PODSTAWY STATYSTYKI OPISOWEJ

MATERIAŁY PRZYGOTOWAWCZE DO UDZIAŁU W EUROPEJSKIM KONKURSIE STATYSTYCZNYM

CZ. I. WPROWADZENIE DO STATYSTYKI OPISOWEJ

100 lat Główny Urząd Statystyczny

European Statistics Competition

1. Wprowadzenie do statystyki opisowej

Termin statystyka współcześnie ma kilka znaczeń:

- zbiór danych liczbowych, przedstawiających kształtowanie się określonych zjawisk i procesów,
- wszelkie prace związane z gromadzeniem i opracowywaniem danych liczbowych,
- charakterystyki opisowe obliczane ze zbiorowości próbnych np. średnia arytmetyczna,
- dyscyplina naukowa mająca własne metody badawcze – nauka o ilościowych metodach badania prawidłowości występujących w zjawiskach masowych scharakteryzowanych za pomocą liczb.

1.2. Podstawowe pojęcia

Zbiorowość statystyczna (populacja generalna) jest to zbiór dowolnych elementów podobnych pod względem określonych właściwości i poddanych badaniu statystycznemu. Przykład: ludność Polski

Próba statystyczna – podzbiór populacji generalnej. Przykład: ludność województwa mazowieckiego.

Jednostka statystyczna – to element (obiekt) zbiorowości statystycznej. Przykład: jedna osoba.

Cecha statystyczna jest to właściwość obiektu tworzącego zbiorowość statystyczną.

Cechy statystyczne dzielą się na:

- **mierzalne** (ilościowe) – podawane liczbowo, przykład: wiek,
- **niemierzalne** (jakościowe) – podawane opisowo, przykład: płeć.

Cechy **ilościowe** można podzielić na:

- **skokowe** (dyskretne) - przyjmują wartości ze skończonych i przeliczalnych przedziałów liczbowych, ale bez wartości pośrednich, przykład: liczba gospodarstw domowych.
- **ciągłe** – przyjmują każdą wartość z określonego przedziału, przykład: wzrost.

1.3. Skale pomiarowe

Skale pomiarowe:

- nominalna,
- porządkowa (rangowa),
- przedziałowa (interwałowa),
- ilorazowa (stosunkowa).

Skale nominalna i porządkowa są skalami jakościowymi. Skale przedziałowa i ilorazowa są skalami ilościowymi.

Skala nominalna jest skalą najniższego poziomu. Liczby w tej skali pełnią jedynie rolę etykiet. O dwóch wartościach zebranych na tej skali można powiedzieć tylko tyle, że albo sa takie same albo się różnią i nie ma relacji porządku między nimi:

$$a = b,$$

$$a \neq b.$$

Szczególnym przypadkiem jest pomiar binarny, który przyjmuje wartości 0 lub 1.

Na wartościach w skali nominalnej można wykonywać: zliczanie, obliczanie częstości występowania, wskazanie klasy najliczniejszej.

Przykłady: PESEL, płeć, grupa krwi.

Skala porządkowa (rangowa) zawiera wszystkie cechy skali nomionalnej, dochodzi informacja o uporządkowaniu wartości (rosnącym albo malejącym) ale nie sa znane odległości pomiędzy pomiędzy wartościami:

$$a = b,$$

$$a \neq b,$$

$$a < b \text{ albo } a > b,$$

$$a \leq b \text{ albo } a \geq b.$$

Na wartościach w skali porządkowej można wykonywać: zliczanie, obliczanie częstości występowania, wskazanie klasy najliczniejszej, uporządkować wartości.

Od skali porządkowej można przejść do skali nominalnej.

Przykłady: stopnie wojskowe i inne, oceny w szkole, poziomy wykształcenia.

Skala przedziałowa (interwałowa) zawiera cechy skali nominalnej i porządkowej, dochodzi informacja o zerze „umownym" („sztucznym"). Można wyznaczyć różnice pomiędzy wartościami – można stwierdzić o ile jedna wartość jest mniejsza (większa od drugiej) ale nie można wskazać ile razy.

Przykłady: temperatura w skali Celsjusza, rok urodzenia.

Na wartościach w skali przedziałowej można wykonywać wszystkie operacje wykonywane na poprzednich skalach oraz można obliczać średnie i miary zróżnicowania.

