Rozdział 1

Analiza Struktury

Jan Żółtowski

Problem 1.1

Kuratorium w Łodzi postanowiło ocenić wpływ warunków szkolnych i pozaszkolnych na wyniki uczniów piszących próbną EMaturę z matematyki[1]. W badaniu ankietowym wzięło udział 361 uczniów. Odpowiadali oni na pytania, wybierając jeden z podanych wariantów odpowiedzi. Z dwudziestu pytań zamieszczonych w kwestionariuszu do analizy wybrano tylko pięć przedstawionych w tabeli 1.1.

Dodatkowo wyniki badania ankietowego wzbogacono, podając dla każdego ucznia liczbę punktów, jaką uzyskał na próbnym egzaminie maturalnym z matematyki.

Tabela 1.1. Wybrane pytania ankiety dla uczniów piszących próbną EMaturę

Lp.	Pytanie	Odpowiedź	
1	2	3	4
1.	Czy posiadasz komputer w domu?	a	Nie
		b	Tak (bez dostępu do internetu)
		c	Tak (z dostępem do internetu)
2.	Czy chodzisz na dodatkowe lekcje z matematyki?	a	Nie, i nie chcę chodzić
		b	Nie, ale chcę chodzić
		c	Tak, raz w tygodniu
		d	Tak, więcej niż raz w tygodniu

[1] Dane uzyskane z badań ankietowych prowadzonych na uczestnikach próbnej EMatury z roku 2009, udostępnione przez organizatora – dr J. Stańdo.

1	2	3	4
3.	W szkole podstawowej rodzice pomagali mi odrabiać lekcje	a	Rzadko
		b	Codziennie, około pół godziny
		c	Codziennie, około godziny
		d	Codziennie, ponad godzinę
4.	Podaj ocenę z matematyki na koniec II klasy 1 2 3 4 5 6		
5.	Jak sądzisz, ile punktów uzyskasz z tej próbnej matury (max. 50)?		

Źródło: kwestionariusz badań uczestników próbnej EMatury 2009.

Organizatorzy zainteresowani są wynikami badania, a zwłaszcza ich właściwą interpretacją. Należy zatem przeprowadzić wszechstronną analizę zebranych danych (przedstawionych w pliku *rozdział1.xls*) i odpowiednio je zinterpretować.

Opis metody rozwiązania problemu: skale pomiarowe

Aby przystąpić do analizy danych i ich opisu, w pierwszej kolejności należy zastanowić się nad badanymi cechami statystycznymi oraz zastosowanymi skalami pomiarowymi, a tym samym sposobem prezentacji zmiennych[2].

Do badania struktury zbiorowości, ze względu na natężenie cechy mierzalnej, nie wystarcza prezentacja materiału statystycznego w postaci szeregów czy też wykresów, zwłaszcza w przypadku dysponowania dużą ilością informacji. Do ilościowej analizy właściwości statystycznych zbiorowości i ich porównania wykorzystuje się charakterystyki liczbowe nazywane **parametrami opisowymi** zbiorowości statystycznej. Są to liczby, które w sposób syntetyczny określają właściwości badanej zbiorowości oraz pozwalają na porównanie kilku zbiorowości ze względu na tę samą cechę lub na analizę kilku różnych cech w ramach tej samej zbiorowości.

Zbiorowość statystyczna (zwana inaczej populacją statystyczną) jest zbiorem dowolnych elementów (osób, rzeczy, zjawisk lub faktów), które są objęte badaniem statystycznym.

[2] Opracowano na postawie [Białek, Depta, 2010, s. 7–10].

Zbiorowość generalna (populacja generalna) jest zbiorem dowolnych elementów (obiektów, zdarzeń) nieidentycznych z punktu widzenia badanej cechy, obejmującym wszystkie elementy będące przedmiotem badania, w odniesieniu do których można formułować wnioski ogólne. Jeżeli elementy zbiorowości generalnej poddaje się badaniu tylko ze względu na jedną cechę, to zbiorowość taką nazywa się jednowymiarową (jednocechową). Zbiorowość nazywana jest wielowymiarową (wielocechową), jeżeli rozpatruje się wiele cech.

