PODSTAWY STATYSTYKI OPISOWEJ

MATERIAŁY PRZYGOTOWAWCZE DO UDZIAŁU W EUROPEJSKIM KONKURSIE STATYSTYCZNYM

CZ. III. MIARY ROZPROSZENIA

100 lat Główny Urząd Statystyczny

European Statistics Competition

1. Miary rozproszenia

Wartości średnie nie wystarczają do scharakteryzowania struktury zbiorowości. Badana zbiorowość statystyczna może powiem charakteryzować się różnym stopniem **zmienności (rozproszenia, zróżnicowania, dyspersji)** badanej cechy.

Dyspersją nazywa się zróżnicowanie jednostek zbiorowości ze względu na wartość badanej cechy. Zmienność ocenia się za pomocą wielu miar statystycznych, wśród których wyróżnia się miary klasyczne i pozycyjne. Miary dyspersji dzieli się takżena bezwzględne (absolutne) i względne (relatywne, stosunkowe).

Klasyczne miary zmienności obliczane są na podstawie wszystkich wartości badanej cechy, pozycyjne – na podstawie niektórych (stojących na określonej pozycji) wartości. Do klasycznych miar zmienności zaliczamy odchylenie przeciętne, odchylenie standardowe, wariancję i współczynnik zmienności (obliczany przy użyciu odchylenia standardowego i średniej arytmetycznej). W grupie **pozycyjnych miar zmienności** wyróżnia się: empiryczny obszar zmienności (rozstęp), odchylenie ćwiartkowe oraz pozycyjny współczynnik zmienności.

Bezwzględne miary zmienności (tj. rozstęp, odchylenie ćwiartkowe, wariancja oraz odchylenie standardowe) są wielkościami mianowanymi, posiadającymi miano badanej cechy. **Względną miarą dyspersji** jest współczynnik zmienności, wyrażany w procentach.

1.1. Klasyczne miary rozproszenia

1.1.1. Odchylenie przeciętne

Odchylenie przeciętne jest to średnia arytmetyczna z bezwzględnych wartości odchyleń wartości zmiennej x_i od średniej arytmetycznej. Informuje ono, w jakim stopniu jednostki badanej zbiorowości różnią się od średniej arytmetycznej ze względu na badaną cechę. Odchylenie przeciętne oznaczamy symbolem d_x i obliczamy w następujący sposób:

$$d_x = \frac{1}{N}\sum_{j=1}^{N}\left|x_j - \bar{x}\right| \ \ dla\ szeregów\ wyliczających,$$

$$d_x = \frac{1}{N}\sum_{i=1}^{k}|x_i - \bar{x}|n_i \ \ dla\ szeregów\ rozdzielczych\ punktowych,$$

$$d_x = \frac{1}{N}\sum_{i=1}^{k}|\dot{x}_i - \bar{x}|n_i \ \ dla\ szeregów\ rozdzielczych\ przedziałowych.$$

gdzie:

N – liczebność badanej zbiorowości,

n_i – liczebność jednostek odpowiadająca poszczególnym wariantom zmiennej,

$\bar{x}$ – symbol średniej arytmetycznej,

x_i- warianty cechy mierzalnej,

$\dot{x}_i$ – środki przedziałów.

1.1.2. Wariancja i odchylenie standardowe

Są to najczęściej wykorzystywane miary zróżnicowania. **Wariancja** jest to średnia arytmetyczna kwadratów odchyleń poszczególnych wartości cechy od ich średniej arytmetycznej. Oznaczamy ją symbolem s^2 i obliczamy w następujący sposób:

$$s^2 = \frac{1}{N}\sum_{j=1}^{N}\left(x_j - \bar{x}\right)^2 \ \ dla\ szeregów\ wyliczających,$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \ \ dla\ szeregów\ rozdzielczych\ punktowych,$$
$$s^2 = \frac{1}{N}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 n_i \ \ dla\ szeregów\ rozdzielczych\ przedziałowych.$$

Wariancja jako miara zróżnicowania ma dwie ważne **właściwości**, a mianowicie:

1. jest różnicą między średnią arytmetyczną kwadratów wartości zmiennej i kwadratem jej średniej arytmetycznej, czyli:

$$s^2 = \overline{x^2} - (\bar{x})^2,$$

2. jeżeli badaną zbiorowość podzielimy na k grup, to wariancja ogólna (całej zbiorowości) jest sumą dwóch składników: **wariancji wewnątrzgrupowej** i **wariancji międzygrupowej**:

$$s^2 = \bar{s}_i^2 + s^2(\bar{x}_i) = \frac{\sum_{i=1}^{k} s_i^2 n_i}{N} + \frac{\sum_{i=1}^{k}(x_i - \bar{\bar{x}})^2 n_i}{N},$$

gdzie:

$N = \sum_{i=1}^{k} n_i$, $i = 1, 2, \dots, k$,

s_i^2- wariancja i-tej grupy,

$\bar{x}_i$ – średnia arytmetyczna i-tej grupy,

$\bar{\bar{x}}$ – średnia ogólna (średnia ze średnich wszystkich grup).