Skala stosunkowa (ilorazowa) zawiera wszystkie cechy niższych skal z tym, że zawiera zero naturalne.

Przykłady: wiek, waga ciała, przejechane kilometry.

Na wartościach w skali stosunkowej można wykonywać wszystkie operacje arytmetyczne.

Zadanie 1.1.

Na podstawie danych zawartych w tablicy 1 określ typ skali do poszczególnych zmiennych.

Tablica 1. Zestawienie danych dotyczące żołnierzy

Lp.	*Imię i nazwisko*	*Stopień wojskowy*	*Kod stopnia*	*Poziom wykształcenia*	*Temperatura ciała*	*Wzrost*	*Waga*	*Lata służby*	*Wiek*	*Kolor włosów*	*Płeć*
1	*Żołnierz1*	*kapitan*	*14*	*wyższe*	*36,3*	*182*	*85*	*10*	*38*	*czarne*	*M*
2	*Żołnierz2*	*major*	*15*	*wyższe*	*36,6*	*178*	*87*	*15*	*43*	*rude*	*M*
3	*Żołnierz3*	*kapral*	*3*	*wyższe*	*36,7*	*185*	*92*	*6*	*27*	*brązowe*	*M*
4	*Żołnierz4*	*pułkownik*	*17*	*wyższe*	*36,5*	*192*	*98*	*25*	*50*	*siwe*	*M*
5	*Żołnierz5*	*porucznik*	*13*	*wyższe*	*36,9*	*182*	*80*	*7*	*26*	*czarne*	*K*
6	*Żołnierz6*	*chorąży*	*9*	*średnie*	*36,2*	*186*	*86*	*7*	*28*	*brązowe*	*M*
7	*Żołnierz7*	*sierżant*	*6*	*średnie*	*37,2*	*196*	*95*	*18*	*45*	*brązowe*	*M*
8	*Żołnierz8*	*kapitan*	*14*	*wyższe*	*37,5*	*176*	*79*	*9*	*40*	*blond*	*K*
9	*Żołnierz9*	*podpułkownik*	*16*	*wyższe*	*36,7*	*185*	*89*	*16*	*50*	*czarne*	*M*
10	*Żołnierz10*	*podporucznik*	*12*	*wyższe*	*36,9*	*189*	*97*	*5*	*26*	*brązowe*	*K*

Źródło: Dane umowne.

Rozwiązanie 1.1:
Skala nominalna: imię i nazwisko, kolor włosów, płeć.

Skala porządkowa: stopień wojskowy, kod stopnia, poziom wykształcenia.

Skala interwałowa: temperatura ciała.

Skala ilorazowa: wzrost, waga, lata służby, wiek.

1.4. Rodzaje badań statystycznych

Rozróżnia się dwa zasadnicze typy badań statystycznych:
1. Badania pełne.
2. Badania częściowe.

Badanie **pełne** obejmuje wszystkie elementy zbiorowości generalnej.

Badanie **częściowe** obejmuje pewną część populacji generalnej tj. podzbiór elementów populacji generalnej określany mianem próby.

W praktyce szersze zastosowanie mają badania częściowe. Powody są następujące:
1. W przypadku populacji nieskończonych, nie istnieje możliwość przeprowadzenia badania pełnego.
2. W przypadku skończonej ale bardzo licznej populacji koszt badania pełnego byłby bardzo wysoki, a czas realizacji bardzo długi.

1.5. Etapy procesu badawczego

1.5.1. Projektowanie badania

Etap projektowania badania stanowi krok przygotowawczy badania. W tym etapie należy ustalić:

1. Cel badania – każde badanie statystyczne musi być podporządkowane konkretnemu celowi. Właściwie ustalony cel stanowi warunek niedopuszczenia do otrzymania zbędnego lub przypadkowego zbioru danych statystycznych.
2. Hipotezy badawcze – określa się je na ogół w badaniach naukowych, hipotezy muszą być możliwe do zweryfikowania, nie mogą dotyczyć np. przyszłości.
3. Zakres podmiotowy badania – należy określić jakich podmiotów dotyczy badanie, np. przedsiębiorstw, gospodarstw domowych, grupy osób.
4. Zakres przedmiotowy badania – należy określić jakie cechy statystyczne będą mierzone w badaniu, np. płeć, wiek, liczba pracujących, itd.
5. Metoda badawcza – określenie metod badawczych, np. metoda ankietowa.

