Statystyka po ludzku

Paweł Tatarzycki

ZŁOTE MYŚLI

Jak bez problemu zdać egzamin ze statystyki?

Autor: Paweł Tatarzycki
Tytuł: Statystyka po ludzku

Data: 08.08.2012

Złote Myśli Sp. z o.o.
ul. Toszecka 102
44-117 Gliwice
www.zlotemysli.pl
email: kontakt@zlotemysli.pl

SPIS TREŚCI

Wstęp

Celem tej publikacji jest „poukładanie" obszernego materiału ze statystyki, ze wskazaniem na praktyczne zastosowania nabywanej wiedzy w tym zakresie. W myśl zasady *stopniowania trudności* – najtrudniejsze, najbardziej złożone zagadnienia omówiono pod koniec tego opracowania. Przykładowo, dobór próby – mimo że jest to elementarne pojęcie statystyki – omówiono w rozdziale ostatnim, co jest konsekwencją wprowadzonej zasady.

Aby ułatwić przejścia do pokrewnych tematów czy trudnych pojęć statystycznych, zastosowano nowatorskie rozwiązanie na wzór hiperłączy internetowych. Rozwiązanie to ma szczególne znaczenie przy powtarzaniu materiału na „za pięć dwunasta", przed kolokwium czy egzaminem. I tak np. odwołanie w kolorze hiperłącza „(zob. Dobór próby)" przyciąga uwagę Czytelnika. W wersji elektronicznej możliwe jest kliknięcie na linku powodujące przejście do podrozdziału „Dobór próby".

W myśl zasady *związku teorii z praktyką* wprowadzany materiał wyjaśniany jest na przykładach, co ułatwia jego zrozumienie, a dodatkowo czyni naukę ciekawszą. Integralną częścią publikacji są przykłady wykonane w arkuszu kalkulacyjnym MS Excel. W tekście publikacji znajdują się informacje typu (zob. *Przykłady…*).

Każdy większy dział „wieńczy" zestaw zadań do samodzielnego wykonania, poprzedzonych rozbudowanym przykładem, zawartych w podrozdziałach „Trening i ewaluacja". Czytelnik może dokonywać analiz, wykorzystując szereg danych praktycznych zebranych w pliku *Dane_do_analizy.xls*. Obok tradycyjnych zadań – w większości działów sprawdzających zamieszczono testy wielokrotnego wyboru, które Czytelnik z łatwością

sprawdzi w specjalnie przygotowanych w tym celu arkuszach MS Excel pt. *Ewaluacja*.

Animacje, czyli prezentacje PowerPoint ukazujące w sposób dynamiczny wykonywanie złożonych czynności obliczeniowych w arkuszu kalkulacyjnym Excela, są pomocne przy studiowaniu rozbudowanych przykładów w działach „Trening i ewaluacja", jak również przy analizie wspomnianych przykładów wykonanych w arkuszu MS Excel.

Do publikacji dołączono ponadto trzy aplikacje wykonane w programie MS Excel:

Bonus 1: „Szeregi statystyczne" – aplikacja do grupowania i prezentacji danych.

Bonus 2: „Rozkłady prawdopodobieństwa" – pozwala błyskawicznie obliczyć prawdopodobieństwo dla zadanej wartości lub odwrotnie – dla wybranych rozkładów.

Bonus 3: „Chi-kwadrat" – wspomaga analizę współzależności danych jakościowych.

1. Charakterystyka etapów badania statystycznego

Badanie statystyczne to złożony proces składający się z kilku etapów. Poniższa tabela zawiera syntetyczne zestawienie podziału badań statystycznych na poszczególne etapy według wybranych autorów.

Tabela 1.1. Etapy badania statystycznego w świetle literatury przedmiotu.

Autorzy	Etapy badania statystycznego
A. Bielecka	1. Planowanie i organizacja badania. 2. Zbieranie danych statystycznych. 3. Opracowanie zebranego materiału statystycznego. 4. Analiza wyników badania.
A. Komosa, J. Musiałkiewicz	1. Przygotowanie badania. 2. Zebranie materiału statystycznego (danych statystycznych). 3. Przygotowanie materiału statystycznego do opracowania. 4. Opracowanie materiału statystycznego. 5. Prezentacja materiału statystycznego. 6. Analiza statystyczna – podstawa wyciągnięcia wniosków.
T. Michalski	1. Przygotowanie badania. 2. Zebranie materiału statystycznego i przygotowanie do opracowania. 3. Opracowanie materiału statystycznego. 4. Prezentacja danych statystycznych i analiza statystyczna.
J. Pociecha	1. Rozpoznanie i sformułowanie problemu. 2. Postawienie hipotezy i ustalenie możliwych rozwiązań. 3. Określenie źródeł informacji. 4. Przygotowanie do gromadzenia danych pierwotnych. 5. Gromadzenie danych. 6. Opracowanie danych i ich analiza. 7. Przygotowanie sprawozdania.
B. Pułaska-Turyna	1. Projektowanie badania. 2. Obserwacja statystyczna. 3. Opracowanie materiału statystycznego. 4. Analiza statystyczna.

M. Sobczyk	1. Przygotowanie (programowanie) badania. 2. Obserwacja statystyczna. 3. Opracowanie i prezentacja materiału statystycznego. 4. Opis lub wnioskowanie statystyczne.
W. Starzyńska	1. Przygotowanie lub programowanie badania statystycznego. 2. Obserwacja statystyczna. 3. Opracowanie surowego materiału statystycznego. 4. Analiza opracowanego materiału statystycznego.

Źródło: Opracowanie własne na podstawie: [3, s. 29], [7, s. 22], [10, s. 28], [14, s. 33], [15, s. 19-20], [19, s. 20], [21, s. 22].

W literaturze przedmiotu najczęściej wymienia się cztery podstawowe etapy badania statystycznego. Mimo pewnych rozbieżności w nazwach, można wymienić następujące podstawowe etapy:

1. Przygotowanie (planowanie, projektowanie, programowanie) badania.
2. Obserwacja statystyczna (zbieranie materiału statystycznego).
3. Opracowanie i prezentacja materiału statystycznego.
4. Analiza statystyczna (opis lub wnioskowanie statystyczne).

Bardziej szczegółową klasyfikację przedstawili A. Komosa i J. Musiałkiewicz [7, s. 22]. Autorzy ci wyodrębnili dodatkowy etap: „przygotowanie materiału statystycznego do opracowania" (np. T. Michalski włącza je do etapu drugiego) oraz oddzielny etap „prezentacja materiału statystycznego" – na ogół jest ona zaliczany do etapu trzeciego (T. Michalski wyjątkowo zalicza ją do ostatniego etapu, związanego z analizą danych [10, s. 28]).

Nieco odmienną klasyfikację etapów badania statystycznego (marketingowego) przedstawia J. Pociecha [14, s. 33]. Po pierwsze etap – szósty stanowi połączenie dwóch wyodrębnionych wcześniej (opracowanie materiału statystycznego i analiza danych). Po drugie – wyodrębniony przez tego autora etap piąty („gromadzenie danych") stanowi jedną z podstawowych czynności zaliczanych do etapu, jakim jest obserwacja statystyczna. Zatem

rozpisany został szczegółowo etap pierwszy, związany z przygotowaniem badania statystycznego (trzy pierwsze wymienione przez tego autora etapy).

W dalszej części tego rozdziału dokładniej scharakteryzowano cztery etapy badań statystycznych według podziału odpowiadającego klasyfikacji M. Sobczyka [19, s. 20]. Autor ten w ramach poszczególnych etapów wymienia następujące czynności:

Tabela 1.2. Czynności wchodzące w skład badania statystycznego w przekroju poszczególnych etapów.

Etap badania statystycznego	**Wykaz czynności wchodzących w skład danego etapu**
I Przygotowanie badania	1. Ustalenie celu badania statystycznego. 2. Określenie przedmiotu badania (zbiorowości i jednostki statystycznej). 3. Właściwe określenie jednostki sprawozdawczej (źródeł danych). 4. Decyzja co do metody badania (pełne czy częściowe).
II Obserwacja statystyczna	1. Ustalenie wartości cech ilościowych lub odmian cech jakościowych u wszystkich jednostek badanej zbiorowości (generalnej bądź próbnej). 2. Kontrola formalna i merytoryczna zebranych danych.
III Opracowanie i prezentacja materiału statystycznego	1. Grupowanie lub klasyfikacja. 2. Zliczanie danych. 3. Tabelaryczna prezentacja materiału statystycznego. 4. Graficzna prezentacja materiału statystycznego.
IV Analiza statystyczna	1. Opis statystyczny. 2. Wnioskowanie statystyczne (badanie częściowe – próba losowa).

Źródło: Opracowanie własne na podstawie: [19, s. 20-30].

2. Opis statystyczny

Opis statystyczny ma sumaryczny charakter, co oznacza, że dotyczy on całej zbiorowości generalnej bądź próbnej, a nie poszczególnych jednostek statystycznych. Opisu statystycznego dokonuje się za pomocą odpowiednich miar [19, s. 30]. W dalszej części tego rozdziału omówiono wybrane miary opisu statystycznego, stosowane w analizie struktury zbiorowości, analizie współzależności oraz analizie dynamiki. Rozdział ten ma zatem analityczny charakter i stanowi wstęp do wnioskowania statystycznego. Dlatego we wszystkich wzorach, gdzie pojawi się liczebność zbiorowości, będzie ona oznaczana literą *n* jako liczebność zbiorowości próbnej (niemniej jednak wzory te znajdują również zastosowanie przy obliczaniu charakterystyk dla całej populacji generalnej).

Tym, na co należy zwrócić uwagę przy studiowaniu niniejszego rozdziału – a o czym niejednokrotnie zdarza się zapominać na egzaminie – jest rodzaj danej cechy statystycznej i związany z nią typ skali pomiarowej. Jak już była mowa, pomiar cech ilościowych na skalach „słabszych" pociąga za sobą znaczną utratę informacji. Im silniejszy typ skali pomiarowej, tym więcej miar statystycznych można obliczyć (zob. tabela 1.5).

Ponadto – w przypadku cech ilościowych – wybór odpowiedniej miary (skorzystanie z prawidłowego wzoru statystycznego) zależy od tego, czy dane są pogrupowane, a jeśli tak, to czy pogrupowano je w szereg rozdzielczy punktowy, czy też szereg rozdzielczy z przedziałami klasowymi.

W związku z powyższym – przy prezentowaniu miar opisu statystycznego podkreślono, czy dany wzór znajduje zastosowanie dla danych niepogrupowanych, czy też pogrupowanych w szereg rozdzielczy (punktowy lub z przedziałami klasowymi). Zwrócono też uwagę na typ skali pomiaru danych, umożliwiający zastosowanie określonej miary.

1.1. Przygotowanie badania

Na tym etapie należy sprecyzować cel badania statystycznego, określić zbiorowość i jednostkę statystyczną, jak również dokonać wyboru metody badania. Jest to ważny etap, ponieważ popełnione tu błędy w dużym stopniu mogą zaważyć na jakości całego badania.

1.1.1. Ustalenie celu badania statystycznego

Na wstępie formułowane są koncepcje dotyczące całości badania statystycznego. Podstawową kwestią jest dokładne określenie celów (ogólnych i szczegółowych) oraz hipotez roboczych [10, s. 28]. A. Bielecka [3, s. 29] wyróżnia dwa zasadnicze cele badania statystycznego, tj.:

1. Cel diagnostyczny – określa, co i dlaczego jest przedmiotem badania statystycznego.
2. Cel praktyczny – precyzuje, komu i czemu badanie ma służyć.

Oto przykłady określenia celu diagnostycznego i praktycznego (por. [3, s. 30]):

Przykład 1. Celem diagnostycznym jest określenie skuteczności wybranych narzędzi marketingowych stosowanych w sprzedaży jogurtów w pewnym supermarkecie – badaniu poddano takie narzędzia, jak: promocje cenowe, degustacje, zamieszczenie oferty w gazetce reklamowej. Cel praktyczny takiego badania to zweryfikowanie hipotezy głoszącej, iż na wzrost popytu znacząco wpływa połączenie promocji cenowej z prezentacją promowanego jogurtu w gazetce reklamowej. Jeśli hipoteza ta okaże się słuszna, to w przyszłości dział marketingu supermarketu zawsze będzie stoso-

wał promocje cenowe dla tej grupy produktów, w połączeniu z wydrukiem oferty promocyjnej w gazetce reklamowej (efekt synergiczny).

Przykład 2. Firma zajmująca się pośrednictwem finansowym planuje wprowadzenie do oferty pośredniczenia w zawieraniu umów odnośnie zmiany Otwartego Funduszu Emerytalnego. Może jednak podpisać umowę wyłącznie z jednym funduszem. Celem diagnostycznym badania będzie określenie częstotliwości i kierunku zmian poszczególnych OFE przez zapisane już do nich osoby oraz identyfikacja kluczowych czynników powodujących te zmiany. Można postawić hipotezę, iż o zmianie OFE decydują głównie czynniki ekonomiczne, takie jak stopa zwrotu czy prowizja od składki. Gdy hipoteza ta okaże się słuszna, to firma podpisze umowę z funduszem o najwyższej stopie zwrotu netto, tj. stopie skorygowanej o koszty prowizji od składek. W przeciwnym razie należy określić czynniki pozaekonomiczne (np. podpisać umowę z funduszem gwarantującym najwyższą stawkę dla akwizytora od podpisanej umowy – czynnik ten może okazać się skutecznym motywatorem dla osób pozyskujących klientów dla danego OFE).

Przykład 3. Firma edukacyjna zamierza rozszerzyć swoją ofertę o nauczanie na odległość (tzw. *e-learning*). Celem diagnostycznym projektowanego badania statystycznego będzie określenie preferencji wśród wybranej grupy studentów odnośnie różnych form nauczania, w tym stosunku do nauczania na odległość. Ponadto celem diagnostycznym jest określenie najbardziej popularnych przedmiotów. Początkowo – z uwagi na znaczne koszty inwestycji w platformę e-learningową – planowane jest wprowadzenie tylko dwóch przedmiotów. Celem praktycznym będzie w tym przypadku zweryfikowanie hipotezy o dużym zainteresowaniu nauczaniem *on-line*, a w przypadku jej poprawności – optymalne dostosowanie oferty do rynku (wybór najbardziej popularnych przedmiotów).

Jak widać, cel diagnostyczny określa obecny stan rzeczy, natomiast cel praktyczny zmierza do wyciągnięcia wniosków i podjęcia odpowiednich kroków w przyszłości.

1.1.2. Określenie przedmiotu badania

Mając ustalone cele badania statystycznego oraz hipotezy robocze – można przejść do kolejnej czynności, jaką jest określenie zbiorowości i jednostki statystycznej.

Zbiorowość statystyczna – zwana też populacją statystyczną lub generalną – to „ogół osób, rzeczy bądź zjawisk będących przedmiotem badań statystycznych" [3, s. 15]. Oto przegląd klasyfikacji populacji statystycznych według wybranych kryteriów:

Tabela 1.3. Klasyfikacja zbiorowości statystycznych pod kątem wybranych kryteriów.

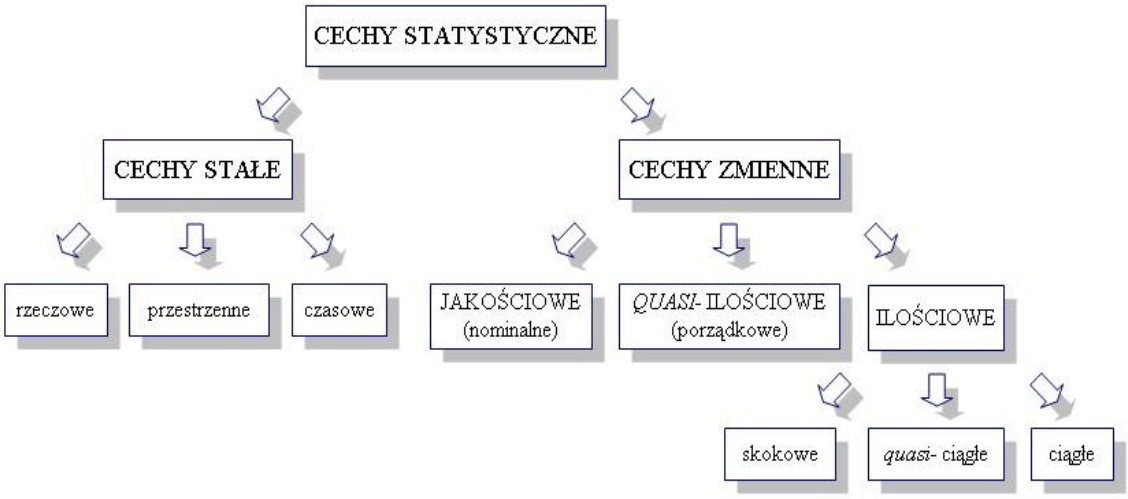
Kryterium klasyfikacji	Rodzaje zbiorowości statystycznych
I Kryterium jednorodności jednostek zbiorowości	1. Zbiorowość jednorodna – wszystkie jednostki są tego samego typu, rodzaju i gatunku. 2. Zbiorowość niejednorodna – jednostki różnią się cechami jakościowymi.
II Charakter jednostek zbiorowości	1. Zbiorowość statyczna – badanie na określony moment. 2. Zbiorowość dynamiczna – badanie w danym przedziale czasowym.
III Ilość badanych cech	1. Zbiorowość jednowymiarowa – badanie ze względu na jedną cechę. 2. Zbiorowość wielowymiarowa – badanie ze względu na wiele cech.
IV Liczba elementów zbiorowości	1. Zbiorowość skończenie liczna – ograniczona możliwa do określenia liczba jednostek. 2. Zbiorowość nieskończenie liczna – nieograniczona pod względem liczebności.
V Zasięg (zakres)	1. Zbiorowość całkowita (populacja generalna). 2. Zbiorowość próbna (próba).

Źródło: Opracowanie własne na podstawie: [2, s. 22-25].

Jednostka statystyczna – zwana też jednostką badania lub obserwacją – to „najmniejszy element zbiorowości statystycznej" [3, s. 15].

Wchodzące w skład badanej zbiorowości jednostki statystyczne odznaczają się pewnymi właściwościami, określanymi mianem **cech statystycznych** [19, s. 12]. Oto szczegółowa klasyfikacja cech statystycznych:

Rysunek 1.1. Klasyfikacja cech statystycznych.