Powyższa własność nosi nazwę **równości wariancyjnej**.

Wariancja jest wielkością nieujemną i mianowaną. Jej mianem jest kwadrat jednostki fizycznej, w jakiej mierzona jest badana zmienna. **Im zbiorowość jest bardziej zróżnicowana, tym wyższa jest wartość wariancji**.

Jeśli wartości badanej cechy mierzone są w skali porządkowej (tzn. są rangowane za pomocą liczb naturalnych), to wariancję obliczamy następująco:

$$s^2 = \frac{n^2 - 1}{12}.$$

Wariancja jest trudna do merytorycznej interpretacji. W celu uzyskania miary zmienności o mianie zgodnym z mianem badanej cechy, oblicza się dodatni pierwiastek kwadratowy z wariancji. Otrzymana w ten sposób miara jest określana mianem **odchylenia standardowego**:

$$s = \sqrt{s^2}.$$

Odchylenie standardowe określa, o ile – średnio biorąc – jednostki zbiorowości różnią się od średniej arytmetycznej badanej zmiennej. Im zbiorowość jest bardziej zróżnicowana, tym wariancja, a więc i odchylenie standardowe, są większe.

Odchylenie standardowe jest charakterystyką często wykorzystywaną w badaniach statystycznych. Ma ono następujące właściwości:

- jest wielkością obliczaną na podstawie wszystkich obserwacji w danym szeregu,
- jego wartość nie zmieni się, jeśli liczebności szeregu wyrazimy w liczbach względnych (procentach) dostatecznie dokładnie ustalonych,
- wartość odchylenia standardowego nie zmieni się, jeśli do wszystkich wartości zmiennej w szeregu dodamy pewną stałą liczbę,
- jeśli wszystkie wartości szeregu pomnożymy przez pewną stałą liczbę większą od zera, to odchylenie standardowe również będzie tylokrotnie większe.

Odchylenie standardowe można wykorzystać do budowy **typowego obszaru zmienności badanej cechy**. Typowy obszar zmienności zawarty jest w przedziale:

$$\bar{x} - s \leq x_{typ} \leq \bar{x} + s.$$

Z odchyleniem standardowym wiąże się tzw. **reguła trzech sigm**. Zgodnie z nią, wystąpienie obserwacji o wartości cechy spoza przedziału $(\bar{x} - 3s;\ \bar{x} + 3s)$ jest mało prawdopodobne.

1.1.3. Standaryzacja wartości cechy

Odchylenie standardowe może być wykorzystane do **standaryzacji wartości cechy**. Standaryzacja jest przekształceniem pierwotnym wartości cechy x_i w wartości nowej cechy z_i według wzoru:

$$z_i = \frac{x_i - \bar{x}}{s},$$

gdzie: $\bar{x}$ i s są odpowiednio średnią arytmetyczną i odchyleniem standardowym pierwotnych wartości cechy.

Wartość standaryzowana informuje o tym, o ile odchyleń standardowych pierwotna wartość cechy jest większa lub mniejsza od średniej arytmetycznej. Standaryzacji podlegają wyłącznie cechy ilościowe, gdyż tylko wtedy można obliczyć średnią arytmetyczną i odchylenie standardowe.

Wartości cechy większej od średniej arytmetycznej odpowiada dodatnia wartość standaryzowana $(z_i > 0)$, a wartościom niższym – ujemna wartość standaryzowana $(z_i < 0)$.

Średnia arytmetyczna zbioru danych standaryzowanych wynosi zero, a odchylenie standardowe jest równe jedności. Dane standaryzowane pochodzące z różnych rozkładów mogą być ze sobą porównywalne.