1.5.2. Obserwacja statystyczna

Obserwacja statystyczna polega na ustaleniu wartości cech statycznych za pomocą pomiaru na skalach pomiarowych. Zbiór danych uzyskanych w wyniku obserwacji statystycznej stanowi materiał statystyczny. Ze względu na źródło pochodzenia materiał statystyczny można podzielić na **pierwotny** i **wtórny**. Pierwotny materiał statystyczny stanowią dane zebrane dla realizacji celu(ów) danego badania. Materiał wtórny natomiast tworzą dane zebrane dla realizacji celów innego badania, ale mogą zostać wykorzystane do realizacji celów danego badania.

Zebrany materiał statystyczny stanowi surowy materiał statystyczny. Z reguły jest on obarczony błędami: **systematycznymi** oraz **losowymi**. Błędy systematyczne wynikają z jednokierunkowych tendencji do zniekształcania rzeczywistości. Błędy losowe (przypadkowe) mogą wynikać z nieświadomości, braku wiedzy,

pobłażliwego podejścia do przekazywania danych, itd. Błędy systematyczne z reguły kumulują się w wynikowej informacji statystycznej. Błędy przypadkowe, co do zasady, powinny się znosić.

Surowy materiał statystyczny może zawierać **błędy logiczne** oraz **rachunkowe**. Błędy logiczne widoczne są w danych sprzecznych z przepisami prawa, np. wynagrodzenie mniejsze od minimalnego. Błędy rachunkowe wynikają na ogół z niewłaściwego zastosowania operacji arytmetycznych (np. sumowania).

W automatycznej kontroli danych sosuje się **błędy twarde** – błędy, które muszą zostać poprawione oraz **błędy uznaniowe** – sytuacje, w których algorytmy kontrolne sygnalizują odstępstwo od założonych wartości, ale dane mogą być poprawne.

1.5.3. Opracowanie i prezentacja materiału statystycznego

Opracowanie materiału statystycznego polega w uproszczeniu na przekształceniu danych statystycznych na informacje. Można tego dokonać poprzez grupowanie danych – polega na podziale zbiorowości na możliwie jednorodne grupy zgodnie z przyjętymi kryteriami. Jeżeli dokonuje się grupowania ze względu na jedną cechę to jest to grupowania proste, a gdy stosujemy więcej cech to jest to grupowanie złożone. W wyniku grupowania statystycznego tworzone są szeregi statystyczne (w przypadku jednej cechy) oraz tablice statystyczne (w przypadku co najmniej dwóch cech).

Szereg statystyczny jest to ciąg wariantów cechy uporządkowanych rosnąco lub malejąco, pogrupowany według określonych kryteriów. Szeregi dzielimy na:

- **wyliczające (szczegółowe)** – uporządkowane rosnąco lub malejąco wartości cechy. Szeregi te są niewygodne w wykrywaniu prawidłowości statystycznych z uwagi na to, że na ogół są długie,
- **rozdzielcze (strukturalne)** – zawierają ciągi wartości badanej cechy wraz z przypisanymi im liczebnościami.

Szeregi rozdzielcze cechy mierzalnej dzielą się na punktowe i przedziałowe. Szeregi punktowe buduje się dla cech skokowych a rozdzielcze przedziałowe dla cech ciągłych.

Szereg przedziałowy powstaje z pogrupowania wartości szeregu szczegółowego w pewną liczbę przedziałów zwanych klasami. Każdemu przedziałowi przypisana jest liczba zaliczających się do niego elementów (liczebność przedziału klasowego). Każdy przedział ograniczony jest granicami: dolnąi górną. Różnica między górną i dolną granicą przedziału nosi nazwę interwału (rozpiętości przedziału).

Kluczowym zagadnieniem podczas konstrukcji szeregów rozdzielczych przedziałowych jest określenie liczby przedziałów klasowych oraz ich rozpiętości. Na ogół przyjmuje się, że rozpiętości muszą być takie same we wszystkich przedziałach klasowych. Ogólnie można stwierdzić, że liczba klas zależy od różnicy między maksymalną (x_{max}) i minimalną (x_{min}) wartością w szeregu oraz liczebności zbiorowości (N).