Zaliczając jednostki statystyczne do danej zbiorowości, w celu uzyskania porównywalności materiału statystycznego, należy określić wszystkie jednostki pod względem:

- **rzeczowym** (co lub kogo poddajemy badaniu statystycznemu);
- **przestrzennym** (jaki obszar jest objęty badaniem);
- **czasowym** (jaki okres obejmuje badanie lub w jakim momencie się ono odbywa).

Wymienione własności określane są jako tzw. **cechy stałe.** Dzielimy je na cechy **rzeczowe**, **przestrzenne** i **czasowe**. Są one wspólne wszystkim jednostkom danej zbiorowości statystycznej, nie podlegają pomiarowi, a jedynie decydują o zaliczeniu jednostki do określonej zbiorowości.

Cechy zmienne to właściwości, które różnią poszczególne jednostki statystyczne i podlegają obserwacji, czyli pomiarowi. Występują one u poszczególnych jednostek w formie jednego z k możliwych wariantów (rodzajów), przy czym $k \geq 2$. Z badaną **cechą statystyczną**, czyli właściwością, którą odznaczają się wszystkie jednostki wchodzące w skład badanej zbiorowości, ściśle związany jest rodzaj skali pomiarowej stosowany w danym przypadku. Najczęściej wyróżniamy: **cechy mierzalne** (inaczej ilościowe, wymierne) oraz **cechy niemierzalne** (inaczej jakościowe, niewymierne).

Cechy mierzalne to takie właściwości, które można zmierzyć i wyrazić liczbą za pomocą określonej jednostki miary (np. wiek w latach, długość w metrach, liczba w sztukach).

Cechy niemierzalne to takie, których nie można zmierzyć, a jedynie stwierdza się występowanie lub niewystępowanie określonego ich wariantu. Są one zwykle opisywane słownie. W zależności od liczby wariantów badanej cechy niemierzalnej wyróżnia się klasyfikację dwudzielną (dychotomiczną), w przypadku występowania dwóch wariantów cechy (np. płeć), oraz klasyfikację wielodzielną (politomiczną), w przypadku liczby wariantów większej niż dwa (np. stopień wykształcenia).

Do **cech niemierzalnych** z reguły zalicza się również cechy ***quasi*-ilościowe,** zwane porządkowymi. Cechy te kwantyfikują zwykle natężenie badanej właściwości przedstawionej w sposób opisowy, porządkując w ten sposób zbio-

rowość (np. ocena wiadomości ucznia: celująca, bardzo dobra, dobra, lub krócej – 6, 5, 4 itp.).

Należy jednak podkreślić, że przyporządkowanie cechy do odpowiedniego jej rodzaju jest ściśle związane ze sposobem określenia wariantu cechy dla danej jednostki. Wiek mierzony w latach (miesiącach itp.) to cecha ilościowa, ale gdy zdefiniujemy warianty wieku w postaci: stary, średni, młody, dziecko, niemowlę, wówczas będzie to cecha porządkowa. Podobnie, jeśli określamy kolor, to z reguły traktujemy go jako cechę niemierzalną (warianty czarny, biały, zielony itd.), ale wystarczy przyjąć, że kolor jest określony jako długość fali świetlnej w nm, aby była to cecha ilościowa.

Do pomiaru cech statystycznych zazwyczaj stosowane są cztery podstawowe skale pomiarowe.