1.1.4. Klasyczny współczynnik zmienności

Przedstawione dotychczas miary nie pozwalają na porównywanie zmienności tej samej cechy w różnych zbiorowościach, czy kilku cech wyrażonych w odpowiednich mianach. W takich przypadkach wykorzystuje się niemianowaną (najczęściej wyrażaną w procentach) miarę zróżnicowania – **współczynnik zmienności**.

Współczynnik zmienności jest ilorazem absolutnej miary zróżnicowania i przeciętnego poziomu wartości cechy. Z uwagi na fakt, że przy analizie rozkładu wartości cechy posługujemy się różnymi miarami dyspersji i przeciętnymi, współczynnik zmienności można obliczyć kilkoma metodami, a mianowicie:

$$V_s = \frac{s}{\bar{x}} * 100,$$

$$V_d = \frac{d_x}{\bar{x}} * 100,$$

gdzie: V jest symbolem współczynnika zmienności, a indeks przy nim informuje o rodzaju bezwzględnej miary dyspersji użytej w obliczeniach

Jeżeli współczynnik zmienności przyjmuje wysokie wartości liczbowe, to fakt ten świadczy o niejednorodności badanej zbiorowości statystycznej. Umownie przyjmuje się, że jeżeli współczynnik V nie przekracza 10%, to cechy wykazują niewielkie zróżnicowanie. Taką zbiorowość uznaje się za jednorodną, co rzutuje na poprawne wyniki analizy statystycznej.

1.2. Pozycyjne miary rozproszenia

1.2.1. Empiryczny obszar zmienności (rozstęp)

Rozstęp jest różnicą między największą i najmniejszą wartością cechy. Oblicza się go ze wzoru:

$$R = x_{max} - x_{min}.$$

Wartość rozstępu zależy jedynie od dwóch skrajnych wielkości: najmniejszej i największej. Brakuje zatem informacji o zróżnicowaniu pozostałych jednostek zbiorowości pod względem badanej cechy. Dlatego też rozstęp stosowany jest głównie w przypadkach, gdy niezbędna jest wstępna orientacja o obszarze zmienności cechy.

Obszar zmienności możemy określić ściśle tylko na podstawie szeregu wyliczającego. Na podstawie szeregu rozdzielczego przedziałowego można jedynie określić jego przybliżoną wartość, jako różnicę między górną granicą ostatniej klasy i dolną granicą klasy pierwszej. Jeżeli jednak szereg rozdzielczy przedziałowy posiada otwarte klasy, to nawet przybliżone określenie obszaru zmienności jest niemożliwe.

Obszar zmienności jest miarą prostą i łatwą do obliczenia. Ma jednak poważną wadę: jego wartość zależy jedynie od dwóch jednostek zbiorowości. Tym samym nie daje informacji, jak dalece różnią się między sobą pozostałe jednostki zbiorowości. Dlatego też obszar zmienności oblicza się zwykle w celu wstępnej orientacji, na jakim „obszarze" rozciągają się wartości badanej zmiennej.

1.2.2. Odchylenie ćwiartkowe

Odchylenie ćwiartkowe oblicza się na podstawie kwartyli. Definiuje się je jako połowę różnicy między kwartylem trzecim Q_3 i pierwszym Q_1, czyli:

$$Q = \frac{Q_3 - Q_1}{2}.$$

Różnicę $Q_3 - Q_1$ określa się mianem **rozstępu kwartylowego** lub **rozstępu międzykwartylowego** i oznaczać będziemy symbolem R_Q.

Odchylenie ćwiartkowe mierzy poziom zróżnicowania jedynie połowy jednostek (50%), pozostałych po odrzuceniu 25% jednostek o wartościach mniejszych od kwartyla pierwszego i 25% jednostek większych od kwartyla trzeciego. Miara ta nie jest więc wrażliwa na skrajne wartości zbioru. Z tego względu nie należy do zbyt często wykorzystywanych miar zmienności.

Jeżeli do opisu tendencji centralnej użyto mediany, a do opisu zmienności odchylenia ćwiartkowego – to możliwe jest określenie **typowego obszaru zmienności badanej cechy**. Obszar ten określa następująca nierówność:

$$Me - Q < x_{typ} < Me + Q.$$

gdzie:

Me – mediana,

Q – odchylenie ćwiartkowe.

Nietypowe w danej zbiorowości są jednostki o wartości niższej od różnicy $Me - Q$ i wyższej od sumy $Me + Q$.

Odchylenie ćwiartkowe jest szczególnie przydatne w analizie statystycznej szeregów rozdzielczych przedziałowych o klasach otwartych. Odchylenie ćwiartkowe interpretuje się jako przeciętne zróżnicowanie badanych jednostek wokół mediany.