Podstawowe wzory wykorzystywane do konstrukcji szeregów rozdzielczych przedziałowych:

- wyznaczanie liczby przedziałów klasowych:

$$k = \sqrt{N}$$
$$k \leq 1 + 3{,}222 log N$$
$$k \leq 5 log N$$

- rozpiętość przedziału: $h = \frac{x_{max} - x_{min}}{k}$

Zadanie 1.2.

Dane są obserwacje dotyczące zysku pewnego sklepu odzieżowego (w tys. zł) w 20 kolejnych dniach roboczych: 17,4; 12,5; 14,4; 13,0; 14,4; 14,4; 13,9; 11,1; 16,5; 15,8; 13,0; 15,8; 18,5; 10,5; 17,4; 15,7;

16,5; 16,9; 11,1; 16,5.

Przedstaw obserwacje statystyczne w postaci: szeregu wyliczającego, szeregu rozdzielczego punktowego, szeregu rozdzielczego przedziałowego.

Rozwiązanie 1.2:

Tablica 2. Szereg wyliczający (prosty, szczegółowy)

10,5
11,1
11,1
12,5
13,0
13,0
13,9
14,4
14,4
14,4
15,7
15,8
15,8
16,5
16,5
16,5
16,9
17,4
17,4
18,5

Tablica 3. Szereg rozdzielczy punktowy

Punkty	*Liczebności*
10,5	*1*
11,1	*2*
12,5	*1*
13,0	*2*
13,9	*1*
14,4	*3*
15,7	*1*
15,8	*2*
16,5	*3*
16,9	*1*
17,4	*2*
18,5	*1*

Szereg rozdzielczy przedziałowy:

Liczba przedziałów:

$$k = \sqrt{N};\ k = 4{,}472135955$$

$$k \leq 1 + 3{,}222\ logN;\ k = 5{,}191918646$$

$$k \leq 5logN;\ k = 6{,}505149978$$

- Wariant $k = 5$

$$x_{min} = 10{,}5$$

$$x_{max} = 18{,}5$$

$$h = \frac{x_{max} - x_{min}}{k} = 1{,}6$$

Tablica 4. Szereg rozdzielczy przedziałowy. Wariant $k = 5$

i	*x0i*	*x1i*	*ni*
1	*10,5*	*12,1*	*3*
2	*12,1*	*13,7*	*3*
3	*13,7*	*15,3*	*4*
4	*15,3*	*16,9*	*7*
5	*16,9*	*18,5*	*3*

- Wariant $k = 4$

$$x_{min} = 10{,}5$$

$$x_{max} = 18{,}5$$

$$h = \frac{x_{max} - x_{min}}{k} = 2$$

Tablica 5. Szereg rozdzielczy przedziałowy. Wariant $k = 4$

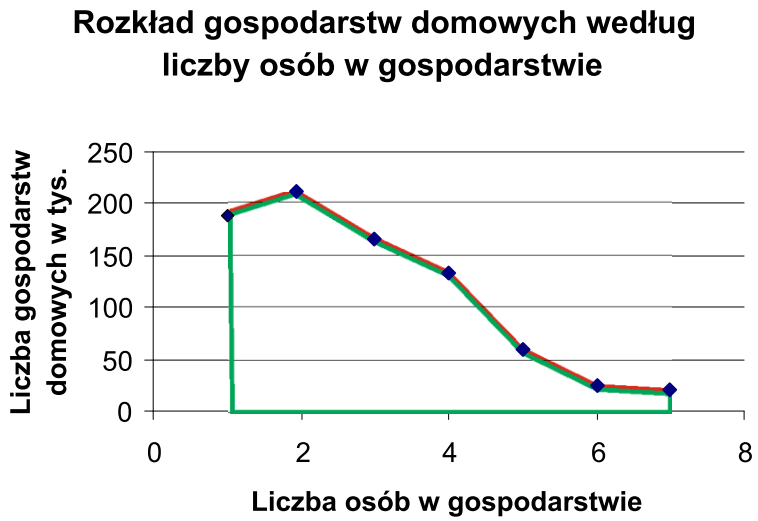
i	*x0i*	*x1i*	*ni*
1	*10,5*	*12,5*	*4*
2	*12,5*	*14,5*	*6*
3	*14,5*	*16,5*	*6*
4	*16,5*	*18,5*	*4*

1.5.4. Prezentacja graficzna szeregów strukturalnych i czasowych

Ważnym narzędziem uwidaczniania prawidłowości występujących w zbiorowości statystycznej jest prezentacja graficzna materiału statystycznego w postaci wykresów. Do najczęściej stosowanych graficznych form prezentacji zalicza się wykresy: punktowe, liniowe i powierzchniowe. Dobór formy uzależniony jest od rodzaju szeregu, w których zapisane są dane oraz charakteru prawidłowości, które wykres ma pokazywać. Prezentacji graficznej szeregów strukturalnych z cechą mierzalną i szeregów czasowych dokonuje się w prostokątnym układzie współrzędnych.