- **Skala nominalna** (można wtedy określić relację: równe lub różne), czyli taka skala, w której pomiar polega na podzieleniu całego zbioru wyników na rozłączne podzbiory, a następnie zidentyfikowaniu jednostki ze względu na posiadanie lub nieposiadanie określonego typu cechy. Klasyfikujemy wówczas jednostkę statystyczną do określonej kategorii, a poszczególnym kategoriom jakościowym badanych cech przypisujemy liczbę lub nazwę. Szczególnymi przypadkami tej skali są skale dwudzielne (dychotomiczne), gdy mamy do czynienia z dwoma wariantami cechy (np. płeć: kobieta lub mężczyzna), i trychotomiczne, gdy występują trzy warianty cechy (np. odpowiedź na pytanie: tak, nie lub nie wiem).
- **Skala porządkowa** (nazywana inaczej skalą rangową, gdy można określić relację: większy lub mniejszy) – umożliwia ona przydział (kwalifikację) jednostek zbiorowości do podzbiorów według stopnia natężenia cechy. Uporządkowanie to może być wykonane w porządku rosnącym lub malejącym i określane jest mianem **rangowania**. Jako przykład zmiennej mierzonej na tej skali można podać poziom wykształcenia (np. ukończona szkoła: podstawowa, gimnazjum, szkoła pogimnazjalna, studia wyższe I stopnia, studia wyższe II stopnia, lub samopoczucie: bardzo dobre, dobre, złe, bardzo złe).
- **Skala przedziałowa** (inaczej interwałowa, jeśli można określić relację: większe o tyle) – występuje wówczas, gdy pomiary badanych cech są wyrażone w postaci liczb rzeczywistych. W skali tej możliwe jest porównywanie analizowanych jednostek statystycznych przez określenie różnicy, czyli przedziału między poszczególnymi jednostkami. Skala przedziałowa ma ustaloną jednostkę miary, nie zawiera jednak naturalnego lub absolutnego punktu zerowego. Na przykład można stwierdzić, że temperatura w danym dniu jest o 10^{o}C wyższa niż dnia poprzedniego. „Punkt zero" w tej skali ustalony jest umownie – nie można zatem stwierdzić, że temperatura 15^{o}C jest trzykrotnie wyższa od 5^{o}C, można tylko określić różnicę. W badaniach ekonomicznych bardzo często wykorzystuje

się tę skalę w pytaniach o wynagrodzenie, np. poniżej 1000 zł, 1000–3000 zł, 3000–5000 zł, powyżej 5000 zł.

▶ **Skala ilorazowa** (stosunkowa) występuje wówczas, gdy można określić relację: tyle razy większe. Zawiera ona zero bezwzględne, czyli absolutne, co oznacza, że jeżeli cecha przyjmuje wartość równą zeru, to jest to jednoznaczne z brakiem jej występowania. Na tej skali można wykonywać wszystkie działania arytmetyczne. Jest ona wykorzystywana do pomiaru cech mierzalnych, zarówno ciągłych, jak i skokowych. Stosowanie tej skali umożliwia porównanie jednostek za pomocą charakterystyk względnych, np. cena 100 zł/kg jest dwukrotnie większa niż 50 zł/kg.

Skale: nominalna i porządkowa należą do tzw. **słabych skal** pomiarowych, natomiast przedziałowa i ilorazowa należą do **skal mocnych**. Skale słabe są stosowane najczęściej do zmiennych jakościowych, a skale mocne do zmiennych mierzalnych – skokowych i ciągłych.

Stosowanie określonych miar statystycznych zależy zarówno od rodzaju cechy, jak i zastosowanej skali pomiarowej. W przypadku cech niemierzalnych, określonych na skali nominalnej, możliwe jest zastosowanie: **wskaźników struktury** (procentów), o ile liczebność próby jest dostatecznie duża (zazwyczaj co najmniej 100 elementów, choć niekiedy wystarcza kilkadziesiąt), a także określenie najczęstszego wariantu cechy, nazywanego **dominantą** lub modalną.

Dla cech *quasi*-ilościowych (porządkowych) możliwe jest dodatkowo określenie pozycyjnych miar położenia, takich jak: mediana, kwantyle, decyle. Ze względu na opisowy charakter tych miar nie można jednak wykonywać na nich działań arytmetycznych. Zwrócić należy także szczególną uwagę na cechy porządkowe, które dzięki przypisaniu im rang liczbowych traktujemy *de facto* jak cechy ilościowe. Przykładem mogą być oceny szkolne. Przyjmując, że ocena bardzo dobra to „5", a dostateczna to „3" itd., w praktyce wyznacza się następnie średnią z ocen czy też odchylenie standardowe.

W przypadku analizy cech mierzalnych można oczywiście wyznaczać klasyczne miary położenia (średnią arytmetyczną, geometryczną czy harmoniczną), zmienności, asymetrii oraz koncentracji.

W tej publikacji zdecydowano się nie wprowadzać teorii związanej ze statystycznymi miarami struktury. Wszystkich zainteresowanych pogłębieniem wiedzy z tej tematyki odsyłamy do literatury przedmiotu [Witkowska (red.), 2004a; Aczel, 2000; Zając, 1974; Ostasiewicz, Rusnak, Siedlecka, 2003; Jóźwiak, Podgórski, 2006; Białek, Depta, 2010].