Źródło: Opracowanie własne na podstawie: [2, s. 26-28], [3, s. 18].

Ogólnie rzecz biorąc, cechy statystyczne można podzielić na dwie grupy [21, s. 15]:

1. CECHY STAŁE – własności wspólne wszystkim jednostkom badanej zbiorowości statystycznej.
2. CECHY ZMIENNE – własności, dzięki którym poszczególne jednostki różnią się między sobą, przy czym dokładny stopień zmienności poszczególnych cech jest możliwy lub niemożliwy do określenia.

Cechy stałe służą do określenia jednostki statystycznej, a tym samym zbiorowości statystycznej, pod względem rzeczowym, przestrzennym i czasowym i nie podlegają badaniu statystycznemu (pełnią rolę „klasyfikatorów") [19, s. 12]. Zatem jednostką statystyczną jest „każdy element wchodzący w skład zbiorowości statystycznej i posiadający – tak jak wszystkie jednostki tej zbiorowości – tę samą lub te same cechy stałe" [2, s. 25]. Wyróżnia się następujące typy cech stałych [2, s. 26-27]:

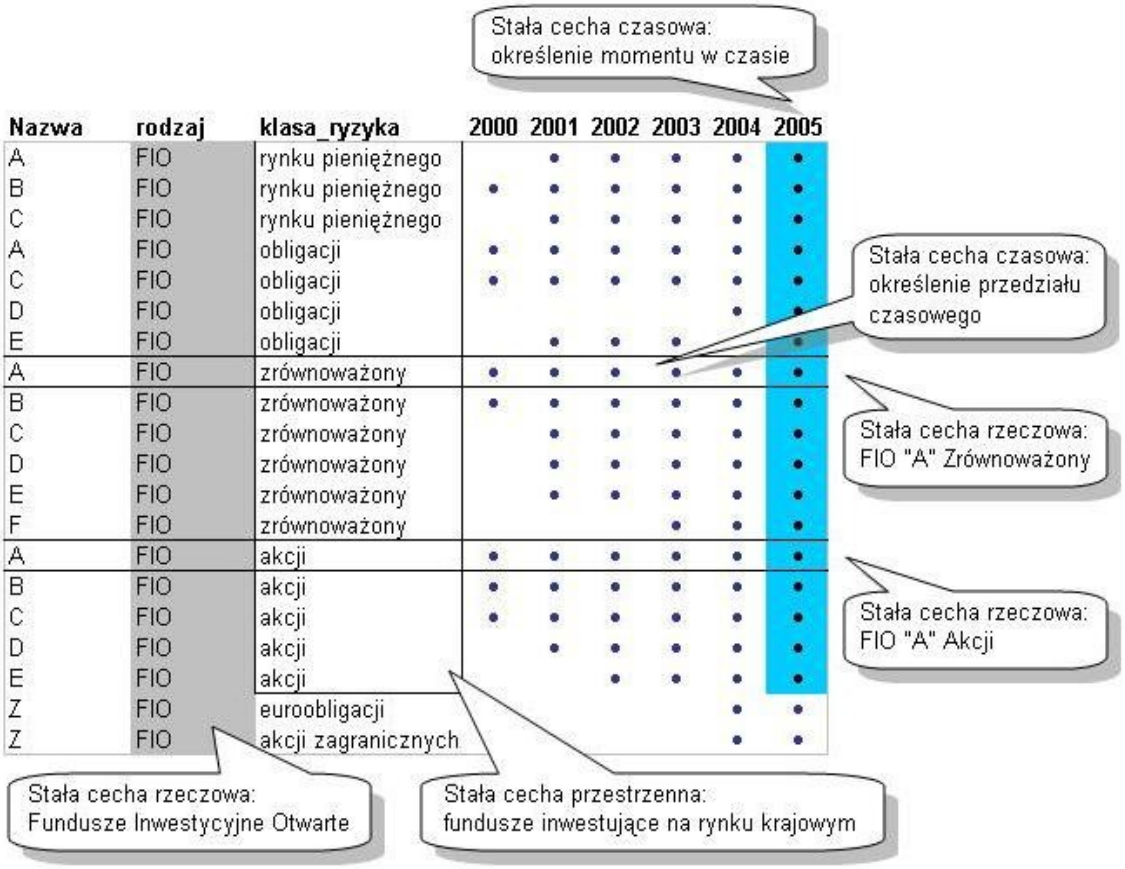
1. **Cechy rzeczowe** (przedmiotowe) – właściwości, którymi charakteryzuje się ściśle określony zbiór osób, rzeczy lub zjawisk. Cecha rzeczowa precyzuje, *kto* lub *co* jest przedmiotem badania statystycznego.
2. **Cechy przestrzenne** – informują o tym, z jakiego miejsca lub obszaru pochodzą jednostki włączone do badania statystycznego.
3. **Cechy czasowe** – określają, z jakiego okresu lub momentu włączono daną jednostkę w skład zbiorowości statystycznej.

M. Sobczyk podkreśla, iż w tej samej zbiorowości można wyodrębnić różne jednostki statystyczne [19, s. 12]. Wybór właściwej jednostki statystycznej zależy głównie od określonego celu badania statystycznego, co ukazują poniższe przykłady:

Przykład 1. Celem badania statystycznego jest określenie struktury liczby uczestników Otwartych Funduszy Inwestycyjnych (FIO), które inwestują powierzone środki na krajowym rynku papierów wartościowych. Raport ma dotyczyć stanu na koniec 2005 roku. Oto jak zostały określone cechy stałe (zob. rys. 1.1):

1. *Cecha rzeczowa* informuje, iż przedmiotem badania jest struktura liczby osób lokujących środki finansowe w Otwartych Funduszach Inwestycyjnych (FIO).
2. *Cecha przestrzenna* zawęża krąg analizy do polskich funduszy inwestujących w krajowe papiery wartościowe.
3. *Cecha czasowa* określa moment w czasie, czyli dane za rok 2005.

Rysunek 1.2. Przykład określenia zbiorowości i jednostek statystycznych według cech stałych.



Źródło: Opracowanie własne.

Z powyższego schematu wynika, iż jednostkami statystycznymi wchodzącymi w skład oznaczonej kolorem niebieskim populacji generalnej są poszczególne Fundusze Inwestycyjne Otwarte, lokujące powierzone środki

wyłącznie na rynku krajowym (stąd nie uwzględniono funduszu „Z") i prowadzące działalność w 2005 roku (nie uwzględniamy w analizie funduszy, które powstały w trakcie 2005 roku) – łącznie 18 jednostek statystycznych. W wyniku analizy statystycznej – zgodnie z celem tego badania – otrzyma się rozkład liczby uczestników FIO w zależności od klasy ryzyka funduszu (zob. miary natężenia i struktury).

Innym celem jest porównanie dynamiki liczby uczestników Funduszu „A" Zrównoważonego z Funduszem „A" Akcji w latach 2000-2005 (zob. analiza dynamiki). Celem praktycznym jest określenie zmian w preferencjach odnośnie tych dwóch funduszy i odpowiednie przygotowanie oferty promocyjnej. Porównywane będą dwie populacje:

1. Jako *cechę rzeczową* przyjęto odpowiednio FIO „A" Zrównoważony (pierwsza populacja) i FIO „A" Akcji (druga populacja).
2. W tym przypadku nie ma potrzeby określania *cechy przestrzennej*, ponieważ wybrane fundusze działają na określonym rynku.
3. *Cecha czasowa* jest wspólna dla obu porównywanych populacji – jest nią zakres czasowy określony na lata 2000-2005.

W tej sytuacji jednostką statystyczną (obserwacją) jest konkretny punkt danych w przekroju czasowym – liczba obserwacji jest równa liczbie lat objętych analizą. Należy zaznaczyć, iż możliwe jest porównywanie funduszy, które działają na rynku w określonym czasie (np. porównanie z FIO „E" Akcji ogranicza analizę do lat 2002-2005).

Przykład 2. Celem badania jest analiza dziennych zmian procentowych indeksu największych polskich spółek WIG 20 w określonym czasie:

1. *Cecha rzeczowa* określa przedmiot analizy, czyli procentowe dzienne zmiany indeksu WIG 20 (można dokonać porównań z innymi indeksami giełdowymi, np. WIG-iem).
2. *Cecha przestrzenna* precyzuje, iż chodzi o GPW w Warszawie.
3. *Cecha czasowa* określa liczbę sesji giełdowych (np. 50 ostatnich sesji).

W tej sytuacji jednostką statystyczną jest sesja giełdowa. Celem analizy może być także ustalenie, jakie spółki w danym dniu wpłynęły pozytywnie na poziom badanego indeksu. Należy wyjaśnić, iż indeks ten jest wypadkową zmian kursów akcji 20 największych spółek wchodzących w jego skład. Oto określenie cech stałych:

1. *Cecha rzeczowa* – procentowe dzienne zmiany kursów akcji spółek WIG 20.
2. *Cecha przestrzenna* – GPW w Warszawie.
3. *Cecha czasowa* – określenie sesji giełdowej (np. ostatnia sesja).

W tej sytuacji jednostką statystyczną nie będzie już sesja giełdowa, lecz spółka zaliczana do indeksu WIG 20. Nietrudno zauważyć, iż istnieje dwadzieścia jednostek statystycznych (w skład WIG 20 wchodzi bowiem dwadzieścia spółek).

Przykład 3. Celem badania statystycznego jest analiza wyników egzaminu ze statystyki w semestrze letnim roku akademickiego 2005/2006 na studiach dziennych uczelni państwowych. Populację generalną określono pod względem cech stałych następująco:

1. *Cecha rzeczowa* – studenci studiów dziennych uczelni państwowych, którzy w semestrze letnim przystąpili do egzaminu ze statystyki (możliwe porównanie ze studiami wieczorowymi i zaocznymi).

2. *Cecha przestrzenna* – osoby studiujące na terytorium RP (wyniki można porównać np. z innymi krajami Unii Europejskiej).
3. *Cecha czasowa* – semestr letni roku akademickiego 2005/2006 (wyniki analizy można np. porównać z analogicznym okresem roku poprzedniego).

Jednostki statystyczne w tym przypadku tworzą studenci studiów dziennych polskich uczelni państwowych, którzy w semestrze letnim w roku akademickim 2005/2006 przystąpili do egzaminu ze statystyki.

Druga grupa cech statystycznych to **cechy zmienne** – podlegają one badaniu statystycznemu [19, s. 12]. Należą do nich trzy kategorie cech, a mianowicie (zob. rys. 1.1):

1. **Cecha jakościowa** (nominalna) to „niemierzalna właściwość, której konkretny wariant występuje lub nie występuje w danej zbiorowości i nie dając wyrażać się liczbowo, daje się opisać jedynie za pomocą określeń słownych” [2, s. 28]. Wariantów cech nominalnych (zob. skala nominalna) nie da się uporządkować (por. [20, s. 22]).
2. **Cecha *quasi*-ilościowa** (*niby*-ilościowa, porządkowa) to „właściwość, która określa natężenie badanej cechy u poszczególnych jednostek danej zbiorowości w sposób opisowy” [2, s. 28]. Warianty cech porządkowych (zob. skala porządkowa) – w przeciwieństwie do wariantów cech nominalnych – można uporządkować (por. [20, s. 22]). Cechy porządkowe – w bardziej ogólnej klasyfikacji – zaliczane są do *cech jakościowych*. Istotne jest to, iż warianty cech jakościowych wyrażone są za pomocą określeń słownych (werbalnych). Przypisywane niekiedy cechom jakościowym (nominalnym lub porządkowym) liczby nie wyrażają bowiem ich wartości – pełnią jedynie rolę „etykiet” (por. [3, s. 18]). Przyjęta w niniejszej publikacji szczegółowa klasyfikacja cech statystycznych – wyodrębniająca cechy *quasi*-ilościowe – ma za zadanie ułatwienie doboru skal pomiarowych w zależności od rodzaju cechy statystycznej.

3. **Cecha ilościowa** to „mierzalna właściwość, występująca z określonym natężeniem u wszystkich jednostek zbiorowości statystycznej" [2, s. 27]. Właściwości cech ilościowych – określanych też mianem *cech mierzalnych* – można mierzyć za pomocą liczb mianowanych typu: metry, kilogramy, sztuki, lata, jednostki pieniężne, czas itp. (por. skala przedziałowa i skala ilorazowa). Do cech ilościowych należą [3, s. 18]:
   - ***cecha skokowa*** – warianty tej cechy wyrażone są za pomocą liczb należących do zbioru przeliczalnego lub skończonego (typową jednostką miary są sztuki/liczby naturalne),
   - ***cecha quasi-ciągła*** (*niby*-ciągła) – cecha ze swej natury skokowa, ale z uwagi na bardzo dużą liczbę przyjmowanych wartości liczbowych traktowana jako cecha ciągła. Różnica między kolejnymi wartościami liczbowymi jest niewielka (np. ceny wyrażone z dokładnością do jednego grosza).
   - ***cecha ciągła*** – cecha, której warianty wyrażone są za pomocą liczb rzeczywistych, gdzie pomiędzy dwiema dowolnymi wartościami liczbowymi danej cechy można teoretycznie zawsze znaleźć wartość pośrednią cechy (typowymi jednostkami miary cech ciągłych są m.in.: czas, metry, kilogramy, wiek).

Należy podkreślić, iż warunkiem zaklasyfikowania danej cechy do *cech skokowych* nie jest fakt, iż jej warianty występują w postaci liczb całkowitych. Przykładem mogą być oceny z egzaminu: 3; 3,5 (3+); 4; 4,5 (4+); 5. Mimo że cecha ta nie przyjmuje wyłącznie liczb całkowitych (np. tak jak miałoby to miejsce w przypadku liczby nieobecności w szkole), to – z uwagi na niewielką liczbę możliwych wariantów – jest ona cechą skokową.

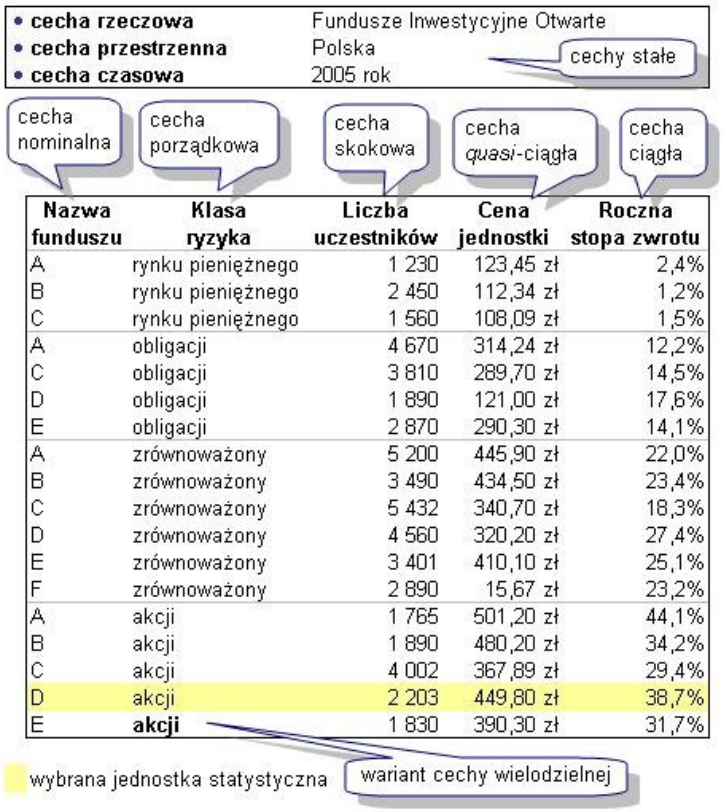
Przy charakterystyce cech statystycznych kilkakrotnie pojawiło się pojęcie *wariantu cechy*. **Wariant cechy statystycznej** jest „informacją uzyskaną o jednostce statystycznej w trakcie badania statystycznego" [7, s. 10]. Z uwagi na liczbę możliwych wariantów, cechy statystyczne dzieli się na [20, s. 22]:

- ***cechy dychotomiczne*** (*zero-jedynkowe*) – cecha może przyjąć tylko dwa warianty.
- ***cechy wielodzielne*** (*politomiczne*) – przyjmują więcej niż dwa warianty.

Liczba wariantów danej cechy może być co najwyżej równa liczbie jednostek wchodzących w skład określonej zbiorowości statystycznej – jest to możliwe w przypadku cech ciągłych. Zazwyczaj jednak liczba wariantów jest mniejsza od liczby jednostek, ponieważ identyczny wariant cechy może występować u kilku jednostek statystycznych (por. [19, s. 13]). Oto przykłady identyfikacji rodzaju cech statystycznych (zmiennych):

Przykład 1. Nawiązując do prezentowanego wcześniej przykładu z Funduszami Inwestycyjnymi Otwartymi (zob. rys. 1.2), należy ustalić – po określeniu jednostki i zbiorowości statystycznej – typy cech statystycznych. Przykład ilustruje rys. 1.3:

Rysunek 1.3. Przykłady cech statystycznych.



Nazwa funduszu	Klasa ryzyka	Liczba uczestników	Cena jednostki	Roczna stopa zwrotu
A	rynku pieniężnego	1 230	123,45 zł	2,4%
B	rynku pieniężnego	2 450	112,34 zł	1,2%
C	rynku pieniężnego	1 560	108,09 zł	1,5%
A	obligacji	4 670	314,24 zł	12,2%
C	obligacji	3 810	289,70 zł	14,5%
D	obligacji	1 890	121,00 zł	17,6%
E	obligacji	2 870	290,30 zł	14,1%
A	zrównoważony	5 200	445,90 zł	22,0%
B	zrównoważony	3 490	434,50 zł	23,4%
C	zrównoważony	5 432	340,70 zł	18,3%
D	zrównoważony	4 560	320,20 zł	27,4%
E	zrównoważony	3 401	410,10 zł	25,1%
F	zrównoważony	2 890	15,67 zł	23,2%
A	akcji	1 765	501,20 zł	44,1%
B	akcji	1 890	480,20 zł	34,2%
C	akcji	4 002	367,89 zł	29,4%
D	akcji	2 203	449,80 zł	38,7%
E	**akcji**	1 830	390,30 zł	31,7%

Źródło: Opracowanie własne (dane umowne).