Pomiędzy odchyleniami: ćwiartkowym, przeciętnym i standardowym, obliczonymi z tego samego szeregu, zachodzi następująca relacja:

$$Q < d < s.$$

Porównując zatem dyspersję różnych szeregów (mających to samo miano i zbliżony średni poziom cechy), należy dla każdego z nich obliczyć tę samą miarę zróżnicowania.

1.2.3. Pozycyjny współczynnik zmienności

Pozycyjny współczynnik zmienności wylicza się według wzoru:

$$V_Q = \frac{Q}{Me} * 100,$$

$$V_{Q_1,Q_3} = \frac{Q_3 - Q_1}{Q_3 + Q_1} * 100.$$

Zadanie 1.1.

Wykorzystując dane z tablicy poniżej wyznacz miary rozproszenia.

Dochód na osobę w rodzinie w tys. zł	*Liczba rodzin*
do 0,4	*25*
0,4-0,8	*50*
0,8-1,2	*40*
1,2-1,6	*35*
1,6-2	*30*

Źródło: dane umowne

Rozwiązanie 1.1:

Tablica 1. Obliczenia pomocnicze do zad. 1.1.

i	x_{0i}	x_{1i}	$\dot{x}_i$	n_i	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 n_i$	$(\dot{x}_i - \bar{x})^3 n_i$
1	0	0,4	0,2	25	-0,79	0,6241	15,6025	-12,325975
2	0,4	0,8	0,6	50	-0,39	0,1521	7,605	-2,965950
3	0,8	1,2	1	40	0,01	0,0001	0,004	0,000040
4	1,2	1,6	1,4	35	0,41	0,1681	5,8835	2,412235
5	1,6	2	1,8	30	0,81	0,6561	19,683	15,943230
Σ	-	-	-	180	-	-	48,778	3,063580

- Wariancja:

$$s^2 = \frac{1}{N}\sum_{i=1}^{5}(x_i - \bar{x})^2 n_i = \frac{48,778}{180} = 0,271$$

Wariancja nie posiada interpretacji

- Odchylenie standardowe:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{5}(x_i - \bar{x})^2 n_i} = \sqrt{\frac{48,778}{180}} = \sqrt{0,271} = 0,52$$

Odchylenie standardowe informuje o ile przeciętnie wartości badanej cechy różnią się od średniej arytmetycznej – w tym przypadku przeciętne dochody na osobę w rodzinie różnią się średnio od średniej arytmetycznej o 0,52 tys. zł.

- Odchylenie przeciętne:

$$d_x = \frac{1}{N}\sum_{i=1}^{5}|x_i - \bar{x}|n_i = \frac{78,3}{180} = 0,435$$

Odchylenie przeciętne informuje o ile przeciętnie wartości badanej cechy różnią się od średniej arytmetycznej – w tym przypadku przeciętne dochody na osobę w rodzinie różnią się średnio od średniej arytmetycznej o 0,435 tys. zł.

- Klasyczny współczynnik zmienności:

$$V_s = \frac{s}{\bar{x}}100 = \frac{0,52}{0,99} = 52,53\%$$

$$V_d = \frac{d_x}{\bar{x}}100 = \frac{0,435}{0,99} = 43,94\%$$

Klasyczny współczynnik zmienności informuje jaką część średniej arytmetycznej stanowi odchylenie standardowe.

- Rozstęp kwartylowy:

$$R_Q = Q_3 - Q_1 = 1,43 - 0,56 = 0,87$$

Połowa środkowych obserwacji mieści się w przedziale o rozpiętości 0,87.

- Odchylenie ćwiartkowe:

$$Q = \frac{R_Q}{2} = \frac{Q_3 - Q_1}{2} = \frac{1,43 - 0,56}{2} = \frac{0,87}{2} = 0,435$$

Jest to połowa rozstępu kwartylowego.

- Pozycyjny współczynnik zmienności:

$$V_Q = \frac{Q}{Me}100 = \frac{0,435}{0,95}100 = 45,8\%$$

Informuje jaką część mediany stanowi odchylenie ćwiartkowe.

$$V_{Q_1,Q_3} = \frac{Q_3 - Q_1}{Q_3 + Q_1}100 = \frac{1,43 - 0,56}{1,43 + 0,56}100 = 43,7\%$$

Współczynnik zmienności dla połowy środkowych obserwacji.

2. Spis tablic