Rozwiązanie problemu za pomocą arkusza kalkulacyjnego EXCEL

Wracając do opisu przedstawionego badania, należy stwierdzić, że zbiorowością statystyczną są w tym przypadku uczniowie szkoły średniej przystępujący do matury w 2009 r. Zbiorowością próbną byli uczniowie szkoły średniej przystępujący do Łódzkiej EMatury w 2009 r. Cechami stałymi w badaniu są:

- rzeczowa – uczniowie szkoły średniej przystępujący do EMatury w 2009 r.;
- czasowa – 8 marca 2009 r.;
- przestrzenna – dla zbiorowości: teren Polski, dla próby: region łódzki.

W przytoczonym fragmencie badań poddano analizie sześć cech zmiennych: pięć zmiennych określonych w kwestionariuszu oraz szóstą zmienną – wynik punktowy z próbnej EMatury. Dla każdej z tych zmiennych ustalimy jej rodzaj i użytą skalę pomiarową. W następnej kolejności, wykorzystując odpowiednie funkcje statystyczne pakietu EXCEL, wyznaczymy te miary statystyczne, które mają dla badanych cech sens oraz podamy ich interpretacje.

Dane zostały wprowadzone do arkusza kalkulacyjnego EXCEL (analizowaną bazę danych zapisano w pliku *rozdz1.xls*) – zob. ekran 1.1.

Plik Edycja Widok Wstaw Format Narzędzia Dane Okno Pomoc

Calibri 11 100%

I9 1

	A	B	C	D	E	F	G	H	I
1	Numer	Liczba pu	Czy posiad	Ile czasu dzi	Czy chodzis	W szkole p	Podaj oce	Jak myś	ni
2	1	8	Tak (z dostępe	Ponad 2 godziny	Nie, ale chcę c	Rzadko	2	20	1
3	2	15	Tak (z dostępe	Od 20 min do 1	Nie, ale chcę c	Codziennie ok	2	17	1
4	3	13	Tak (z dostępe	Ponad 2 godziny	Nie, ale chcę c	Rzadko	3	5	1
5	4	5	Tak (z dostępe	Mniej niż 20minu	Nie, ale chcę c	Codziennie ok	2	5	1
6	5	20	Tak (z dostępe	Od 20 min do 1	Tak, raz w tygc	Codziennie pc	4	30	1
7	6	22	Nie	Ponad 2 godziny	Tak, raz w tygc	Rzadko	5	30	1
8	7	16	Tak (z dostępe	Ponad 2 godziny	Nie, ale chcę c	Rzadko	4	30	1
9	8	16	Tak (z dostępe	Ponad 2 godziny	Tak, więcej raz	Codziennie ok	2	18	1
10	9	32	Tak (bez dostę	Ponad 2 godziny	Tak, raz w tygc	Codziennie pc	4	30	1
11	10	14	Tak (z dostępe	Ponad 2 godziny	Tak, więcej raz	Rzadko	2	15	1
12	11	15	Tak (z dostępe	Ponad 2 godziny	Nie, ale chcę c	Codziennie ok	2	28	1
13	12	15	Tak (z dostępe	1-2 godzin dzier	Nie, ale chcę c	Codziennie pc	3	16	1
14	13	18	Tak (z dostępe	Ponad 2 godziny	Tak, raz w tygc	Rzadko	3	15	1
15	14	8	Tak (bez dostę	Mniej niż 20minu	Nie, i nie chcę	Rzadko	3	0	1
16	15	36	Tak (z dostępe	Od 20 min do 1	Nie, ale chcę c	Rzadko	3	43	1

Ekran 1.1. Baza danych EMatura 2009

Źródło: opracowanie własne.

Zmienna określająca posiadanie przez ucznia komputera w domu (warianty odpowiedzi: nie; tak, bez dostępu do internetu; tak, z dostępem do internetu) jest typową zmienną niemierzalną, a do jej pomiaru stosujemy skalę nominalną. Należy zatem w pierwszej kolejności zliczyć, ile osób zadeklarowało określony wariant cechy. Możemy w tym przypadku wykorzystać statystyczną funkcję: LICZ.JEZELI(zakres; kryteria)[3]. Należy jednak pamiętać, aby „zablokować" zakres adresowy komórek $C$2:$C$362 przed skopiowaniem go do komórek L8 i L9. Jeśli tego nie zrobimy i w komórce L7 pozostawimy zakres C2:C362, to po skopiowaniu formuły w komórce L8 pojawi się zakres zmiennej zawierający komórki od C3:C363. Uwaga ta dotyczy również pozostałych badanych zmiennych. Liczbę osób posiadających dany wariant cechy można również wyznaczyć, wykorzystując tabele przestawne.