Zbiorowość statystyczna została określona pod względem rzeczowym (co jest przedmiotem badania), przestrzennym (teren badania) oraz czasowym (moment badania określony na 2005 rok). Tak określona zbiorowość składa się z 18 jednostek statystycznych, którymi są poszczególne Fundusze Inwestycyjne Otwarte lokujące środki finansowe na krajowym rynku w 2005 roku. Wybraną jednostkę statystyczną zaznaczono żółtym kolorem. Każda jednostka posiada szereg właściwości, czyli zmiennych cech statystycznych. Dwie pierwsze, „Nazwa funduszu" i „Klasa ryzyka", mają jakościowy charakter, ponieważ ich warianty dają się opisać w sposób słowny. Pogrubionym kolorem zaznaczono jeden z wariantów cechy „Klasa ryzyka" – cecha ta jest cechą *quasi*-ilościową (porządkową), ponieważ jej warianty można uporządkować pod kątem stopnia ryzyka (niemniej jednak w innych analizach, gdzie ryzyko nie ma znaczenia, cecha ta jest cechą nominalną). „Stopa zwrotu" nie jest cechą *quasi*-ciągłą, ponieważ teoretycznie można ją wyznaczyć z nieskończenie dużą precyzją – jest to iloraz ceny jednostki uczestnictwa z końca do ceny z początku 2005 roku. Natomiast ceny z definicji podaje się z dokładnością do 1 grosza.

Przykład 2. Celem badania statystycznego jest analiza rynku mieszkań w tzw. standardzie deweloperskim w Polsce. Oto zestaw cech statystycznych branych pod uwagę:

1. Nazwa województwa – cecha jakościowa nominalna.
2. Ilość pokoi – cecha ilościowa skokowa.
3. Cena mieszkania (zł/m^2) – cecha ilościowa *quasi*-ciągła.

Przykład 3. Przedmiotem badania statystycznego jest określenie czynników wpływających na wyniki egzaminu ze statystyki. Jako cechę zależną przyjęto liczbę punktów uzyskanych na egzaminie (cecha ilościowa *quasi*-

ciągła – punkty mierzone w skali od zera do 100 z dokładnością do 0,1). Oto zestaw zmiennych objaśniających:

1. Liczba nieobecności na zajęciach – cecha ilościowa skokowa.
2. Przeciętna liczba godzin poświęconych nauce statystyki tygodniowo – jw.
3. Preferencje co do przedmiotu *statystyka* (nudny, ciekawy) – cecha porządkowa.
4. Płeć studenta – cecha jakościowa (nominalna).

Reasumując, zbiorowość statystyczną tworzą poszczególne jednostki statystyczne, posiadające określone cechy statystyczne. O ile cechy stałe – wspólne wszystkim jednostkom badania statystycznego – służą do określenia zbiorowości, o tyle cechy zmienne podlegają badaniu. Należy ustalić, czy będzie ono obejmowało wszystkie jednostki, czy tylko wybrane z nich, a następnie dokonać wyboru adekwatnej metody badania.

1.1.3. Wybór metody badania statystycznego

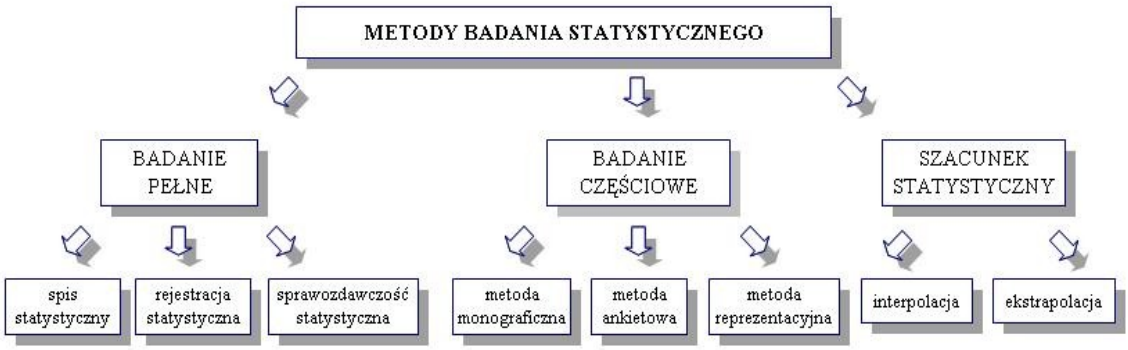
Kolejną czynnością w fazie wstępnej jest określenie *metody badania statystycznego*. Wybór metody zależy od takich czynników, jak (por. [19, s. 16]):

- cel badania statystycznego,
- rodzaj zbiorowości statystycznej,
- stopień szczegółowości badania,
- ilość dostępnych środków finansowych,
- stosowane metody analizy (opis lub wnioskowanie statystyczne).

Badanie statystyczne obejmuje wszystkie jednostki statystyczne lub tylko wybrane z nich, czyli *próbę*. **Próba** to pewien podzbiór populacji generalnej, którego elementy zostały dobrane w sposób losowy bądź nielosowy (por. [20, s. 20]). Innymi słowy: próba to „liczebność jednostek badania" [5, s. 19].

Klasyfikacja metod badania statystycznego – ze względu na liczbę jednostek objętych badaniem – przedstawia się następująco:

Rysunek 1.4. Klasyfikacja metod badań statystycznych ze względu na liczbę jednostek objętych badaniem.



Źródło: Opracowanie na podstawie: [7, s. 14].

Ogólnie rzecz biorąc, można wyodrębnić trzy grupy metod badania statystycznego:

1. BADANIE PEŁNE (całkowite, wyczerpujące) – polega na tym, że informacje o badanych cechach statystycznych są gromadzone od wszystkich jednostek statystycznych wchodzących w skład zbiorowości statystycznej [7, s. 15].
2. BADANIE CZĘŚCIOWE (niepełne, fragmentaryczne) – obejmuje wybrane jednostki zbiorowości statystycznej [19, s. 16].

3. SZACUNEK STATYSTYCZNY (szacunek wartości) – interpolacyjny lub ekstrapolacyjny szacunek statystyczny zaliczany jest niekiedy w literaturze przedmiotu do metod badania częściowego (zob. [3, s. 32]):
   - *interpolacja* polega na znajdowaniu nieznanych wartości funkcji w dowolnym punkcie przedziału (x_1, x_n) na podstawie dostępnych wartości funkcji, należących do tego przedziału (np. ustalanie wartości *kwartyli*).
   - *ekstrapolacja* polega na ustaleniu nieznanych wartości funkcji w dowolnym punkcie leżącym poza przedziałem wartości posiadanych: x_{n+1}, x_{n+i} (np. prognozowanie).

Do **metod badania pełnego** należą (zob. [7, s. 15-18]):

1. **Spis statystyczny** jest to badanie polegające na zbieraniu informacji o wartościach cechy statystycznej bezpośrednio od wszystkich jednostek tworzących zbiorowość statystyczną. Informacje te są zbierane przez specjalnie do tego celu przeszkolone osoby (rachmistrzów spisowych). Jednocześnie informacje te są utrwalane na formularzach spisowych, przygotowanych przez instytucję organizującą spis. Rachmistrze spisowi dokonują zatem bezpośredniej obserwacji statystycznej. Spisy statystyczne dostarczają szczegółowych informacji o badanej zbiorowości. Ze względu na bardzo wysokie koszty omawiana metoda znajduje zastosowanie w badaniach najważniejszych zjawisk społeczno-gospodarczych (np. Narodowy Spis Powszechny Ludności i Mieszkań z 2002 roku przeprowadzony przez Główny Urząd Statystyczny).
2. **Rejestracja statystyczna** polega na wpisywaniu zdarzeń i faktów do odpowiednich rejestrów. Rejestracja statystyczna ma węższy zakres tematyczny aniżeli spis statystyczny. Ponadto różni się ona od niego sposobem gromadzenia informacji – przy rejestracji statystycznej nie występuje bezpośrednia obserwacja statystyczna, lecz informacje będące przedmiotem rejestracji są zgłaszane w punktach rejestracyjnych. Wyróżnia się:

- *doraźną rejestrację statystyczną* – polega ona na tym, że w wyznaczonym czasie określone osoby zgłaszają się w wyznaczonych miejscach i udzielają informacji objętej tematyką rejestracji (np. ewidencja działalności gospodarczej),
- *bieżącą rejestrację statystyczną* – polega ona na ciągłym, bieżącym, systematycznym notowaniu zdarzeń i faktów określonych przez instytucję prowadzącą rejestrację (np. ewidencja ludności).

3. **Sprawozdawczość statystyczna** to najbardziej powszechny rodzaj pełnych badań statystycznych – polega na przekazywaniu przez jednostki sprawozdawcze określonych informacji liczbowych i opisowych w postaci standardowych sprawozdań. Instytucja organizująca badanie statystyczne powinna opracować odpowiednie formularze statystyczne wraz z instrukcjami ich wypełniania, jak również określić termin ich przekazywania (jako przykład można podać opracowane dla celów podatkowych formularze PIT adresowane do osób fizycznych czy też formularze ZUS wypełniane przez przedsiębiorców).

Zbiorowości statystycznej nie można poddać *badaniu pełnemu* w takich sytuacjach, jak (por. [2, s. 23], [3, s. 31-32]):

- badany element ulega zniszczeniu (badanie pełne oznaczałoby w tej sytuacji zniszczenie wszystkich elementów),
- badanie pełne jest zbyt kosztowne (np. z uwagi na dużą populację generalną),
- badanie pełne jest zbyt czasochłonne (np. duża dynamika zmian badanego zjawiska wymaga podjęcia szybkich decyzji),
- badana zbiorowość jest nieskończenie duża (w praktyce za taką populację można też uznać bardzo liczne populacje, np. liczbę potencjalnych internautów – w tej sytuacji można mówić wyłącznie o badaniu częściowym).

W powyższych sytuacjach odpowiednim badaniem jest badanie częściowe. W literaturze statystycznej wymienia się następujące **metody badania częściowego**:

1. **Metoda monograficzna** polega na wszechstronnym opisie i szczegółowej analizie pojedynczej jednostki statystycznej lub niewielkiej liczby charakterystycznych (typowych) jednostek badanej zbiorowości. Dzięki niewielkiej grupie jednostek można w badaniu uwzględnić stosunkowo dużą liczbę cech statystycznych (zob. cechy zmienne). Podstawowe znaczenie w tej metodzie ma opis w oparciu o dane liczbowe [10, s. 25]. Przykładem może być opis wybranej placówki wychowawczo-oświatowej.
2. **Metoda ankietowa** polega na tym, że podmiot organizujący badanie zwraca się do określonej grupy osób (respondentów) z zaproszeniem do dobrowolnego wypowiedzenia się w określonej sprawie. Zaproszenie to może mieć charakter *powszechny* (ankieta kierowana do szerokiego grona osób, np. za pośrednictwem Internetu) lub *selektywny* (ankieta kierowana do wąskiej grupy respondentów, np. za pośrednictwem prasy specjalistycznej). Z uwagi na fakt, iż ankieta wypełniana jest przez respondenta, powinna być ona zredagowana w taki sposób, aby każdy ankietowany jednoznacznie rozumiał stawiane mu pytania i potrafił udzielić na nie odpowiedzi [7, s. 19-20] (zob. Gromadzenie danych ze źródeł pierwotnych).
3. **Metoda reprezentacyjna** opiera się na próbie pobranej ze zbiorowości generalnej w sposób losowy. Z teoretycznego i praktycznego punktu widzenia metoda ta jest najbardziej prawidłową formą badania częściowego. Zastosowanie rachunku prawdopodobieństwa przy uogólnianiu wyników z próby losowej na całą zbiorowość (zob. wnioskowanie statystyczne) pozwala na określenie wielkości popełnianego błędu. Możliwości tej nie stwarzają pozostałe metody badania częściowego, tj. metoda monograficzna i ankietowa [19, s. 17-18].

Przyjmując jako kryterium klasyfikacji *częstotliwość przeprowadzania badania statystycznego,* można wyróżnić trzy rodzaje badań statystycznych [7, s. 15]:

1. **Badania doraźne** (sporadyczne, jednorazowe, *ad hoc*) – są prowadzone wówczas, gdy zapotrzebowanie na określony rodzaj informacji pojawia się bardzo rzadko i jest spowodowane nieprzewidzianymi przyczynami (np. badanie preferencji nabywców danego produktu).
2. **Badania okresowe** są badaniami powtarzalnymi, które przeprowadza się w określonych momentach (np. publikowany na koniec każdego kwartału ranking Otwartych Funduszy Emerytalnych).
3. **Badania ciągłe** polegają na tym, że obserwacja i rejestracja określonych zdarzeń i faktów odbywa się w sposób ciągły. Badania ciągłe dotyczą jedynie niektórych, ściśle określonych faktów i zdarzeń (np. analiza procesu produkcyjnego pod względem jakości – konstrukcja tzw. *kart kontrolnych*).

W wypadku podjęcia decyzji o wyborze metody badania częściowego pojawia się kwestia **doboru próby**. Z uwagi na złożony charakter tego zagadnienia – metody doboru próby omówiono w ostatnim rozdziale (zob. Dobór próby). W tym miejscu warto podkreślić, iż w przypadku metody reprezentacyjnej dobór próby powinien być wyłącznie losowy.

1.2. Obserwacja statystyczna

Po ustaleniu celu badania statystycznego (diagnostycznego i praktycznego), określeniu zbiorowości i jednostki statystycznej (pod względem rzeczowym, przestrzennym i czasowym), jak również dokonaniu wyboru odpowiedniej metody badania (pełnego lub częściowego) – można przystąpić do drugiego etapu, jakim jest *obserwacja statystyczna*.

Ogólnie rzecz biorąc, **metody pozyskiwania danych** można podzielić na dwie grupy (por. [19, s. 20], [21, s. 20]):

1. Metody korzystania z publikowanych źródeł informacji (odpłatne lub nieodpłatne pozyskiwanie informacji od jednostek sprawozdawczych).
2. Metody przeprowadzania własnego badania statystycznego (zob. gromadzenie informacji ze źródeł pierwotnych).

Zebrane w wyniku obserwacji statystycznej dane określa się mianem **materiału statystycznego** [19, s. 20], przy czym – w zależności od przyjętej metody gromadzenia danych – rozróżnia się [10, s. 32]:

1. ***Materiał statystyczny pierwotny*** – informacje do prowadzenia danego badania statystycznego uzyskiwane są drogą odrębnego badania. Informacje te pochodzą z tzw. *źródeł pierwotnych* w wyniku pomiaru bezpośredniego (zob. kwestionariusz).
2. ***Materiał statystyczny wtórny*** – materiał zaczerpnięty spoza statystycznych źródeł, zwanych *źródłami wtórnymi*, który został wykorzystany w badaniach statystycznych.

Wybrane wtórne źródła danych znajdują się w pliku *dane_do_analizy.xls*, stanowiącym integralną część niniejszego opracowania. Plik ten zawiera wybrane dane finansowe i dane społeczno-gospodarcze. Poniżej przedstawiono przykłady wtórnych źródeł informacji:

Przykład 1. Jednostką sprawozdawczą dostarczającą co kwartał informacji o trzyletnich stopach zwrotu Otwartych Funduszy Emerytalnych jest Komisja Nadzoru Ubezpieczeń i Funduszy Emerytalnych (http://www.knuife-.gov.pl/).

Przykład 2. Spółki notowane na Giełdzie Papierów Wartościowych w Warszawie (http://www.gpw.pl) mają obowiązek sporządzania okresowych raportów finansowych.

Przykład 3. Jednostką sprawozdawczą prezentującą m.in. poziom stóp procentowych jest Narodowy Bank Polski (http://www.nbp.pl).

Przykład 4. Instytucją prezentującą dane o przestępczości w Polsce jest Komenda Główna Policji (http://www.kgp.gov.pl).

W tym miejscu warto zwrócić uwagę na szereg zniekształceń rzeczywistości, wynikających z błędnej interpretacji oficjalnych informacji pochodzących właśnie ze źródeł wtórnych. Oto następujące sytuacje:

Sytuacja 1. Oficjalny ranking najlepiej sprzedających się płyt CD (np. z oprogramowaniem edukacyjnym) nie musi odzwierciedlać nawet kolejności miejsc w rankingu. Dzieje się tak za sprawą „drugiego" – nieoficjalnego – obrotu nielegalnym oprogramowaniem, w wyniku czego ustalenie najbardziej popularnych programów komputerowych wymaga przeprowadzenia odrębnych badań wśród wybranej grupy respondentów (anonimowość ankiety sprzyja zakreślaniu odpowiedzi, jaki program ostatnio kupił ankietowany – nie wnika się przy tym, z jakiego źródła on pochodzi).

Sytuacja 2. Ustalenie faktycznej liczby rozwiedzionych rodzin jest praktycznie niemożliwe w oparciu o dane ze źródeł wtórnych – wiadomo bowiem, iż część rodzin rozwodzi się fikcyjnie („na papierze") w celu otrzymania zasiłku dla matki samotnie wychowującej dziecko. W tym przypadku wiarygodnych informacji mogłaby dostarczyć anonimowa ankieta.

Sytuacja 3. Kwestią kłopotliwą jest określenie skali ruchu turystycznego w pewnej nadmorskiej miejscowości w oparciu o wpływy z podatku klimatycznego (np. 1 zł za dobę). Takie informacje nie uwzględniają osób, które specjalnie przyjeżdżają na jeden dzień do tej miejscowości (np. na organi-

zowany koncert), czy też turystów znajdujących zakwaterowanie bez rejestracji i tym samym niepłacących podatku klimatycznego.

Ponadto należy pamiętać, iż źródła wtórne niekiedy dostarczają tylko pobieżnych informacji. I tak śledząc dostępne statystyki odwiedzin pewnego portalu internetowego można dowiedzieć się, ile procent odwiedzających to kobiety, jaka jest struktura wiekowa itp. Niestety, takie zbiorcze informacje nie pozwalają na określenie zależności np. pomiędzy wiekiem a płcią osób odwiedzających portal – tu konieczne jest dotarcie do danych niepogrupowanych.