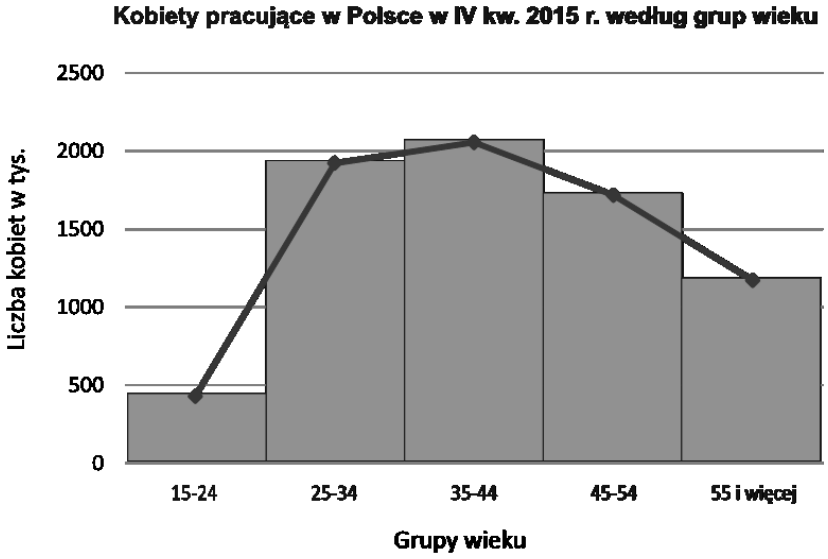
W przypadku graficznej prezentacji szeregów strukturalnych z cechą ilościową na osi poziomej - OX odkładamy wartości cechy (szereg punktowy) lub środki przedziałów (szereg przedziałowy), a na osi pionowej - OY liczebności cząstkowe.

Sam szereg rozdzielczy punktowy przedstawiamy za pomocą wykresu punktowego bądź wykresu liniowego. Wykres punktowy ma postać punktów, z których każdy prezentuje określoną liczbę jednostek zbiorowości, posiadających ten sam wariant cechy ilościowej. Jeżeli te punkty połączymy to otrzymamy **diagram**. Z kolei jeżeli diagram domkniemy liniami prostopadłymi do osi poziomej poprowadzonymi przez najniższą i najwyższą wartość cechy to otrzymamy **wielobok liczebności** (rys. 1).



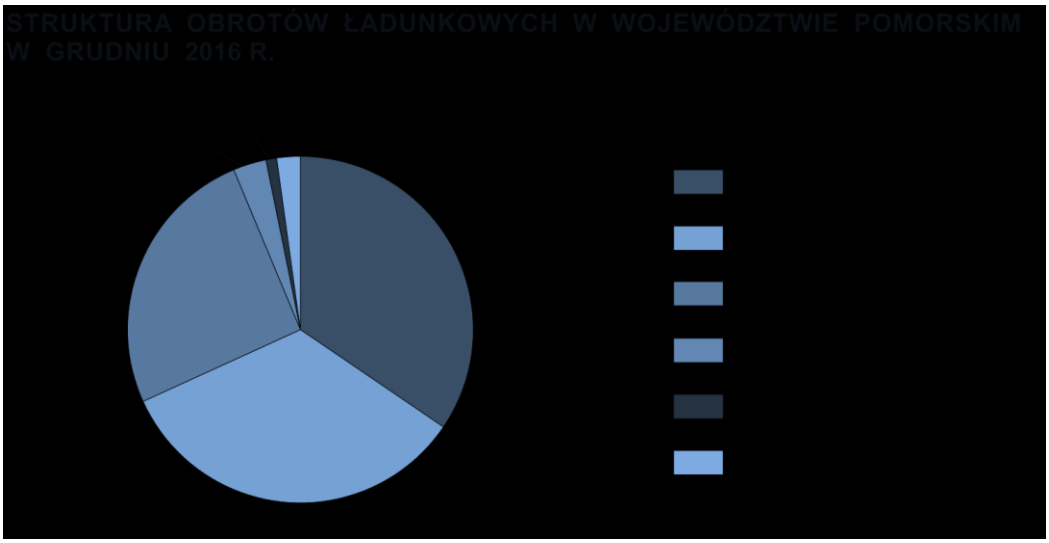
Rysunek 1. Wykres prezentujący szereg rozdzielczy punktowy