Z uwagi na użytą skalę nominalną, w przypadku tej zmiennej możemy jedynie określić najczęstszy wariant, w jakim występuje badana zmienna, oraz wyznaczyć wskaźniki struktury (np. w procentach), dzieląc liczbę osób posiadających dany wariant cechy przez łączną liczbę osób (zob. ekran 1.2).

L7 =LICZ.JEŻELI($C$2:$C$362;K7)

	K	L	M	N
5	**Czy posiadasz komputer w domu?**			
6		liczba osób n_i	w_i	
7	Nie	6	1,66%	
8	Tak (bez dostępu do Internetu)	26	7,20%	
9	Tak (z dostępem do Internetu)	329	91,14%	
10	**Σ**	361	100%	
11				

Ekran 1.2. Sposób wyznaczenia liczebności poszczególnych wariantów badanej cechy z wykorzystaniem programu EXCEL

Źródło: obliczenia własne.

W badanej grupie uczniów zdecydowana większość, a mianowicie 329 osób (czyli 91,1% ogółu ankietowanych), ma w domu komputer z dostępem do internetu; komputer bez dostępu do internetu posiadało 26 uczniów (7,2% ogółu), brak komputera deklarowało zaś zaledwie 6 uczniów (1,66% ogółu). W chwili obecnej komputer staje się dla uczniów wręcz niezbędnym wyposażeniem i źródłem nie tylko informacji czy komunikacji, lecz także rozrywki.

Zmienna określająca uczestnictwo w dodatkowych lekcjach matematyki jest także zmienną jakościową (niemierzalną). W przypadku tej zmiennej możliwe

[3] Składnia funkcji: **zakres** – argument wymagany, określa jedną lub więcej komórek, które mają zostać zliczone, **kryteria** – argument wymagany – oznacza liczbę, wyrażenie, odwołanie do komórki lub ciągu tekstowego, określającego, które komórki będą zliczane.

jest jedynie wyznaczenie częstości występowania poszczególnych wariantów cechy oraz wyznaczenie wskaźników struktury (procentów) – zob. tabela 1.2.

Tabela 1.2. Uczestnictwo uczniów piszących próbną maturę w dodatkowych lekcjach

Wariant odpowiedzi	n_i	w_i (w %)
Nie i nie chcę chodzić	61	16,90
Nie, ale chcę chodzić	100	27,70
Tak, raz w tygodniu	158	43,77
Tak, więcej niż raz w tygodniu	42	11,63
Suma	**361**	**100,00**

Źródło: obliczenia własne.

Najwięcej przebadanych uczniów, bo 158 (co stanowiło 43,8% ogółu ankietowanych), uczęszczało na dodatkowe lekcje matematyki raz w tygodniu; 100 uczniów (27,8% ogółu) nie chodziło na dodatkowe lekcje matematyki, ale wyraziło chęć uczestniczenia w nich. Dla 61 uczniów (16,9%) chodzenie na dodatkowe lekcje nie jest konieczne (ich zdaniem), natomiast 42 uczniów (11,6% ogółu) uczestniczyło w dodatkowych lekcjach matematyki częściej niż raz w tygodniu.

Zmienna określająca, „jak często w szkole podstawowej rodzice pomagali uczniom odrabiać lekcje", ma charakter cechy porządkowej i do jej opisu zastosujemy skalę porządkową (tab. 1.3).

Tabela 1.3. Pomoc rodziców przy odrabianiu lekcji w szkole podstawowej dla grupy maturzystów

Pomoc rodziców	n_i	w_i (w %)	n_{sk}
Rzadko	219	60,66	219
Codziennie około pół godziny	54	14,96	273
Codziennie około godziny	56	15,51	329
Codziennie ponad godzinę	32	8,86	361
Suma	**361**	**100,00**	X

Źródło: obliczenia własne.

W tym przypadku warianty cechy zostały uporządkowane zgodnie z hierarchią „od wartości najmniejszej do największej". Zakładamy, że rodzice, którzy pomagali dziecku „codziennie około pół godziny" przy odrabianiu lekcji,