Powyższe przykłady pokazują, iż mimo bogactwa informacji pochodzących ze źródeł wtórnych, niekiedy niezbędne jest dotarcie do informacji pochodzących ze źródeł pierwotnych. W kolejnym podrozdziale dokładniej omówiono organizację **własnego badania statystycznego** (gromadzenie informacji ze źródeł pierwotnych).

Tym, na co należy zwrócić uwagę przy studiowaniu niniejszego rozdziału – a o czym niejednokrotnie zdarza się zapominać na egzaminie – jest rodzaj danej cechy statystycznej i związany z nią typ skali pomiarowej. Jak już była mowa, pomiar cech ilościowych na skalach „słabszych" pociąga za sobą znaczną utratę informacji. Im silniejszy typ skali pomiarowej, tym więcej miar statystycznych można obliczyć (zob. tabela 1.5).

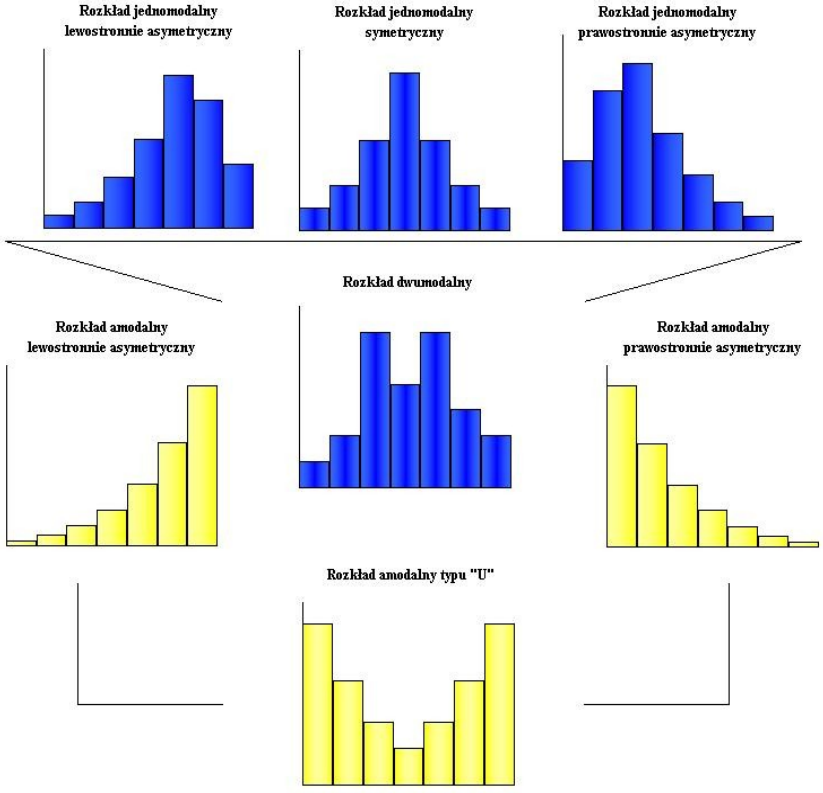
Ponadto – w przypadku cech ilościowych – wybór odpowiedniej miary (skorzystanie z prawidłowego wzoru statystycznego) zależy od tego, czy dane są pogrupowane, a jeśli tak, to czy pogrupowano je w szereg rozdzielczy punktowy, czy też szereg rozdzielczy z przedziałami klasowymi.

W związku z powyższym – przy prezentowaniu miar opisu statystycznego podkreślono, czy dany wzór znajduje zastosowanie dla danych niepogrupowanych, czy też pogrupowanych w szereg rozdzielczy (punktowy lub z przedziałami klasowymi). Zwrócono też uwagę na typ skali pomiaru danych, umożliwiający zastosowanie określonej miary.

2.1. Opis struktury zbiorowości

Przedstawiona w poprzednim rozdziale graficzna prezentacja materiału statystycznego z wykorzystaniem wykresów ukazujących strukturę badanej zbiorowości (zob. wykresy strukturalne) pozwala na wstępną ocenę empirycznego rozkładu zbiorowości ze względu na daną cechę statystyczną. W tym miejscu warto usystematyzować możliwe rozkłady empiryczne. Można je bowiem sklasyfikować w zależności od siły i kierunku ewentualnej asymetrii, jak również z punktu widzenia ilości ośrodków dominujących.

Rysunek 2.1. Typologia rozkładów empirycznych cechy ciągłej.



Źródło: Opracowanie na podstawie: [9, s. 65].

Szczególne miejsce wśród rozkładów cech zajmuje rozkład normalny, należący do klasy rozkładów jednomodalnych symetrycznych. Jednak w prak-

tyce empiryczne rozkłady cech są na ogół bardziej „smukłe” bądź bardziej „spłaszczone” aniżeli teoretyczny rozkład normalny (zob. eksces). Można tu zatem mówić o pewnym stopniu dopasowania danych empirycznych do rozkładu normalnego (zob. Hipotezy nieparametryczne).

Rozkłady cechy są w różnym stopniu lewo- bądź prawostronnie asymetryczne. O sile i kierunku asymetrii informują miary asymetrii. Z uwagi na siłę asymetrii rozróżnia się rozkłady umiarkowanie asymetryczne (jeden ośrodek dominujący) bądź rozkłady skrajnie asymetryczne (amodalne). Rozkłady skrajnie asymetryczne to takie, „w których prawie wszystkie jednostki mają niskie bądź wysokie wartości cechy” [19, s. 33]. Rozkłady typu „U” – zwane też siodłowymi – stanowią niejako złożenie rozkładu lewo- i prawostronnie asymetrycznego (w tym przypadku zamiast o wartości dominującej można mówić o tzw. „antymodzie”, tj. wartości będącej przeciwieństwem dominanty).

Rozkłady dwumodalne (bimodalne) posiadają dwa wyraźnie widoczne ośrodki dominujące, przy czym żaden z nich nie skupia wartości skrajnych (por. rozkład siodłowy). Przykładem takiego rozkładu może być rozkład częstości kursowania autobusów komunikacji miejskiej (ośrodkami dominującymi są godziny porannego i popołudniowego szczytu). Analogicznie można wyznaczyć rozkład trimodalny (trzy ośrodki dominujące) oraz – uogólniając – rozkłady wielomodalne (są to raczej teoretyczne przypadki).

Istnieje szereg miar statystycznych, służących do opisu zbiorowości statystycznej. Dlatego w literaturze przedmiotu zwykle klasyfikuje się je z punktu widzenia dwóch następujących kryteriów (por. [3, s. 96]):

Pierwszy – podział miar ze względu na zakres danych niezbędnych do ich wyznaczenia:

- *miary klasyczne*, do wyliczenia których niezbędne są wszystkie jednostki objęte badaniem statystycznym,

- *miary pozycyjne*, dla wyznaczenia których potrzebne są tylko wybrane obserwacje ze względu na zajmowaną pozycję w uporządkowanym zbiorze danych.

Ten podział miar statystycznych ma swoje implikacje w praktyce. Np. w przypadku danych pogrupowanych w szereg rozdzielczy klasowy z otwartym dolnym lub górnym przedziałem klasowym – zastosowanie znajdują miary pozycyjne.

Drugi podział pozwala na klasyfikację miar ze względu na rodzaj informacji, jakie one wnoszą o empirycznym rozkładzie cechy statystycznej. I tak wyróżnia się tu (por. [19, s. 35]):

1. *Miary położenia* (średnie, przeciętne) – służą do określenia wartości cechy, wokół której skupiają się wszystkie pozostałe wartości tej cechy.
2. *Miary dyspersji* (zmienności, rozproszenia) – badają stopień zróżnicowania wartości cechy, w tym wokół miar średnich.
3. *Miary asymetrii* (skośności) – służą do badania kierunku i siły ewentualnej asymetrii rozkładu zbiorowości ze względu na daną cechę statystyczną.
4. *Miary koncentracji* – pozwalają określić stopień koncentracji wokół wartości średniej, jak również ustalić stopień koncentracji jednostek statystycznych ze względu na wartości badanej cechy (np. koncentracja wysokości wynagrodzeń, obrotów ze sprzedaży itp.).

Poniżej przedstawiono typologię miar statystycznych według obu przedstawionych klasyfikacji:

Tabela 2.1. Typologia miar opisu statystycznego.

Zakres zastosowań	Miary klasyczne	Miary pozycyjne
Miary położenia	średnia arytmetyczna, średnia harmoniczna	mediana, kwartyle, percentyle, dominanta,
Miary dyspersji	wariancja, odchylenie standardowe/przeciętne, współczynnik zmienności klasyczny, typowy obszar zmienności	rozstęp, odchylenie ćwiartkowe, współczynnik zmienności pozycyjny, typowy obszar zmienności
Miary asymetrii	współczynnik asymetrii klasyczny	współczynnik asymetrii pozycyjny
	mieszany współczynnik asymetrii	
Miary koncentracji	eksces, współczynnik koncentracji Lorenza	–

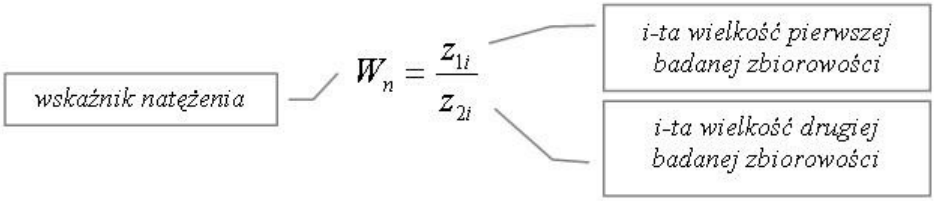
Źródło: Opracowanie na podstawie: [9, s. 54].

Kolejne podrozdziały odpowiadają klasyfikacji miar statystycznych ze względu na informacje, jakich wyznaczone charakterystyki dostarczają o rozkładzie empirycznym badanej cechy.

2.1.1. Miary natężenia i struktury

Miarą natężenia jest *wskaźnik natężenia*, zaś struktury *wskaźnik struktury*. Obie te miary odzwierciedlają zależności, proporcje i relacje występujące pomiędzy liczbami absolutnymi [2, s. 72].

Wskaźnik natężenia (*współczynnik natężenia*) to „wzajemny stosunek liczebności dwóch zbiorowości pozostających w logicznej zależności” [2, s. 72]. Wartość wskaźnika natężenia wyznacza się według wzoru:



Współczynnik natężenia jest wielkością mianowaną – określa on liczbę jednostek pierwszej zbiorowości przypadającą na określoną jednostkę drugiej zbiorowości [7, s. 89].

Wskaźniki natężenia pojawiły się już we wcześniejszej części tego opracowania. Klasycznym przykładem jest gęstość zaludnienia (zob. rys. 1.15), czyli liczba mieszkańców przypadająca na 1 km^2 powierzchni danego obszaru. Inne ekonomiczne przykłady tego typu wskaźników to (por. [7, s. 89]):

- liczba mieszkań oddanych do użytku na 1000 mieszkańców według województw,
- cena 1 m^2 powierzchni mieszkania w danym województwie,
- wskaźnik wydajności pracy, tj. wartość przychodów na 1 zatrudnionego,
- wskaźnik rotacji aktywów (wartość przychodów ze sprzedaży na 1 zł majątku przedsiębiorstwa),
- wartość księgowa na 1 akcję,
- PKB *per capita*, tj. Produkt Krajowy Brutto na 1 mieszkańca.

Ponadto w rozdziale pierwszym pojawił się wskaźnik natężenia niezwiązany z ekonomią, a mianowicie wskaźnik natężenia liczebności. Jeśli jako rozpiętość bazowego przedziału klasowego przyjmie się wartość „1”, to wówczas otrzyma się relację liczebności *i-tej* klasy (n_i) do jej rozpiętości (h_i). Innym przykładem wskaźnika natężenia – niezwiązanego z dziedziną ekonomii – jest prędkość, czyli relacja drogi do czasu mierzona np. liczbą przebytych kilometrów na godzinę czy też w m/s (np. siła wiatru). Oto przykład obliczania wskaźników natężenia:

Przykład. W tabeli poniżej zawarte są informacje o zatrudnieniu i wielkości przychodów ze sprzedaży w trzech oddziałach firmy. Na podstawie tych informacji obliczono wskaźniki wydajności pracy:

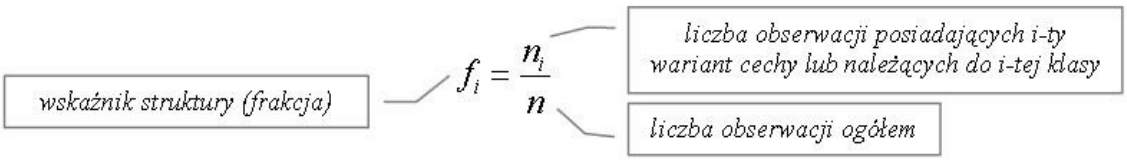
Tabela 2.2. Wydajność pracy w poszczególnych oddziałach przedsiębiorstwa.

Oddziały	Przychody (zł mies.)	Liczba zatrudnionych	Wydajność pracy (zł/os.)
I	10 000	10	10 000 / 10 = 1 000
II	20 000	40	20 000 / 40 = 500
III	40 000	20	40 000 / 20 = 2 000
Σ	**70 000**	**70**	**70 000 / 70 = 1 000**

Źródło: Obliczenia własne na podstawie danych umownych.

Najwyższą wydajnością pracy odznacza się oddział trzeci (2000 zł mies. przychodu na 1 zatrudnionego). Wyniki te należałoby odnieść do przeciętnej płacy miesięcznej. Należy zauważyć, iż przeciętna wydajność pracy w firmie na poziomie 1000 zł mies. na 1 zatrudnionego nie jest średnią arytmetyczną wydajności trzech oddziałów – bowiem aby obliczyć średnią wydajność pracy, należy zastosować wzór na średnią harmoniczną.

Wskaźniki struktury – określane również mianem frakcji lub częstości względnych – ukazują udziały poszczególnych części (klas) w danej zbiorowości [10, s. 100]. Wskaźniki te pojawiły się już przy prezentacji graficznej (zob. diagram i histogram). Pojawiło się wtedy pojęcie *częstości względnej* (frakcji), czyli relacji liczebności danej części (klasy) zbiorowości do ogólnej liczby obserwacji (por. [21, s. 32]):



Powyższy wskaźnik można też wyrazić w postaci procentowej – wystarczy poszczególne frakcje przemnożyć przez 100:

$$f_i = \frac{n_i}{n} \cdot 100$$

Frakcje sumują się do jedności lub – w ujęciu procentowym – do 100 procent. Niekiedy w literaturze podaje się wzór pozwalający na wyrażenie wskaźników struktury w promilach (zob. [7, s. 92], [10, s. 101]).

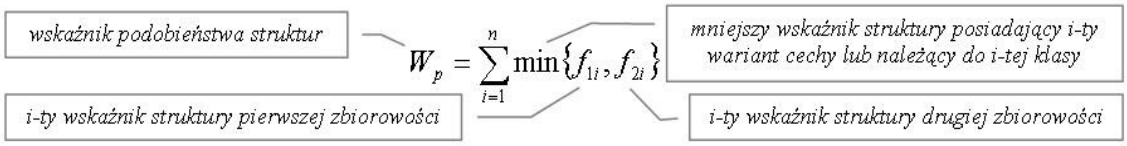
Należy podkreślić, iż wskaźniki struktury można wyznaczyć dla cech mierzonych na każdym rodzaju skali pomiarowej – do ich obliczenia niezbędne są bowiem liczebności obserwacji posiadających dany wariant cechy bądź należących do określonego przedziału klasowego (por. [20, s. 87]). Jest to zatem uniwersalna miara statystyczna. Oto przykład obliczenia wskaźników struktury na podstawie danych umownych, dotyczących ankiety internetowej odnośnie liczby godzin uczenia się statystyki tygodniowo (zob. *Dane_do_analizy.xls*, zakładka: *Ankiety*). Poniższa tabela zawiera niezbędne obliczenia:

Tabela 2.3. Wskaźniki struktury liczby godzin nauki statystyki tygodniowo w czasie sesji i poza sesją.

Liczba godzin tygodniowo x_i	**Liczebności**		**Wskaźniki struktury**	
	sesja n_{1i}	**poza sesją** n_{2i}	**sesja** f_{1i}	**poza sesją** f_{2i}
do 2 godzin	1	7	1/15 = 0,067	7/15 = 0,467
2 – 4 godziny	2	7	2/15 = 0,133	7/15 = 0,467
5 – 10 godzin	3	1	3/15 = 0,200	1/15 = 0,067
ponad 10 godzin	9	0	9/15 = 0,600	0/15 = 0,000
Σ	15	15	1	1

Źródło: Obliczenia własne na podstawie danych umownych.

Do porównania struktur dwóch zbiorowości można zastosować ***wskaźnik podobieństwa struktur*** (por. [20, s. 88-89]):



Nawiązując do powyższego przykładu: do wyznaczenia wskaźnika podobieństwa struktur potrzebne będzie wprowadzenie dodatkowej kolumny (por. tabela 2.3):

Tabela 2.4. Wskaźnik podobieństwa struktur godzin nauki statystyki tygodniowo w czasie sesji i poza sesją.

Liczba godzin tygodniowo x_i	**liczebności**		**wskaźniki struktury**		$\min\{f_{1i}, f_{2i}\}$
	sesja n_{1i}	**poza sesją** n_{2i}	**sesja** f_{1i}	**poza sesją** f_{2i}	
do 2 godzin	1	7	0,067	0,467	0,067
2 – 4 godziny	2	7	0,133	0,467	0,067
5 – 10 godzin	3	1	0,200	0,067	0,000
ponad 10 godzin	9	0	0,600	0,000	0,000
Σ	15	15	1	1	0,133

Źródło: Obliczenia własne na podstawie danych umownych.

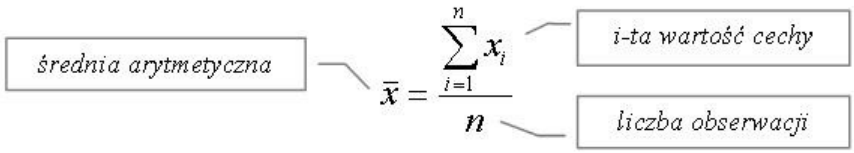
Wartość omawianego wskaźnika jest wielkością unormowaną, tzn. zawiera się w przedziale [0,1]. Im większe podobieństwo struktur porównywanych zbiorowości, tym wartość wskaźnika bliższa jedności (dla struktur identycznych wskaźnik osiąga wartość równą 1). Wskaźnik na poziomie 0,133 świadczy o dużym zróżnicowaniu struktur liczby godzin nauki statystyki w sesji i poza sesją.

