dr hab. Agnieszka Bitner
Rzeczoznawca majątkowy

Statystyka opisowa w wycenie nieruchomości
Część I - wyznaczanie miar zbioru danych

WPROWADZENIE

Dział statystyki zajmujący się opisem zbioru danych za pomocą miar oraz graficzną prezentacją danych nazywa się statystyką opisową. Miary (wskaźniki) charakteryzują za pomocą jednej liczby wybrane własności zbioru danych. Taką własnością zbioru danych jest na przykład poziom badanej cechy, a jego miarą może być mediana albo średnia arytmetyczna. Inną własnością jest zróżnicowanie danych w zbiorze, które mierzy wariancja. Wyznaczone miary w sposób syntetyczny przekazują informacje o całym zbiorze danych. Podanie miar jest najprostszym sposobem opisu zbiorów danych oraz umożliwia ich porównywanie. Szczególnym rodzajem zbioru danych jest baza nieruchomości transakcyjnych. Najczęściej do jej scharakteryzowania stosuje się jedną miarę, którą jest średnia cena transakcyjna w wybranym okresie czasu. Jest to najprostsza charakterystyka rynku i w wielu wycenach nieruchomości korzysta się wyłącznie z tej miary. Często nie podaje się nawet niepewności (błędu) oszacowania średniej. Operaty, w których analizuje się inne miary stanowią wyjątek. Nieznajomość innych miar powoduje utratę ważnych informacji o analizowanym zbiorze danych. Na przykład, informacji o rozproszeniu czy asymetrii położenia względem średniej cen w bazie nieruchomości. Znajomość miar rozproszenia i asymetrii umożliwia właściwy dobór i interpretację miar położenia cen, takich jak średnia i mediana. Dla rozproszonych i asymetrycznych danych lepszą od średniej, miarą poziomu cen, jest mediana.

Wskaźniki opisujące badany zbiór danych można podzielić na kilka grup, w zależności od tego jaką własność chcemy zmierzyć w badanym zbiorze. Są to: miary położenia, miary zmienności (rozproszenia), miary asymetrii oraz miary koncentracji (spłaszczenia). Celem artykułu jest omówienie tych podstawowych miar, które powinny być wykorzystywane do opisu zbioru danych transakcyjnych. Głównym narzędziem do analiz rynku używanym przez rzeczoznawców majątkowych jest program MS Excel. Z tego powodu w artykule podano funkcje arkusza kalkulacyjnego służące do wyznaczania miar. Do najważniejszych miar należą miary położenia, dlatego zostaną one przedstawione jako pierwsze.

MIARY OPISUJĄCE ZBIÓR DANYCH

Miary położenia

Rozważmy uporządkowany według wartości zbiór danych. Dla tego zbioru możemy określić graniczne wartości oraz położenie wybranych wartości w zbiorze danych za pomocą kwartyli. Szczególnym kwartylem jest mediana – miara centralnego położenia (centralnej tendencji) w zbiorze danych. Oprócz mediany stosuje się jeszcze dwie miary centralnej tendencji: średnią arytmetyczną zwaną po prostu średnią oraz dominantę. Miary te zostaną omówione poniżej wraz z podaniem symboli, którymi są najczęściej oznaczane.

- *Wartość minimalna – min* – wartość najmniejsza w zbiorze
- *Wartość maksymalna – max* – wartość największa w zbiorze
- *Pierwszy kwartyl* – q_1 – wartość, poniżej której znajduje się jedna czwarta danych w zbiorze
- *Drugi kwartyl* (*mediana*) – q_2 (*Me*) – wartość, poniżej której znajduje się połowa danych w zbiorze
- *Trzeci kwartyl* – q_3 – wartość, poniżej której znajduje się trzy czwarte danych zbiorze
- *Dominanta* (*moda*) – *D* – wartość, która w zbiorze występuje najczęściej
- *Średnia* (*przeciętna*) – $\bar{x}$ – średnia arytmetyczna

Warto przypomnieć, że wartości minimalną i maksymalną oraz kwartyle wyznaczamy dla uporządkowanego rosnąco zbioru danych.

Przykład

Rozważmy uporządkowany zbiór dwudziestu danych:

12, 13, 17, 21, 24, 24, 26, 27, 28, 31, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Miary położenia wyznaczone dla tego zbioru są następujące:

min = 12

max = 58

q_1 = 24

q_2 = 31,5 = Me

q_3 = 41,5

D = 24

$\bar{x}$ = 32,5

Wartości minimalna i maksymalna są oczywiste. Wytłumaczenia wymagają pozostałe miary. Medianą, czyli wartością poniżej której znajduje się połowa danych jest, w przykładowym zbiorze danych, średnia arytmetyczna wartości 31 i 32, czyli 31,5. Właściwie medianą mogłaby być każda wartość pomiędzy 31 i 32, w praktyce jest to średnia z tych wartości. Ogólnie, w przypadku zbiorów o parzystej liczbie danych mediana jest średnią arytmetyczną wartości środkowych znajdujących się na miejscach $\frac{n}{2}$ oraz $\frac{n}{2}+1$. W przypadku nieparzystej liczby danych mediana jest po prostu wartością środkową położoną na pozycji $\frac{n+1}{2}$.

