić jednoznacznej odpowiedzi na dane pytanie (tzw. „trudno powiedzieć") lub w ogóle nie podjęła próby rozwiązania zadania w zeszycie ćwiczeń;

χ) zmienna nie ma wartości gdyż dana osoba odmówiła udzielenia odpowiedzi na dane pytanie (tzw. „odmowa odpowiedzi") lub odpowiedziała w sposób formalnie niepoprawny na dane pytanie w zeszycie ćwiczeń (np. zakreśliła kilka odpowiedzi w pytaniu, w którym należało zaznaczyć tylko jedną poprawną odpowiedź).

W zależności od typu zmiennej kody te przyjmują różne wartości:

typ zmiennej	nie dotyczy	trudno powiedzieć / nie podjęto próby rozwiązania	odmowa odpowiedzi / wielokrotne zaznaczenie
liczba całkowita o rozmiarze N	7, gdy N = 1 97, gdy N = 2, 997, gdy N = 3, itd.	8, gdy N = 1 98, gdy N = 2, 998, gdy N = 3, itd.	9, gdy N = 1 99, gdy N = 2, 999, gdy N = 3, itd.
liczba rzeczywista o rozmiarze N i dokładności D	7, gdy N – D = 1, 97, gdy N – D = 2, 997, gdy N – D = 3, itd.	8, gdy N – D = 1, 98, gdy N – D = 2, 998, gdy N – D = 3, itd.	9, gdy N – D = 1, 99, gdy N – D = 2, 999, gdy N – D = 3, itd.
data	9997-00-00 00:00:00	9998-00-00 00:00:00	9999-00-00 00:00:00
tekst	NIE DOTYCZY	TRUDNO POWIEDZIEĆ	ODMOWA ODPOWIEDZI
waga	jak liczba rzeczywista		
terc	jak liczba całkowita		

20 Dlaczego zmienna typu *liczba całkowita* na skali *nominalnej* musi posiadać wartość w kolumnie *etykiety*?

Przyjęto tak dlatego, że połączenie skali nominalnej z reprezentacją w postaci liczby całkowitej sugeruje, że same wartości liczbowe nic nie znaczą i że nie sposób bez podania etykiet domyśleć się ich znaczenia. Zdarzają się jednak *zmienne*, w wypadku których znaczenie danej wartości nie jest nam do niczego potrzebne, chodzi bowiem tylko o odróżnienie jednych wartości od innych (np. identyfikator osoby wprowadzającej dane do zbioru danych). W takim wypadku należy zmienić typ *zmiennej* na *tekst*.

21 Czym są zmienne typu *terc*?

Są to specjalne zmienne, służące do przechowywania informacji o gminie, powiecie lub województwie za pomocą kodów gmin/powiatów/województw nadawanych przez Główny Urząd Statystyczny.

Jeśli jakaś zmienna jest typu *terc*, wtedy:

- w kolumnie *rozmiar* należy wpisać:
  - 6 – jeśli wartości zmiennej to kody gmin;
  - 4 – jeśli wartości zmiennej to kody powiatów;
  - 2 – jeśli wartości zmiennej to kody województw;
- w kolumnie *dokładność* należy podać rok, z którego pochodzą kody

gmin/powiatów/województw (podział administracyjny Polski na poziomie gmin i powiatów ulega co roku niewielkim zmianom, stąd niezbędne jest określenie roku, względem którego sprawdzana będzie poprawność kodów TERC).

- wartości w kolumnach *minimum, maksimum* oraz *etykiety* nie mają znaczenia (nie są brane pod uwagę podczas weryfikacji).

Wartości zmiennej typu *terc*:

- należy uzupełniać zerami wiodącymi do długości stosownej dla rozmiaru zmiennej (np. kod powiatu bolesławieckiego w latach 1999-2013 – *212* należy uzupełnić do *0212*);
- dysponując 7-cyfrowymi kodami gmin należy obciąć ostatnią cyfrę (nie identyfikuje ona gminy, a jedynie oznacza jej rodzaj).

Wartości zmiennej typu *terc* są weryfikowane względem bazy danych o podziale administracyjnym Polski znajdującej się na serwerach IBE, aktualizowanej co roku na podstawie danych GUS.

22 Nie podoba mi się wzór codebooków obowiązujący w IBE, czy muszę się do niego stosować?

Do wzorca codebooka przyjętego w IBE nie trzeba się stosować gdy spełnione są jednocześnie dwa warunki:

- zbiór danych z badania nie będzie w przyszłości upubliczniany (w szczególności za pośrednictwem strony www IBE);
- nie ma potrzeby, by ZAI dokonało weryfikacji poprawności zbioru danych dostarczonego przez wykonawcę badania.

Data	Osoba	Opis modyfikacji
2011-11-30	Mateusz Żółtak	Pierwotna wersja dokumentu
2011-12-02	Katarzyna Wądołowska	• korekta wartości specjalnych omawianych w pytaniu 18 „Co oznacza komunikat maksymalna wartość zmiennej koliduje z wartościami specjalnymi?” • dopisanie pytania 19 „Jakie są kody specjalne?”
2011-12-16	Matusz Żółtak	• dopisanie pytania 14 „Jak poprawnie zapisać codebook w formacie CSV za pomocą Ms Excel”
2012-01-25	Mateusz Żółtak	• uszczegółowienie opisu kodu „nie dotyczy” w pytaniu 18 „Jakie są kody specjalne?” • dopisanie pytania 16 „Ile trwa weryfikacja zbioru danych?” • dopisanie pytania 17 „Jak można skrócić czas weryfikacji zbioru danych?”
2012-03-20	Mateusz Żółtak	• korekta pytania 12 „Jak wypełnić kolumnę *warunek*?” - mocniejsze zaakcentowanie konieczności oddzielania spacją nazwy zmiennej od operatora i operatora od wartości
2013-02-03	Mateusz Żółtak	• ujednolicenie pytań 18 „Co oznacza komunikat *maksymalna wartość zmiennej koliduje z wartościami specjalnymi*?” i 19 „Jakie są kody specjalne?” - istnieją tylko trzy, a nie pięć wartości specjalnych • doprecyzowanie w pytaniu 18 „Co oznacza komunikat *maksymalna wartość zmiennej koliduje z wartościami specjalnymi*?” zakresu wartości „użytkowych” dla typów *liczba rzeczywista* oraz *waga* • dodanie pytania 20 „Czy są zmienne typu *terc*?”
2013-12-09	Mateusz Żółtak	• rozbudowanie pytania 19 „Jakie są kody specjalne” o informacje o wszystkich typach zmiennych; w związku z tym zmiana tytułu na „Co to są kody specjalne i jakie przyjmują wartości?”