2.1.2. Miary położenia

Miary położenia (średnie, tendencji centralnej) w syntetyczny sposób charakteryzują badaną zbiorowość statystyczną. Z uwagi na swój syntetyczny charakter nadają się one do porównań zbiorowości w czasie i przestrzeni. Główną zaletą tych miar – w odróżnieniu od wskaźników struktury – jest wyrażanie ich wielkości w liczbach mianowanych, tj. w takich jednostkach miary, w jakich wyrażona jest wartość danej cechy statystycznej [7, s. 116-117].

Klasyczną miarą położenia jest ***średnia arytmetyczna***. Należy zaznaczyć, iż miara ta jest dostępna tylko dla cech mierzonych za pomocą skali przedziałowej bądź ilorazowej. W statystyce matematycznej (zob. Wnioskowanie statystyczne) istotne jest rozróżnienie średniej arytmetycznej dla próby od średniej arytmetycznej dla populacji generalnej *m* (por. [3, s. 99]).

To, z jakiego wzoru należy obliczyć średnią arytmetyczną, zależy od tego, czy dane zostały pogrupowane w szereg rozdzielczy czy też nie. I tak, dla danych niepogrupowanych średnią arytmetyczną wyznacza się ze wzoru:



Oto przykład obliczania średniej arytmetycznej według powyższego wzoru:

Przykład. W ankiecie dla Czytelników (zob. rys. 1.6) w pytaniu nr 6 poproszono respondentów m.in. o ocenę jakości treści niniejszego opracowania na pięciostopniowej skali Stapela. Oto oceny uzyskane na podstawie piętnastu ankiet internetowych (dane umowne):

5, 4, 4, 5, 3, 4, 2, 4, 3, 5, -1, -4, 1, -2, -5

W rozbudowanym przykładzie zamieszczonym w rozdziale pierwszym (*Trening i ewaluacja*) powyższe dane uśredniono za pomocą *Raportu tabeli przestawnej* (zob. aplikacja MS Excel: *Przykłady – grupowanie danych*). Ponadto w programie MS Excel wśród funkcji statystycznych (*Wstaw…*, *Funkcja…*, a następnie określenie funkcji statystycznych) dostępna jest wbudowana funkcja obliczająca średnią arytmetyczną dla danych niepogrupowanych:

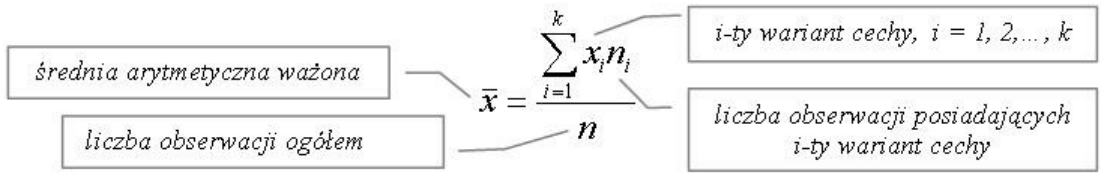
ŚREDNIA(*zakres_danych*)

Aby tradycyjnie obliczyć średnią arytmetyczną, należy zsumować uzyskane punkty, a następnie podzielić je przez liczbę obserwacji, tj. $n = 15$ (liczba otrzymanych ankiet):

$$\bar{x} = \frac{28}{15} = 1{,}866$$

Przeciętna liczba punktów wskazuje na pozytywną ocenę prezentowanych treści.

Dla danych pogrupowanych w szereg rozdzielczy punktowy oblicza się *ważoną średnią arytmetyczną* według poniższego wzoru:



Przykład. Pewna szkoła prywatna ocenia swoją ofertę edukacyjną według sporządzonej listy kryteriów. W ankiecie przeprowadzonej na reprezentatywnej grupie 200 studentów zadano pytanie: *Który z wymienionych czynników jest dla Pana/Pani najistotniejszy?* (tylko jedna opcja odpowiedzi):

a) cena kursu,
b) zróżnicowanie oferty edukacyjnej,
c) wiedza i umiejętności kadry dydaktycznej,
d) możliwość nauki przez Internet,
e) dogodna lokalizacja,
f) materiały dydaktyczne wliczone w cenę kursu.

Ocena oferty według każdego z powyższych kryteriów została dokonana przez właściciela szkoły w skali od 0 do 10. Aby obliczyć średnią arytmetyczną ważoną, konieczne jest wprowadzenie dodatkowej kolumny $x_i\, n_i$. Oto niezbędne obliczenia:

Tabela 2.5. Średnia ważona ocena atrakcyjności oferty edukacyjnej szkoły prywatnej.

Czynniki i	Ocena x_i	Liczba wskazań n_i	Obliczenia pomocnicze $x_i n_i$
a)	7	92	$7 \times 92 = 644$
b)	4	29	116
c)	8	38	304
d)	0	17	0
e)	4	14	56
f)	0	10	0
Σ		**200**	**1120**

Źródło: Obliczenia własne na podstawie danych umownych.

Na podstawie sporządzonej tabeli pomocniczej można stosunkowo łatwo obliczyć niezbędne sumy $x_i n_i$, a następnie podstawić do wzoru na średnią ważoną:

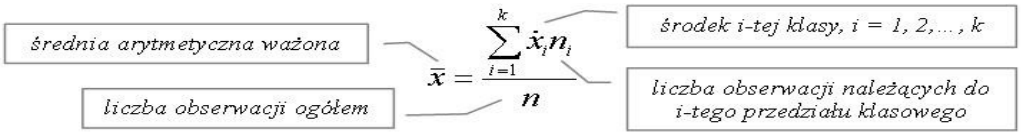
$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{1120}{200} = 5{,}6$$

Z uwagi na dysjunktywny charakter pytania ankiety (wymagane wskazanie tylko jednego czynnika) liczba wskazań jest równa liczbie respondentów (n = 200). Uzyskana ważona ocena punktowa – gdzie wagami n_i są liczby wskazań – sugeruje, iż oferta szkoły jest przeciętna. W związku z tym należałoby podjąć pewne działania zmierzające do uczynienia tej oferty bardziej atrakcyjną (np. poszerzenie oferty o dodatkowe kursy).

Podstawowym błędem jest niestosowanie odpowiedniego wzoru dla danych pogrupowanych, tj. nieuwzględnianie wag, czyli liczebności cząstkowych n_i. W związku z tym – zamiast dzielenia przez liczbę wszystkich obserwacji n (w powyższym przykładzie liczbę wskazań), niektórzy studenci dzielą przez liczbę wariantów k (na zasadzie analogii do wzoru na tradycyjną średnią). Należy więc pamiętać o uwzględnianiu wag w przypadku danych pogrupowanych w szereg punktowy bądź z przedziałami klasowymi.

Dla danych pogrupowanych w szereg rozdzielczy z przedziałami klasowymi średnią arytmetyczną ważoną oblicza się w analogiczny sposób jak

średnią dla szeregu punktowego, przy czym zamiast wartości x_i zastosowanie znajdują środki przedziałów klasowych:



Środki przedziałów klasowych były już wyznaczane przy prezentacji materiału statystycznego (zob. diagram). Stanowią one średnią arytmetyczną dolnej i górnej granicy przedziału klasowego.

Przykład. Inwestor rozważa zakup akcji spółki Żywiec. W związku z tym interesuje go przeciętna wartość tygodniowych stóp zwrotu tych akcji, uzyskanych w pierwszym półroczu 2006 r. (zob. *Dane_do_analizy.xls*, zakładka: *Akcje*). Dane pogrupowaneow szereg rozdzielczy z przedziałami klasowymi (zob. *Przykłady – grupowanie danych*). Na podstawie pogrupowanych danych należy wyznaczyć ważoną średnią arytmetyczną tygodniowych stóp zwrotu akcji spółki Żywiec. W tabeli poniżej znajdują się niezbędne obliczenia:

Tabela 2.6. Oczekiwana stopa zwrotu z inwestycji w akcje spółki Żywiec (proc. tygodniowo).

I	**Stopy zwrotu** x_i	**Liczba tygodni** n_i	**Środki klas** $\dot{x}_i$	**Obliczenia pomocnicze** $\dot{x}_i \cdot n_i$
1	–10,00 – –7,51	1	–8,75	1 × (–8,75) = –8,75
2	–7,50 – –5,01	1	–6,25	–6,25
3	–5,00 – –2,51	1	–3,75	–3,75
4	–2,50 – –0,01	9	–1,25	–11,25
5	0,00 – 2,49	11	1,25	13,75
6	2,50 – 4,99	1	3,75	3,75
7	5,00 – 7,50	1	6,25	6,25
Σ		**25**		**–6,25**

Źródło: Obliczenia własne na podstawie danych pochodzących z Serwisu Internetowego Gazety Parkiet, http://www.parkiet.com/dane/dane_atxt.jsp

Należy wyjaśnić, iż wartość górnego przedziału klasowego odpowiada wartości dolnego przedziału następnej klasy (różnice z dokładnością do 0,01 informują, że przedziały są lewostronnie domknięte). Przykładowo, środek pierwszego przedziału klasowego obliczono następująco:

$$\dot{x}_i = \frac{-10 + (-7{,}5)}{2} = -8{,}75$$

Wartość średnią obliczono w oparciu o wyznaczone sumy w powyższej tabeli:

$$\bar{x} = \frac{\sum_{i=1}^{k} \dot{x}_i n_i}{n} = \frac{-6{,}25}{25} = -0{,}25$$

Przeciętna tygodniowa stopa zwrotu akcji spółki Żywiec wyniosła –0,25 proc., stąd w pierwszym półroczu 2006 r. inwestycje w te walory nie przyniosły zysków w dłuższym horyzoncie czasu (niewielka strata).

Wagami we wzorach na średnie ważone – oprócz liczebności n_i – mogą też być wskaźniki struktury (frakcje – f_i). Wówczas wzory będą miały postać:

a) szereg punktowy:

$$\bar{x} = \sum_{i=1}^{k} x_i f_i$$

b) szereg klasowy:

$$\bar{x} = \sum_{i=1}^{k} \dot{x}_i f_i$$

Przykład. Praktycznym przykładem zastosowania pierwszego z zaprezentowanych powyżej wzorów na średnią ważoną (szereg punktowy) jest określenie oczekiwanej stopy zwrotu portfela akcji. Wagami są udziały poszczególnych walorów. Oto sposób obliczeń:

Tabela 2.7. Oczekiwana roczna stopa zwrotu portfela akcji.

Spółki I	Stopa zwrotu (proc.) x_i	Struktura portfela f_i	Obliczenia pomocnicze $x_i f_i$
A	33	0,24	$33 \times 0{,}24 = 7{,}92$
B	40	0,15	6,00
C	14	0,05	0,70
D	22	0,27	5,94
E	18	0,29	5,22
Σ		1,00	25,78

Źródło: Obliczenia własne na podstawie danych umownych.

Średnia stopa zwrotu portfela wyniosła 25,78 proc. rocznie. Jak widać, wartość średniej została odczytana bezpośrednio z tabeli, bez konieczności dodatkowych obliczeń.

Ponieważ miary klasyczne dla danych pogrupowanych w szereg rozdzielczy punktowy oraz dla danych pogrupowanych w szereg z przedziałami klasowymi wyznacza się w sposób analogiczny, stąd w dalszej części teoretycznej będą pojawiać się przykłady obliczeń tego typu miar dla szeregu z przedziałami klasowymi (kontynuacja przykładu z tygodniowymi stopami zwrotu akcji spółki Żywiec).

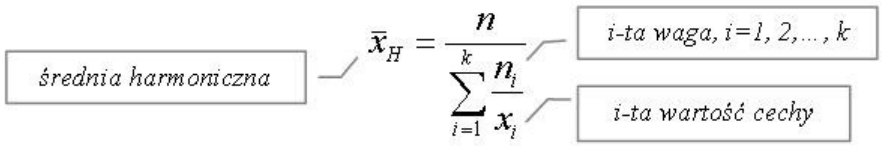
Jeżeli dane występują w postaci wskaźników natężenia, to do wyznaczenia ich wartości przeciętnej – jak już zasygnalizowano – stosuje się ***średnią harmoniczną***. Rozróżnia się średnią harmoniczną prostą oraz ważoną (por. [21, s. 54]):

a) średnia harmoniczna prosta:

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

średnia harmoniczna; n – liczba obserwacji; x_i – i-ta wartość cechy

b) średnia harmoniczna ważona:



Przykład 1. Student postanowił przeznaczyć 300 zł na korepetycje ze statystyki. Wybrał losowo trzech korepetytorów (n = 3), oferujących odpowiednio ceny za godzinę korepetycji: 25 zł, 40 zł i 50 zł. U każdego z nich postanowił zakupić lekcje za kwotę 100 zł. Przeznaczone kwoty pozwoliły odpowiednio na zakup 4 godzin u pierwszego korepetytora, 2,5 godziny u drugiego oraz 2 godzin u trzeciego (w sumie 8,5 godziny). Ponieważ poszczególne kwoty są sobie równe (po 100 zł), stąd przeciętną cenę jednej godziny korepetycji można obliczyć ze wzoru na prostą średnią harmoniczną:

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{3}{\frac{1}{25} + \frac{1}{40} + \frac{1}{50}} = \frac{3}{0{,}085} = 35{,}29$$

Przeciętna cena korepetycji to 35,29 zł/godz. Wartość tę można uzyskać, dzieląc łączne wydatki na korepetycje (300 zł) przez zakupioną liczbę godzin ogółem (8,5 godz.). Średnią harmoniczną prostą można wyznaczyć w Excelu, posługując się funkcją:

ŚREDNIA.HARMONICZNA(25; 40; 50)

Możliwe jest oczywiście podanie zakresu komórek, do których wpisano ceny korepetycji (w trzech sąsiadujących wierszach lub kolumnach).

Przykład 2. Wracając do przykładu dotyczącego wydajności pracy (wartość przychodów na 1 zatrudnionego): można stwierdzić, że mamy tu do czynienia ze średnią harmoniczną ważoną. Jako wagi n_i cechy będącej relacją dwóch wielkości należy przyjąć wartości jej licznika – w tym przykładzie będą to przychody wyrażone w zł (w mianowniku występuje liczba zatrudnionych). Oto sposób obliczenia średniej harmonicznej ważonej:

Tabela 2.8. Przeciętna wydajność pracy w przedsiębiorstwie posiadającym trzy oddziały regionalne.

Oddziały	**Wydajność pracy (zł/os.)** x_i	**Przychody (zł)** n_i	**Liczba zatrudnionych** n_i/x_i
I	1 000	10 000	10 000 / 1 000 = 10
II	500	20 000	20 000 / 500 = 40
III	2 000	40 000	40 000 / 2 000 = 20
Σ		70 000	70

Źródło: Obliczenia własne na podstawie danych umownych.

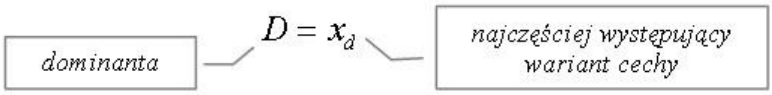
Na podstawie obliczeń pomocniczych zawartych w powyższej tabeli można wyznaczyć w prosty sposób średnią harmoniczną ważoną:

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{70000}{70} = 1000$$

Suma wag stanowi ogólną wartość przychodów przedsiębiorstwa (n = 70 000). Wartość średniej harmonicznej informuje, że przeciętna wydajność pracy w badanym przedsiębiorstwie to 1000 zł na 1 zatrudnionego.

Kolejną grupę – obok klasycznych – stanowią *pozycyjne miary średnie*. Ich niewątpliwą zaletą jest to, że mogą być one – w przeciwieństwie do średniej arytmetycznej – wyznaczone również dla cech mierzonych za pomocą skal słabszych (zob. skala nominalna i skala porządkowa), przy czym dominantę można określić nawet dla cechy mierzonej na skali nominalnej. Inną zaletą jest to, że miary te można obliczyć w oparciu o ograniczony zbiór danych (ma to znaczenie, gdy np. skrajne przedziały klasowe nie są domknięte).

Dominantą (modalną, modą) w zbiorze danych jakościowych jest występujący najczęściej i-ty wariant cechy (por. [3, s. 116-117]):



Przykład. Właściciel szkoły prywatnej chce określić najistotniejszy czynnik decydujący o atrakcyjności oferty edukacyjnej. W tym celu poproszono grupę losowo wybranych studentów o określenie jednego z sześciu sugerowanych czynników. Po zliczeniu odpowiedzi okazało się, że aż 92 respondentów (wielkość próby to n = 200 studentów) wskazało na cenę (zob. tabela 2.5). Zatem cena okazała się czynnikiem najważniejszym.

W przypadku danych ilościowych dominantę można wyznaczyć przy założeniu, że rozkład cechy jest jedno- lub wielomodalny, nie zaś amodalny (zob. rys. 2.1). Sposób obliczania dominanty zależy od tego czy dane pogrupowano w szereg rozdzielczy punktowy czy też z przedziałami klasowymi (dominanty nie można obliczyć dla danych niepogrupowanych). W szeregu rozdzielczym punktowym wartość dominanty można wskazać od razu, tak jak w przypadku danych jakościowych.

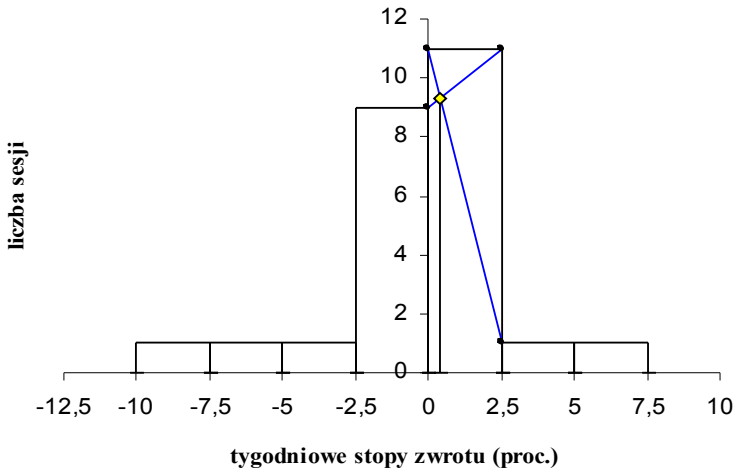
Przykład. Rozkład liczby kont *e-mail* (zob. rys. 1.18) jest rozkładem jednomodalnym prawostronnie asymetrycznym (zob. rys. 1.18). Na podstawie sporządzonego histogramu łatwo zauważyć, iż najwięcej ankietowanych internautów posiadało jedno konto *e-mail*.