Natomiast do prezentacji szeregów rozdzielczych przedziałowych zastosowanie ma wykres powierzchniowy zwany **histogramem.** Stanowią go pola przystających do siebie prostokątów (słupków), przy czym jeden z boków każdego prostokąta ma długość odpowiadającą rozpiętości, a drugi liczebności przedziału klasowego. Można również ten szereg przedstawić za pomocą wykresu liniowego zwanego **krzywą liczebności** (rys. 2). Krzywa powstaje z połączenia punktów, których współrzędnymi są środki przedziałów klasowych i liczebności poszczególnych przedziałów.



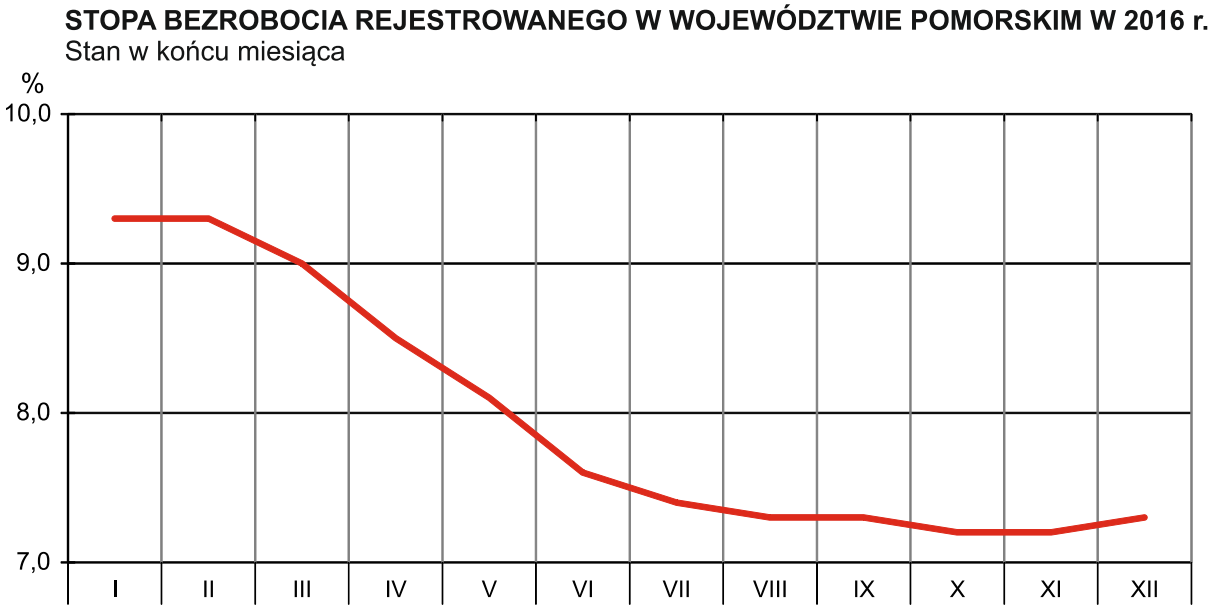
Rysunek 2. Wykres prezentujący szereg rozdzielczy przedziałowy

Do prezentacji graficznej szeregów rozdzielczych z cechą jakościową stosujemy wykresy powierzchniowe sporządzane niezależnie od układu współrzędnych. Wykorzystywane są w tym celu najczęściej pola figur płaskich takich jak koło, kwadrat, prostokąt czy trójkąt równoboczny (rys. 3).



Rysunek 3. Wykres prezentujący szereg rozdzielczy z cechą jakościową.

Do graficznej prezentacji szeregów dynamicznych wykorzystuje się wykresy liniowe, przy czym na osi OX zaznaczamy jednostki czasu, a na osi OY wielkość obserwowanego zjawiska (rys. 4).



Rysunek 4. Wykres prezentujący szereg czasowy

2. Spis ilustracji

3. Spis tablic