W tym miejscu warto podkreślić, iż dominanta to wartość cechy, a nie odpowiadająca jej liczebność. Niejednokrotnie zamiast podania wartości dominanty (w tym przypadku jedno konto *e-mail*) zdarza się, że student podaje liczebność (w tym przykładzie liczba internautów).

W szeregu rozdzielczym z przedziałami klasowymi wyznaczenie wartości dominanty wymaga zastosowania wzoru interpolacyjnego (zob. szacunek statystyczny). Bardzo pomocne jest graficzne wyznaczenie dominanty. W tym celu należy sporządzić histogram (dla równych przedziałów klasowych jest to *histogram liczebności* lub *histogram częstości względnych*),

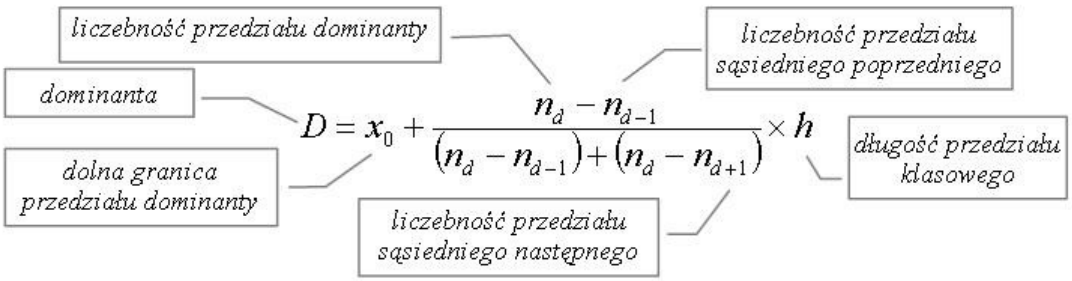
a następnie ustalić punkt przecięcia się linii, tak jak pokazano to na rys. 2.2:

Rysunek 2.2. Rozkład tygodniowych stóp zwrotu akcji spółki Żywiec w I półroczu 2006 r.



Źródło: Opracowanie na podstawie danych pochodzących z Serwisu Internetowego Gazety Parkiet, http://www.parkiet.com/dane/dane_atxt.jsp

Po zrzutowaniu argumentów punktu, w którym przecięły się wyznaczone linie, na oś OX otrzymano wartość dominanty (por. [3, s. 119]). Analitycznie wielkość tę można wyznaczyć ze wzoru dla danych pogrupowanych w szereg rozdzielczy z równymi przedziałami klasowymi:



Przykład. Na podstawie danych dotyczących tygodniowych stóp zwrotu akcji spółki Żywiec należy obliczyć dominantę, czyli najczęstszą tygodniową stopę zwrotu. W oparciu o sporządzony histogram (zob. rys. 2.2) nietrudno stwierdzić, iż przedziałem dominanty jest przedział: [0-2,5 proc.). Do obliczenia dominanty niezbędne są następujące informacje (zob. tabela 2.6):

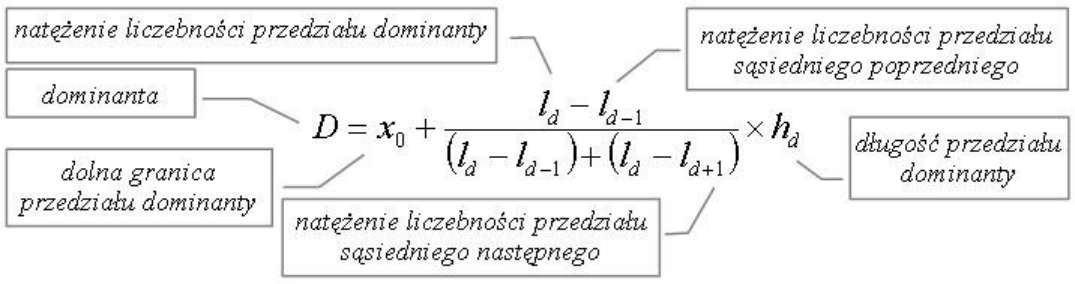
a) dolna granica przedziału dominanty: $x_0 = 0$,
b) liczebność przedziału dominanty: $n_d = 11$,
c) liczebność przedziału sąsiedniego poprzedzającego: $n_{d-1} = 9$,
d) liczebność przedziału sąsiedniego następnego: $n_{d+1} = 1$,
e) rozpiętość przedziału klasowego (wszystkie przedziały są sobie równe): $h = 2{,}5$.

Po podstawieniu do wzoru należy pamiętać, że otrzymaną liczbę na końcu dodajemy do dolnej granicy (w tym przykładzie nie ma to znaczenia, bo wartość ta jest równa zeru):

$$D = x_0 + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} \times h = 0 + \frac{11-9}{(11-9)+(11-1)} \times 2{,}5 = 0 + \frac{2}{12} \times 2{,}5 = 0{,}417$$

Zatem w pierwszym półroczu 2006 r. najczęstsza tygodniowa stopa zysku z akcji spółki Żywiec była wielkością dodatnią (0,42 proc.), tj. ok. 1,7 proc. miesięcznie.

Szczególną ostrożność przy wyznaczaniu miar pozycyjnych, w tym dominanty, należy zachować w przypadku szeregu rozdzielczego z nierównymi przedziałami klasowymi. Zwrócono już na ten fakt uwagę przy omawianiu wykresów statystycznych. Wracając do przykładu z rozkładem wieku budynków mieszkalnych w Polsce (stan na 2002 r.): w tym wypadku można obliczyć dominantę na podstawie rys. 1.20. Jak stwierdzono, dominanta zawiera się w przedziale 1971-1979 (zob. tabela 1.18). Znajduje tu zastosowanie wzór analogiczny do wzoru na dominantę w szeregu rozdzielczym z równymi przedziałami klasowymi, przy czym pojawią się tu wskaźniki natężenia liczebności l_i:



Podstawiamy do wzoru następujące wartości:

a) dolna granica przedziału dominanty: $x_0 = 1971$,
b) natężenie liczebności przedziału dominanty: $l_d = 3493$,
c) natężenie liczebności przedziału sąsiedniego poprzedzającego: $l_{d-1} = 1582$,
d) natężenie liczebności przedziału sąsiedniego następnego: $l_{d+1} = 2857$,
e) rozpiętość przedziału dominanty: $h_d = 8$.

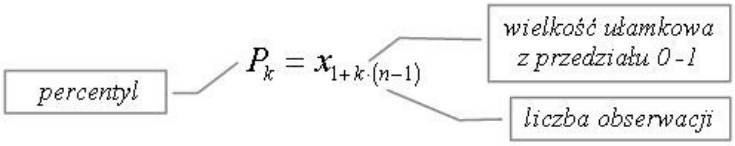
$$D = 1971 + \frac{3493 - 1582}{(3493 - 1582) + (3493 - 2857)} \times 8 = 1971 + 6 = 1977$$

Jak wynika z obliczeń przeprowadzonych na podstawie danych Narodowego Spisu Powszechnego z 2002 r. – najwięcej mieszkań w Polsce wybudowano w 1977 r. Są to na ogół piętrowe budynki, wznoszone z betonowych płyt.

W szeregach rozdzielczych z nierównymi przedziałami klasowymi wyznaczenie dominanty niejednokrotnie może okazać się sprawą trudną. Podstawowy błąd polega na nieodpowiednim sporządzeniu histogramu (dla liczebności zwykłych zamiast dla natężenia liczebności) i co się z tym wiąże niestosowaniu wzoru uwzględniającego wskaźniki natężenia liczebności – stąd kluczowe znaczenie ma prawidłowe sporządzenie histogramu.

Dla danych opartych minimum na skali porządkowej można – obok dominanty – obliczyć kwantyle. ***Kwantyle*** to „wartości cechy badanej w zbiorowości, które dzielą ją na określone części pod względem liczby jednostek. Części te mogą być równe lub pozostawać do siebie w określonych proporcjach” [19, s. 43]. W szczególności wśród kwantyli wyróżnia się *percentyle* (dzielące zbiorowość na 100 części), *decyle* (10 części) i *kwartyle* (4 części).

W przypadku danych indywidualnych (niepogrupowanych) istotne jest to, aby warianty cechy były uporządkowane rosnąco. Ogólnie *k-tym percentylem* w uporządkowanym zbiorze wartości cechy jest taka wartość, poniżej której znajduje się k-ty procent wartości z tego zbioru (por. [13, s. 29]):

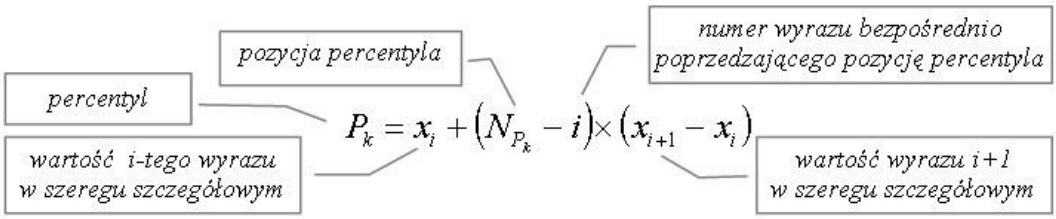


Przykładowo, 28 percentyl (k = 0,28) dzieli zbiorowość w ten sposób, że 28 proc. jednostek statystycznych posiada wartości nie większe niż wartość tego kwantyla.

W wielu sytuacjach wartość danego percentyla nie pokrywa się z wartością danego wyrazu w uporządkowanym rosnąco szeregu statystycznym, lecz z wielkością znajdującą się pomiędzy dwoma wyrazami:

$$P_k \in (x_i, x_{i+1})$$

W tej sytuacji należy skorzystać z bardziej zaawansowanego wzoru interpolacyjnego:



Pozycję percentyla ustala się analogicznie jak numer obserwacji w pierwszym prezentowanym wzorze na k-ty percentyl:

$$N_{P_k} = 1 + k \cdot (n-1)$$

Jedynie w przypadku szczególnym, gdzie pozycja percentyla jest liczbą całkowitą, jej wartość można wyznaczyć od razu: $P_k = x_i$.

Medianę, będącą drugim kwartylem (5 decylem, 50 percentylem), można obliczyć z następujących (uproszczonych) wzorów:

a) liczba obserwacji nieparzysta:

$$Me = x_{\frac{1}{2}\cdot(n+1)}$$

b) liczba obserwacji parzysta:

$$Me = \frac{1}{2}\cdot\left(x_{\frac{1}{2}n} + x_{\frac{1}{2}n+1}\right)$$

Wielkość ta dzieli populację na dwie części. Dla parzystej liczby obserwacji jest to wyraz środkowy uporządkowanego ciągu (szereg szczegółowy), zaś dla nieparzystej liczby obserwacji – średnia arytmetyczna z dwóch środkowych wartości tego ciągu. Oto przykłady:

Przykład 1. Wyznaczyć medianę i pozostałe kwartyle przeciętnej ceny jednego metra kwadratowego mieszkania 1-pokojowego na rynku wtórnym w większych miastach Polski (zob. *Dane_do_analizy.xls*; zakładka: *Mieszkania*).

Punktem wyjścia jest uporządkowanie danych rosnąco:
1. Poznań: 3606 zł/m^2.
2. Gdańsk: 3630 zł/m^2.
3. Wrocław: 4500 zł/m^2.
4. Kraków: 5843 zł/m^2.
5. Warszawa: 5993 zł/m^2.

Z uwagi na nieparzystą liczbę danych (n = 5) – medianę wyznacza się według wzoru:

$$Me = x_{\frac{1}{2}\cdot(n+1)} = x_{\frac{1}{2}\cdot(5+1)} = x_3 = 4500$$

Wartością środkową, czyli medianą, okazała się przeciętna cena 1 metra kw. mieszkania 1-pokojowego we Wrocławiu. W dwóch porównywanych miastach ceny w analogicznym okresie okazały się niższe (Poznań, Gdańsk), a w pozostałych dwóch – wyższe (Kraków, Warszawa).

Pozostałe kwartyle, tj. kwartyl pierwszy (dolny) i trzeci (górny) można wyznaczyć z ogólnego wzoru na k-ty percentyl:

a) kwartyl pierwszy (25 percentyl):

$$P_{0,25} = x_{1+0,25\cdot(5-1)} = x_{1+1} = x_2 = 3630$$

b) kwartyl trzeci (75 percentyl):

$$P_{0,75} = x_{1+0,75\cdot(5-1)} = x_{1+3} = x_4 = 5843$$

W przypadku jednej czwartej miast objętych analizą cena 1 metra kw. kawalerki nie przekroczyła 3630 zł (Poznań) – w pozostałych miastach ceny w badanym okresie były wyższe. Analogicznie interpretuje się kwartyl trzeci: ceny 1 metra kw. kawalerki w 75 proc. analizowanej zbiorowości nie przekroczyły 5843 zł – w pozostałych 25 proc. porównywanych miast były one wyższe (Warszawa). Analizę tę można uogólnić na większą liczbę miast.

Przykład 2. W pierwszym pytaniu kwestionariusza ankiety dla Czytelników (wzór kwestionariusza zaprezentowano na rys. 1.6) respondenci mieli określić czy niniejsza publikacja pomogła im w przygotowaniu się do egzaminu. Dane umowne zawiera arkusz *Dane_do_analizy.xls* (zakładka *Ankiety*). Przyjęto następujący sposób kodowania danych:

–2 – zdecydowanie nie,
–1 – raczej nie,
0 – trudno powiedzieć,
+1 – raczej tak,
+2 – zdecydowanie tak.

Należy obliczyć medianę i pierwszy kwartyl na podstawie wybranych ankiet. Tak jak w przykładzie poprzednim, najpierw należy posortować odpowiedzi rosnąco:

Numer obserwacji i	1	2	3	4	5	6	7	8	9	10	11	12
Wartości wyrazów x_i	-2	-1	-1	0	0	0	1	1	1	1	2	2

Z uwagi na parzystą liczbę objętych analizą formularzy (n = 12) – do obliczenia mediany znajduje zastosowanie drugi z prezentowanych wyżej wzorów:

$$Me = \frac{1}{2}\cdot\left(x_{\frac{1}{2}n} + x_{\frac{1}{2}n+1}\right) = \frac{1}{2}\cdot\left(x_{\frac{1}{2}\cdot 12} + x_{\frac{1}{2}\cdot 12+1}\right) = \frac{1}{2}\cdot\left(x_6 + x_7\right) = \frac{1}{2}\cdot\left(0+1\right) = 0{,}5$$

Zatem połowa respondentów nie miała zdania (0) lub stwierdziła, że e-book nie był pomocny w przygotowaniu się do egzaminu ze statystyki (–2, –1). Jednocześnie co drugi ankietowany przyznał, że publikacja okazała się przydatna w zdaniu egzaminu (+1, +2). Jeśli chodzi o kwartyl pierwszy, to w tym przykładzie szukana wartość znajduje się pomiędzy trzecim (i = 3) a czwartym wyrazem uporządkowanego rosnąco ciągu liczb:

$$N_{P_{0,25}} = 1 + 0{,}25\cdot\left(n-1\right) = 1 + 0{,}25\cdot\left(12-1\right) = 3{,}75 \in \left(3,\, 4\right)$$

W tej sytuacji należy posłużyć się wzorem interpolacyjnym.

$$P_{0,25} = x_3 + \left(N_{P_{0,25}} - 3\right)\times\left(x_4 - x_3\right) = -1 + \left(3{,}75-3\right)\times\left(0-\left(-1\right)\right) = -1 + 0{,}75\times 1 = -0{,}25$$

Zdaniem co czwartego Czytelnika publikacja nie była lub raczej nie była mu pomocna w przygotowaniu się do egzaminu.

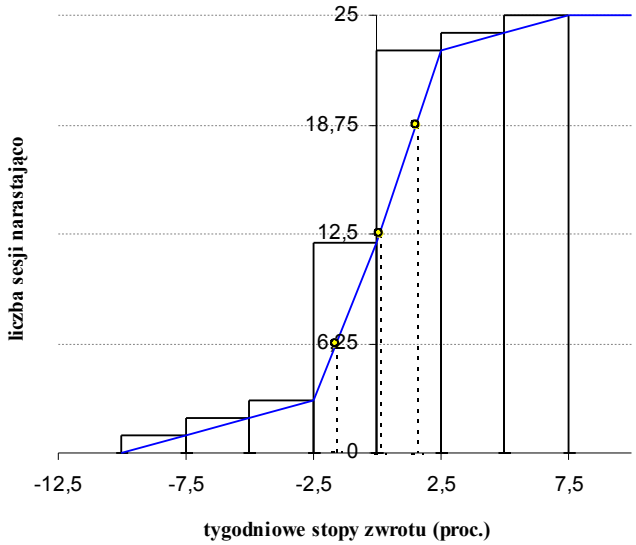
Dane w postaci szeregu punktowego należy tak traktować, jak dane w postaci omówionego szeregu szczegółowego (analogiczny sposób wyznaczania percentyli). W programie MS Excel wbudowana jest funkcja, którą można stosować do wyznaczania wartości k-tego percentyla dla danych niepogrupowanych:

PERCENTYL(*zakres_danych*; *k*)

Dla danych pogrupowanych w szereg rozdzielczy z przedziałami klasowymi – jak już zasygnalizowano – kwartyle można wyznaczyć graficznie poprzez narysowanie wykresu *kumulanty* (zob. rys. 1.23). Poniżej przedsta-

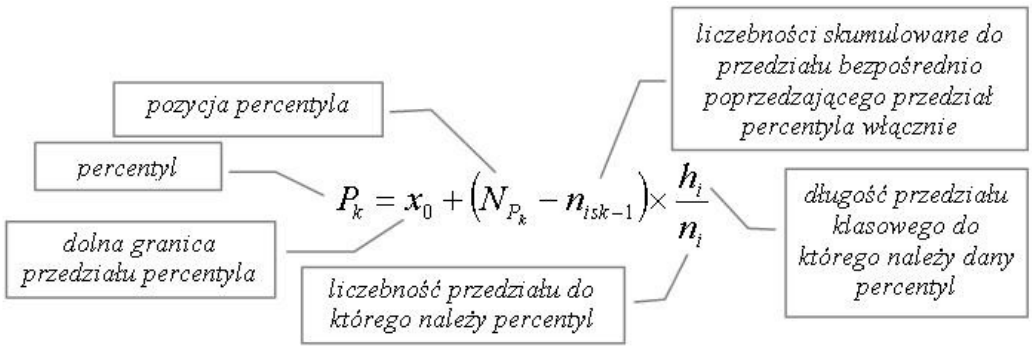
wiono sposób graficznego wyznaczania wartości kwartyli (analogicznie można wyznaczyć dowolny percentyl) dla danych będących kontynuacją przykładu dotyczącego tygodniowych stóp zysku cen akcji spółki Żywiec:

Rysunek 2.3. Wykres kumulanty tygodniowych stóp zwrotu akcji spółki Żywiec w I półroczu 2006 r.

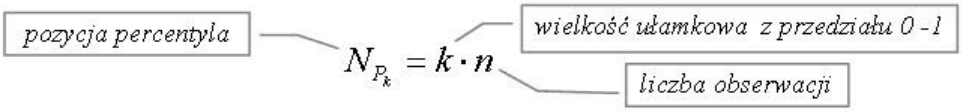


Źródło: Opracowanie na podstawie danych pochodzących z Serwisu Internetowego Gazety Parkiet, http://www.parkiet.com/dane/dane_atxt.jsp.

Po zrzutowaniu punktów przecięcia się pozycji kwartyli (poziome linie przerywane) z kumulantą – otrzyma się wartości kwartyli (odczyt z osi OX). Wielkości te można obliczyć, stosując wzór interpolacyjny dla danych pogrupowanych w szereg rozdzielczy z przedziałami klasowymi (uogólnienie interpolacyjnego wzoru dla danych niepogrupowanych):



Pozycję percentyla wyznacza się natomiast ze wzoru:



Przy obliczaniu kwartyli najpierw należy ustalić ich pozycje:

1. *Pierwszy kwartyl* to wartość cechy, dzieląca daną zbiorowość w ten sposób, że 25 proc. jednostek przyjmuje wartości mniejsze lub równe tej wartości, a pozostałe – większe; stąd pozycja tego kwartyla wynosi $0{,}25 \cdot n$.
2. *Drugi kwartyl* (mediana) to wartość cechy, dzieląca populację na połowę – stąd pozycja $0{,}5 \cdot n$.
3. *Trzeci kwartyl* to wartość cechy, dzieląca populację w proporcji: 75 proc. jednostek przyjmuje wartości nie większe od trzeciego kwartylu, a pozostałe 25 proc. wartości większe – dlatego pozycja tego kwartyla to $0{,}75 \cdot n$.

Następnie należy określić przedziały klasowe, w których znajdują się poszczególne kwartyle. Pomocne jest tu graficzne wyznaczenie kwartyli (zob. rys. 2.3). Niemniej jednak przedział kwartyla można wyznaczyć bezpośrednio z tabeli danych (zob. tabela 2.9). Jeśli suma liczebności przekroczy poziom pozycji kwartyla, to w danym przedziale zawiera się kwartyl, którego szukamy. Oto określenie przedziału mediany (pozycja mediany to 12,5):

Tabela 2.9. Tygodniowe stopy zwrotu z inwestycji w akcje spółki Żywiec (liczba sesji narastająco).

I	Stopy zwrotu x_i	Liczba tygodni n_i	Liczba sesji narastająco	Komentarz
1	–10,00 – –7,51	1	1	*Wartości mniejsze od pozycji mediany: 12 < 12,5*
2	–7,50 – –5,01	1	2	
3	–5,00 – –2,51	1	3	
4	–2,50 – –0,01	9	12	
5	0,00 – 2,49	11	23	*Pozycja mediany przekroczona: 23 > 12,5*
6	2,50 – 4,99	1	24	
7	5,00 – 7,50	1	25	
Σ		**25**		

Źródło: Obliczenia własne na podstawie danych pochodzących z Serwisu Internetowego Gazety Parkiet, http://www.parkiet.com/dane/dane_atxt.jsp.

Mając już określone przedziały kwartyli, w kolejnym kroku należy określić dolną granicę, liczebność i rozpiętość przedziału danego kwartyla (zakładamy tu równe klasy). Potrzebne są także liczebności skumulowane – do przedziału poprzedzającego włącznie. Oto zestawienie danych niezbędnych do obliczenia *pierwszego kwartyla*:

a) pozycja pierwszego kwartyla: 6,25
b) dolna granica przedziału pierwszego kwartyla: –2,5
c) liczebność przedziału pierwszego kwartyla: 9
d) suma liczebności trzech klas poprzedzających przedział pierwszego kwartyla: 3
e) rozpiętość przedziału pierwszego kwartyla: 2,5

Podstawiamy do wzoru:

$$Q_1 = x_0 + \left(0{,}25 \cdot n - n_{isk-1}\right) \times \frac{h_i}{n_i} = -2{,}5 + \left(6{,}25 - 3\right) \times \frac{2{,}5}{9} = -1{,}597$$

Jedna czwarta tygodniowych stóp zwrotu to spadki na poziomie minimum 1,6 proc.

A oto analogiczne dane niezbędne do wyznaczenia mediany:

a) pozycja mediany: 12,5
b) dolna granica przedziału mediany: 0
c) liczebność przedziału mediany: 11
d) suma liczebności czterech klas poprzedzających przedział mediany: 12
e) rozpiętość przedziału mediany: 2,5

$$Me = x_0 + \left(0{,}5 \cdot n - n_{isk-1}\right) \times \frac{h_i}{n_i} = 0 + \left(12{,}5 - 12\right) \times \frac{2{,}5}{11} = 0{,}114$$

Połowa osiągniętych tygodniowych stóp zysku przekroczyła poziom 1,1 proc.

W przedziale czwartym znajduje się także trzeci kwartyl, stąd w porównaniu z medianą zmieni się tu tylko pozycja kwartyla:

$$Q_3 = x_0 + \left(0{,}75 \cdot n - n_{isk-1}\right) \times \frac{h_i}{n_i} = 0 + \left(18{,}75 - 12\right) \times \frac{2{,}5}{11} = 1{,}534$$

W przypadku 25 proc. tygodni miały miejsce stopy zysku przekraczające 1,5 proc.

Pomiędzy wyznaczonymi miarami tendencji centralnej mogą zachodzić następujące zależności (por. [7, s. 121]):

a) rozkład symetryczny:

$$\bar{x} = Me = D$$

b) rozkład lewostronnie asymetryczny:

$$\bar{x} < Me < D$$

c) rozkład prawostronnie asymetryczny:

$$D < Me < \bar{x}$$

Z powyższego porównania wynika, że miary pozycyjne są znacznie mniej „czułe” na obserwacje nietypowe, stąd jest postulowane ich zastosowanie w przypadku rozkładów cechy o znacznej asymetrii. Ponadto – jak już wspomniano – zastosowanie tych miar nie wymaga zaangażowania do obliczeń wszystkich obserwacji, co jest ważne w przypadku niedomkniętych skrajnych przedziałów klasowych.

Średnią arytmetyczną można zastosować w przypadku, gdy rozkład cechy nie jest skrajnie asymetryczny czy wielomodalny. Dużym atutem tej miary jest jej stosunkowo proste obliczanie. Poza tym stanowi ona podstawę do wyznaczania innych miar klasycznych.

3. Wnioskowanie statystyczne

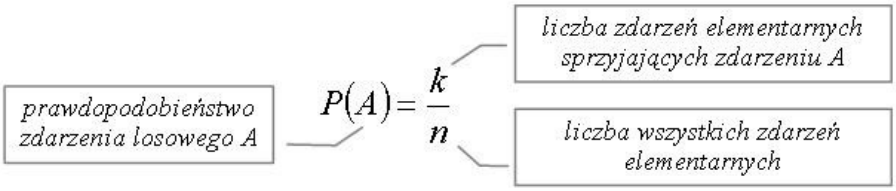
Wnioskowanie statystyczne opiera się na rachunku prawdopodobieństwa, a reguły tego wnioskowania określają metody wchodzące w skład statystyki matematycznej, w tym metody estymacji (szacowania) nieznanych parametrów strukturalnych oraz metody weryfikacji (sprawdzania) hipotez statystycznych [8, s. 10]. Estymację przedziałową oraz weryfikację hipotez statystycznych poprzedzono krótkim wprowadzeniem do rachunku prawdopodobieństwa, jak również omówiono wybrane skokowe i ciągłe rozkłady prawdopodobieństwa. Rozkłady te w większości przypadków znajdują bowiem zastosowanie w metodach wnioskowania statystycznego.

3.1. Wybrane zagadnienia z rachunku prawdopodobieństwa

Na wstępie należałoby zdefiniować pojęcie prawdopodobieństwa. **Prawdopodobieństwo** to „numeryczne wyrażenie szansy wystąpienia jakiegoś zdarzenia" [21, s. 166]. Jest to miara unormowana, tj. należąca do przedziału [0-1]. Jeżeli prawdopodobieństwo jest równe zeru, to wówczas dane zdarzenie nie wystąpi, gdy jest równe 1 – to zdarzenie jest pewne. Natomiast zdarzenia, dla których wartości prawdopodobieństwa należą do zbioru (0,1) nie są ani pewne, ani niemożliwe – przypisane im ułamki są prawdopodobieństwem zajścia danego zdarzenia.

Zgodnie z klasyczną definicją prawdopodobieństwa: prawdopodobieństwo zdarzenia losowego A – przy założeniu, że wszystkie zdarzenia elementarne są jednakowo możliwe – jest ilorazem liczby zdarzeń elementarnych sprzyjających temu zdarzeniu i liczby wszystkich zdarzeń elementarnych

[19, s. 78]. Klasyczną definicję prawdopodobieństwa zdarzenia A można wyrazić wzorem:



Oto dwa proste przykłady ilustrujące sposób obliczania prawdopodobieństwa zgodnie z klasyczną definicją:

Przykład 1. Gra „szczęśliwy numerek" polega na wylosowaniu jednej liczby spośród 49. W tej sytuacji liczba zdarzeń elementarnych wynosi n = 49 (może zostać wylosowana liczba od 1 do 49). Tylko jedna z nich okaże się wygrywającą, stąd k = 1. Prawdopodobieństwo wygranej to:

$$P(A)=\frac{k}{n}=\frac{1}{49}$$

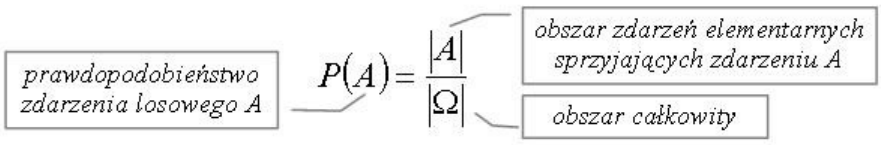
Przykład 2. Wśród 200 złożonych w pewnej miejscowości wniosków o dotacje unijne 25 okazało się źle wypełnionych. Należy obliczyć prawdopodobieństwo błędnego wypełnienia wniosku. Dane:

n = 200 wniosków,
k = 25 wniosków źle wypełnionych.

Prawdopodobieństwo zdarzenia A, polegającego na wylosowaniu wniosku posiadającego wady, wynosi:

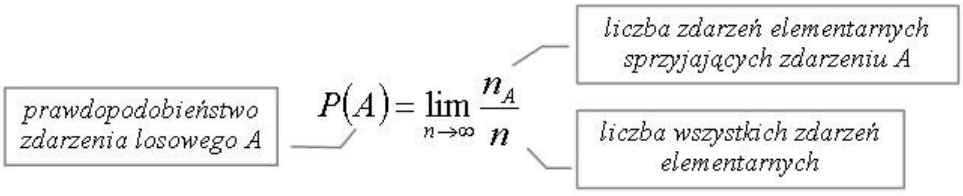
$$P(A)=\frac{k}{n}=\frac{25}{200}=0{,}125=12{,}5\%$$

Rozwinięciem klasycznej definicji prawdopodobieństwa jest *definicja graficzna*:



Obszar całkowity to przestrzeń zdarzeń elementarnych o określonej jednostce miary (długość, pole, objętość). Obszar A spełnia warunki określone zdarzeniem A. Przedstawiona definicja znajduje zastosowanie np. w rozkładach ciągłych, gdzie pole pod tzw. funkcją gęstości wynosi 1. W przypadku cech ciągłych skorzystanie z klasycznej definicji prawdopodobieństwa jest bezzasadne, ponieważ w tej sytuacji prawdopodobieństwo przyjęcia określonej wartości przez zmienną losową jest równe zeru.

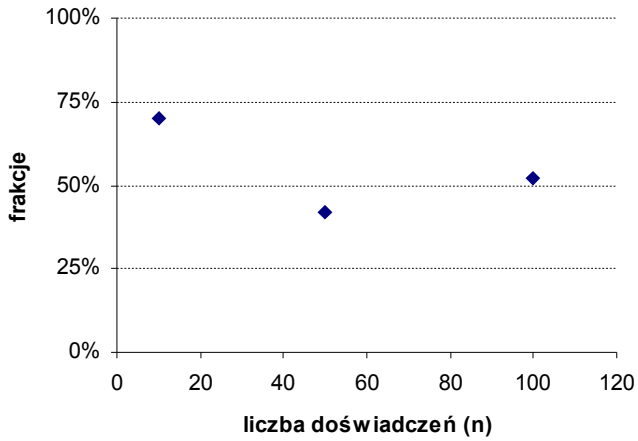
Trzecia, *statystyczna definicja prawdopodobieństwa* – zwana też częstościową lub frekwencyjną – mówi, że prawdopodobieństwo zdarzenia A jest granicą częstości tego zdarzenia, gdy liczba doświadczeń n rośnie nieograniczenie [19, s. 81]. Można to zapisać następująco:



Statystyczna definicja prawdopodobieństwa pozwala przypuszczać, że wraz ze wzrostem próby losowej frakcja (zob. wskaźnik struktury) wyznaczona na jej podstawie jest coraz bliższa wartości prawdopodobieństwa określonej według definicji częstościowej. Można tu posłużyć się prostym przykładem:

Przykład. Funkcja *los()* programu MS Excel generuje liczby z przedziału [0,1]. Jako n_A można określić wartości mniejsze bądź równe 0,5. Im więcej prób, tym wartości empiryczne (frakcje) będą bliższe teoretycznej wartości 0,5 (zob. *Przykłady – zbieżność prawdopodobieństwa*).

Rysunek 3.1. Zbieżność prawdopodobieństwa do teoretycznej wartości 0,5.

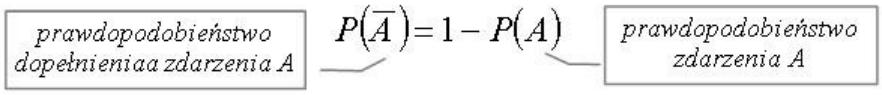


Źródło: Opracowanie własne.

Symulację przeprowadzono dla 10, 50 i 100 prób. Im więcej prób, tym różnice pomiędzy frakcjami a wartością teoretyczną 50 proc. są coraz mniejsze. Jest to zgodne z przedstawioną statystyczną definicją prawdopodobieństwa.

Mając już zdefiniowane prawdopodobieństwo, możemy sprecyzować, czym jest **zdarzenie losowe** *A* – jest to podzbiór przestrzeni zdarzeń elementarnych (Ω), zawierający wyróżnione ze względu na daną cechę zdarzenia elementarne, czyli wyniki doświadczenia losowego (por. [21, s. 167]). Nawiązując do powyższego przykładu: interesującymi nas zdarzeniami elementarnymi były wygenerowane za pomocą funkcji *los()* liczby nieprzekraczające 0,5.

Kolejną kwestią jest **algebra zdarzeń**. Na szczególną uwagę zasługuje tu *prawdopodobieństwo dopełnienia zdarzenia A* (zwanego też zdarzeniem przeciwnym do A). Prawdopodobieństwo dopełnienia można zapisać następująco [1, s. 79]:



Powyższa reguła będzie stosowana przy omawianiu rozkładów prawdopodobieństwa (zob. Charakterystyka wybranych rozkładów prawdopodobieństwa).

Przykład. Należy obliczyć prawdopodobieństwo tego, że losowo wybrany wniosek o dotację UE został prawidłowo wypełniony, wiedząc, że co ósmy zawiera błędy. Oznaczamy:

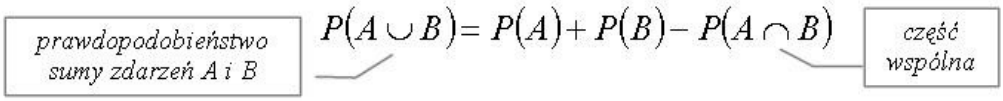
P(A) – prawdopodobieństwo tego, że wniosek został źle wypełniony.

Podstawiamy do wzoru:

$$P(\overline{A}) = 1 - P(A) = 1 - \frac{1}{8} = \frac{7}{8}$$

Zatem prawdopodobieństwo prawidłowego wypełnienia wniosku wynosi 7/8.

Następną ważną regułą w algebrze zdarzeń jest tzw. *reguła sumowania*. Prawdopodobieństwo sumy dwóch zdarzeń można przedstawić następująco [1, s. 79]:



Warto tu wskazać na przypadek szczególny, jakim są zdarzenia wykluczające się wzajemnie. W tej sytuacji brak jest części wspólnej:

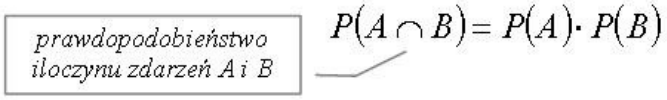
$$P(A \cap B) = 0$$

stąd:

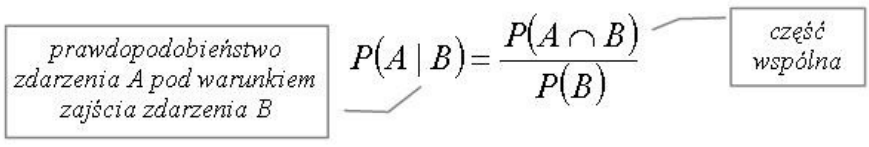
$$P(A \cup B) = P(A) + P(B)$$

W rachunku prawdopodobieństwa istotny jest podział zdarzeń losowych na:

1. **Zdarzenia niezależne** – zajście jednego z tych zdarzeń nie ma wpływu na prawdopodobieństwo zajścia drugiego z nich. Oto warunek niezależności zdarzeń:



2. **Zdarzenia zależne** – prawdopodobieństwo zajścia zdarzenia A zależy od zajścia zdarzenia B. Można tu mówić o tzw. *prawdopodobieństwie warunkowym* zdarzenia A przy założeniu, że zaszło zdarzenie B:

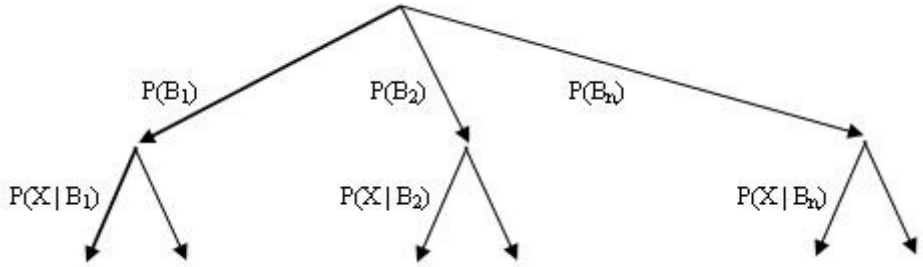


Z powyższego równania można wyprowadzić wzór na iloczyn zdarzeń A i B:

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

W przypadku gdy zdarzenia są zależne – warto posłużyć się tzw. *drzewem stochastycznym*:

Rysunek 3.2. Drzewo stochastyczne.



Źródło: Opracowanie własne.

Zdarzenia na poszczególnych „gałęziach" drzewa są parami przeciwstawne, stąd np.:

$$P(B_1) + P(B_2) + \ldots + P(B_n) = 1$$

Na podstawie powyższego schematu można wyprowadzić ogólny wzór na prawdopodobieństwo całkowite:

prawdopodobieństwo zajścia zdarzenia X

$$P(X) = \sum_{i=1}^{n} P(B_i) \cdot P(X \mid B_i)$$

prawdopodobieństwo zdarzenia X pod warunkiem zajścia zdarzenia B_i

Mając obliczone prawdopodobieństwo zajścia zdarzenia X – można skorzystać z tzw. *wzoru Bayesa*:

prawdopodobieństwo zajścia zdarzenia B_i przy założeniu, że zaszło zdarzenie X

$$P(B_i \mid X) = \frac{P(B_i) \cdot P(X \mid B_i)}{P(X)}$$

prawdopodobieństwo zdarzenia X pod warunkiem zajścia zdarzenia B_i

prawdopodobieństwo zajścia zdarzenia X

Wzór ten pozwala na wyznaczenie prawdopodobieństw zdarzeń B_i, gdy wiemy, że zaszło zdarzenie X.

Przykład. Prawdopodobieństwo zdania egzaminu ze statystyki w pierwszym terminie uzależnione jest od tego, czy student korzysta z dodatkowych form nauczania. Z badań przeprowadzonych wśród wybranej grupy studentów wynika, iż czterech na dziesięciu studentów skorzystało z dodatkowych form nauczania. Wśród tej grupy osób aż 70 proc. zdało egzamin w pierwszym terminie. Natomiast egzamin w pierwszym terminie zdał tylko co drugi student niekorzystający z dodatkowych form nauczania. Należy obliczyć:

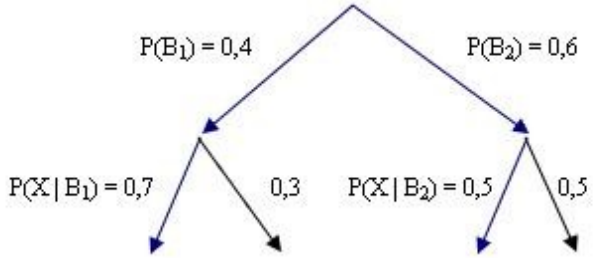
a) prawdopodobieństwo zdania egzaminu ze statystyki w pierwszym terminie,
b) prawdopodobieństwo, że losowo wybrany student, który zdał egzamin w pierwszym terminie korzystał z dodatkowych form nauczania.

Wprowadzamy następujące oznaczenia:

P(X) – prawdopodobieństwo zdania egzaminu ze statystyki w pierwszym terminie,
$P(B_1)$ – prawdopodobieństwo, że student korzystał z dodatkowych form nauczania,
$P(B_2)$ – prawdopodobieństwo, że student nie korzystał z dodatkowych form nauczania.

Dane przedstawiono na drzewie stochastycznym:

Rysunek 3.3. Drzewo stochastyczne – przykład liczbowy.



Źródło: Dane umowne.

a) obliczamy prawdopodobieństwo całkowite:

$$P(X) = P(B_1) \cdot P(X \mid B_1) + P(B_2) \cdot P(X \mid B_2) = 0{,}4 \cdot 0{,}7 + 0{,}6 \cdot 0{,}5 = 0{,}28 + 0{,}3 = 0{,}58$$

b) korzystamy ze wzoru Bayesa:

$$P(B_1 \mid X) = \frac{P(B_1) \cdot P(X \mid B_1)}{P(X)} = \frac{0{,}4 \cdot 0{,}7}{0{,}58} = \frac{28}{58} = 0{,}483$$

Prawdopodobieństwo zdania egzaminu w pierwszym terminie wynosi 58 proc. Prawdopodobieństwo, że losowo wybrany student, który zdał egzamin w pierwszym terminie, korzystał z dodatkowych form nauczania wynosi 48,3 proc.

To, czy zdarzenia są od siebie zależne, czy też nie, będzie miało wpływ na wybór rozkładu prawdopodobieństwa, a także na dobór niektórych testów statystycznych.

Opis struktury zbiorowości dotyczył empirycznych rozkładów cech jakościowych i ilościowych. W przypadku teoretycznych rozkładów prawdopodobieństwa można mówić o tzw. zmiennej losowej. Mianem **zmiennej losowej** określa się „każdą jednoznacznie określoną funkcję rzeczywistą wy-

znaczoną na zbiorze zdarzeń elementarnych" [9, s. 88]. Zmienne losowe dzielą się na (por. [8, s. 30]):

1. Skokowe (por. cecha skokowa) – w przypadku zmiennych losowych skokowych (dyskretnych) można mówić o *rozkładzie masy prawdopodobieństwa*:

$$P(X = x_i) = p_i$$

2. Ciągłe (por. cecha ciągła i *quasi*-ciągła) – w przypadku zmiennych losowych ciągłych mówimy o tzw. *rozkładzie gęstości prawdopodobieństwa*:

$$P(a < X < b) = \int_a^b f(x)\,dx = p_i$$

Teoretyczne rozkłady prawdopodobieństwa posiadają syntetyczne charakterystyki (por. [8, s. 35]):

- wartość oczekiwana (por. średnia arytmetyczna),
- wariancja bądź odchylenie standardowe (pierwiastek kwadratowy z wariancji).

Sposób obliczania wymienionych charakterystyk zawiera tabela:

Tabela 3.1. Podstawowe charakterystyki rozkładów zmiennych losowych.

	Zmienne losowe skokowe	**Zmienne losowe ciągłe**
Wartość oczekiwana	$E(X) = m = \sum_{i=1}^{k} x_i p_i$	$E(X) = m = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx$
Wariancja	$D^2(X) = \sigma^2 = \sum_{i=1}^{k} (x_i - m)^2 \cdot p_i$	$D^2(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - m)^2 \cdot f(x)\,dx$

Źródło: Opracowanie własne na podstawie: [8, s. 35].

W kolejnym podrozdziale omówiono wybrane rozkłady skokowe i ciągłe. Należy zaznaczyć, iż charakterystyki są obliczane nie ze wzorów prezentowanych w tabeli 3.1, lecz ze wzorów uproszczonych.

3.2. Charakterystyka wybranych rozkładów prawdopodobieństwa

W niniejszym podrozdziale omówiono wybrane rozkłady prawdopodobieństwa. Obliczeń można dokonać w załączonym dodatku *Rozkłady prawdopodobieństwa*. W tym podrozdziale położono nacisk na odpowiedni wybór rozkładu, a także na umiejętność odczytu żądanych wartości z tablic statystycznych. Oto klasyfikacja omówionych w dalszej części rozkładów prawdopodobieństwa:

Tabela 3.2. Klasyfikacja rozkładów prawdopodobieństwa.

	Rozkłady skokowe	**Rozkłady ciągłe**
Zmienne niezależne	1. Rozkład dwumianowy. 2. Rozkład dwupunktowy. 3. Rozkład geometryczny. 4. Rozkład Poissona.	1. Rozkład jednostajny. 2. Rozkład normalny. 3. Rozkład T-Studenta. 4. Rozkład Chi-kwadrat. 5. Rozkład F.
Zmienne zależne	1. Rozkład hipergeometryczny	

Źródło: Opracowanie własne.

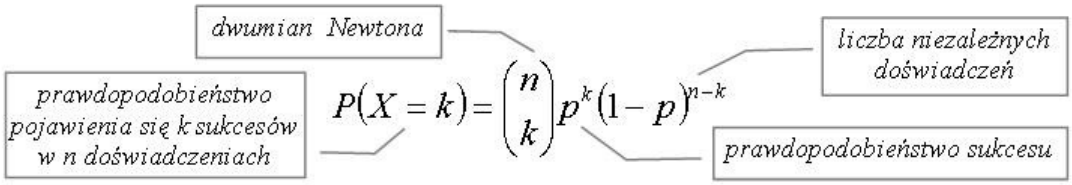
3.2.1. Rozkład dwumianowy

Rozkład dwumianowy (Bernoulliego) zmiennej losowej X znajduje zastosowanie wówczas, gdy (por. [21, s. 195]):

1. Przeprowadza się *n* jednakowych doświadczeń.
2. Dla każdego doświadczenia możliwe są dwa wyniki: sukces lub porażka.

3. Prawdopodobieństwo sukcesu *p* w kolejnych doświadczeniach nie zmienia się (doświadczenia niezależne).
4. Liczba doświadczeń *n* jest niewielka (zał. n < 30).

Funkcja prawdopodobieństwa rozkładu dwumianowego jest następująca:



Dwumian Newtona oblicza się według wzoru:

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

Oto podstawowe charakterystyki rozkładu:

a) wartość oczekiwana:

$$m = np$$

b) odchylenie standardowe:

$$\sigma = \sqrt{np(1-p)}$$

Dystrybuantą zmiennej losowej X o rozkładzie dwumianowym jest funkcja postaci (por. [9, s. 95]):

$$F(k) = P(X \leq k)$$

Analogicznie można określić dystrybuantę dla pozostałych rozkładów skokowych.

Przykład. Student na „chybił-trafił" rozwiązuje test wielokrotnego wyboru ze statystyki, gdzie tylko jedna spośród czterech opcji odpowiedzi jest prawidłowa. Test liczy 10 pytań. Proszę obliczyć prawdopodobieństwo tego, że ponad 40 proc. odpowiedzi będzie prawidłowych. Wypisujemy dane:

a) liczba sukcesów polegających na właściwym zaznaczeniu odpowiedzi: $P(X > 4)$,
b) liczba niezależnych prób (pytań w teście): $n = 10$,
c) prawdopodobieństwo sukcesu: $p = 0{,}25$.

Możemy skorzystać ze wzoru na prawdopodobieństwo dopełnienia zdarzeń:

$$P(X > 4) = 1 - P(X \leq 4) = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)]$$

Następnie obliczamy prawdopodobieństwa cząstkowe ze wzoru na funkcję prawdopodobieństwa rozkładu dwumianowego. Oto sposób obliczeń dla $k = 0$:

$$P(X = 0) = \binom{10}{0} \cdot (0{,}25)^0 \cdot (1 - 0{,}25)^{10-0}$$

$$\binom{10}{0} = \frac{10!}{0!(10-0)!} = \frac{10!}{1 \cdot 10!} = 1$$

Wracamy do wzoru:

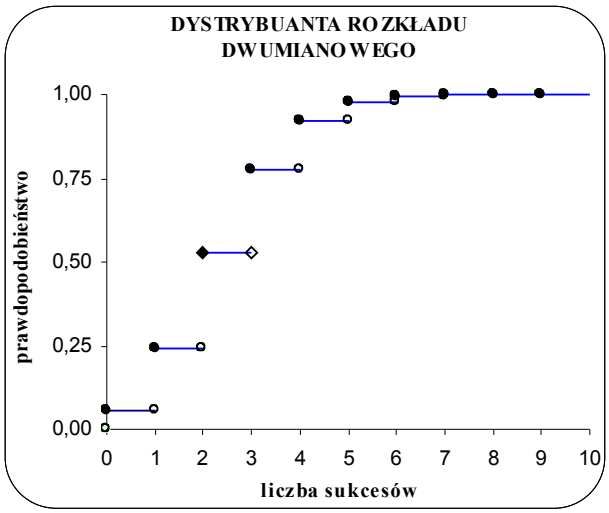
$$P(X = 0) = 1 \cdot 1 \cdot (1 - 0{,}25)^{10} = (0{,}75)^{10} = 0{,}0563$$

Analogicznie obliczamy prawdopodobieństwa dla k = 1, k = 2, k = 3 i k = 4. Suma prawdopodobieństw cząstkowych to:

$$P(X \leq 4) = 0{,}9219$$

Powyższe prawdopodobieństwo jest wartością dystrybuanty rozkładu dwumianowego w punkcie 4. Oto wykres dystrybuanty tego rozkładu:

Rysunek 3.4. Dystrybuanta rozkładu dwumianowego.



Źródło: Opracowanie własne.

Prawdopodobieństwo tego, że student poprawnie wskaże ponad 40 proc. odpowiedzi, wynosi (przy założeniu, że za dane pytanie jest zero punktów lub jeden punkt):

$$P(X > 4) = 1 - P(X \leq 4) = 1 - 0{,}9219 = 0{,}0781$$

Jedynie ośmiu studentów na stu uzyska ponad 40 proc. poprawnych odpowiedzi zakreślając odpowiedzi na „chybił-trafił”.

Szczególnym przypadkiem rozkładu dwumianowego jest **rozkład dwupunktowy** (zerojedynkowy). W tej sytuacji ma miejsce:

a) prawdopodobieństwo sukcesu:

$$P(X=1)=p$$

b) prawdopodobieństwo porażki:

$$P(X=0)=1-p=q$$

Charakterystyki tego rozkładu są następujące:

a) wartość oczekiwana:

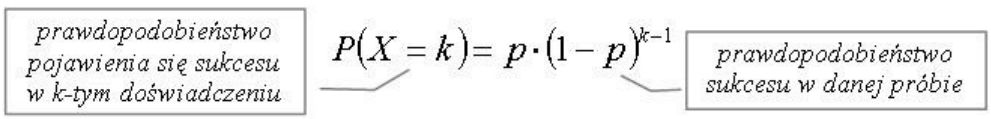
$$m=p$$

b) odchylenie standardowe:

$$\sigma=\sqrt{p\cdot(1-p)}$$

Nawiązując do powyższego przykładu: możemy stwierdzić, że prawdopodobieństwo sukcesu, jakim jest losowy wybór prawidłowej opcji odpowiedzi wynosi 0,25. Jednocześnie prawdopodobieństwo porażki, tj. zaznaczenia nieprawidłowej odpowiedzi, wynosi 0,75.

O ile rozkład dwumianowy określa liczbę *k* sukcesów wśród *n* powtórzeń doświadczenia (np. *n* rzutów monetą), o tyle **rozkład geometryczny** wyznacza prawdopodobieństwo pojawienia się pierwszego sukcesu:



Charakterystyki:

a) wartość oczekiwana:

$$m = \frac{1}{p}$$

b) odchylenie standardowe:

$$\sigma = \sqrt{\frac{1-p}{p^2}}$$

Przykład. Średnio rzecz biorąc, co piąty internauta odwiedzający pewien sklep internetowy robi w nim zakupy. Należy obliczyć prawdopodobieństwo tego, że pierwsza transakcja pojawi się przy trzecim wejściu na stronę. Ile powinno być wejść na stronę, aby została dokonana transakcja kupna-sprzedaży?:

Wypisujemy dane:

p = 0,2 (co piąty internauta)
k = 3

Podstawiamy do wzoru na funkcję prawdopodobieństwa rozkładu geometrycznego:

$$P(X=3) = 0{,}2 \cdot (1-0{,}2)^{3-1} = 0{,}2 \cdot (0{,}8)^2 = 0{,}128$$

Prawdopodobieństwo tego, że pierwsza transakcja zostanie zawarta po trzecim wejściu na stronę, wynosi 12,8 proc.

Aby odpowiedzieć na pytanie, ile powinno być średnio wejść na stronę, by została dokonana transakcja kupna-sprzedaży, obliczamy wartość oczekiwaną:

$$E(X) = \frac{1}{p} = \frac{1}{0{,}5} = 5$$

Należy oczekiwać, iż średnio przy pięciu wejściach na stronę zostanie zakupiony jakiś produkt ze sklepu internetowego. Oczywiście pierwszy internauta może od razu nabyć pewien produkt, ale też może zdarzyć się sytuacja, w której nawet pięć wejść nie gwarantuje zbytu produktów. Warto więc obliczyć jeszcze odchylenie standardowe:

$$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0{,}2}{(0{,}2)^2}} = 4{,}47$$

Górną granicę typowego obszaru zmienności uzyskamy, dodając do wartości oczekiwanej obliczone powyżej odchylenie standardowe. Zatem o nietypowej sytuacji możemy mówić w przypadku, gdy na stronę wejdzie więcej niż dziewięciu internautów i nie zostanie zawarta transakcja kupna-sprzedaży.

Dlaczego warto mieć pełną wersję?

Pełną wersję książki zamówisz na stronie wydawnictwa Złote Myśli

http://statystyka.zlotemysli.pl