Kwartyle w literaturze wyznacza się w różny sposób, oczywiście wszystkie sposoby określają kwartyle tak, żeby podzieliły one uporządkowany rosnąco zbiór danych na cztery części. W artykule do wyznaczenia kwartyli zastosowano algorytm zgodny z Excelem. Algorytm ten zostanie wyjaśniony na przykładzie obliczenia trzeciego kwartyla w analizowanym zbiorze 20 danych. Trzeci kwartyl to wartość, poniżej której znajduje się 75 %

danych, czyli 15 danych badanego zbioru. Jest to zatem wartość większa od 41 a mniejsza od 43. Miejsce trzeciego kwartyla określa, zgodnie z algorytmem, wzór $(nr\ kwartyla/4) * (n-1)+1$. Czyli, w przykładzie, trzeci kwartyl znajduje się na pozycji $\left(\frac{3}{4}\right) * (20-1)+1 = 15{,}25$ to znaczy, w jednej czwartej odległości między wartościami 41 i 43. Różnica $43-41=2$ po pomnożeniu przez 0,25 wynosi 0,5. Do wartości 41 należy zatem dodać 0,5 żeby otrzymać trzeci kwartyl q_3 = 41,5.

Dominanta to wartość, która w zbiorze występuje najczęściej. W rozważanym przykładzie wartość 24 występuje dwukrotnie, pozostałe wartości jednokrotnie, czyli dominanta wynosi 24. Zazwyczaj zbiór jednostkowych cen transakcyjnych nieruchomości nie ma dominanty, ponieważ ceny jednostkowe są bardzo zróżnicowane. Zbiór danych może zatem nie mieć dominanty (mody), może też mieć więcej niż jedną dominantę.

Średnią zbioru danych jest suma wszystkich wartości elementów tego zbioru podzielona przez ich liczbę, czyli

$$\overline{x} = \frac{x_1+x_2+\cdots+x_n}{n} \tag{1}$$

lub

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

gdzie: *n* – liczba elementów zbioru, $\sum$ oznacza znak sumowania.

Średnia jest najczęściej stosowaną miarą w badaniach rynku nieruchomości. Posługujemy się nią w celu wyznaczenia średniej ceny danego typu nieruchomości w badanym okresie czasu, na przykład w danym miesiącu, obliczając średnią powierzchnię danego typu nieruchomości. Miara ta jest podstawą metody skorygowanej ceny średniej w podejściu porównawczym. Średnia ma własność, która wyróżnia ją spośród pozostałych miar położenia, mianowicie uwzględnia wszystkie informacje zawarte w zbiorze danych. W przeciwieństwie do mediany, która pokazuje względne położenie danych w stosunku do wartości środkowej, nie uwzględniając wartości elementów zbioru położonych poniżej i powyżej wartości środkowej. Powyższa własność mediany może być jej zaletą, jeśli w analizie chcemy ustrzec się wpływu niewielu nietypowych wartości i w takim przypadku może być ona lepszą miarą niż średnia.

Miary zmienności (rozproszenia)

Miary zmienności informują jak bardzo zróżnicowane są wartości w zbiorze danych. Tej informacji nie przekaże żadna z miar centralnej tendencji. Można łatwo wymienić dwa symetryczne zbiory danych, które będą miały takie same średnie, mediany i dominanty, natomiast poszczególne wartości w zbiorach będą się znacznie różniły. To zróżnicowanie pokażą miary zmienności. Do podstawowych miar rozproszenia należą: odstęp międzykwartylowy, rozstęp, wariancja, odchylenie standardowe, współczynnik zmienności.

- *Odstęp międzykwartylowy* – różnica między trzecim i pierwszym kwartylem
$$Q = q_3 - q_1 \tag{3}$$
- *Rozstęp* – różnica między największą i najmniejszą wartością w zbiorze danych
$$R = max - min \tag{4}$$
- *Wariancja* – przeciętne kwadratowe odchylenie poszczególnych wartości w zbiorze od ich średniej

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (5)$$

- *Odchylenie standardowe* – pierwiastek kwadratowy z wariancji

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (6)$$

- *Współczynnik zmienności*

$$V = \frac{s}{\bar{x}}100\% \quad (7)$$

Wariancja podobnie jak średnia uwzględnia wszystkie informacje zawarte w zbiorze danych, ponieważ we wzorze występują wszystkie wartości x_i, w przeciwieństwie do odstępu międzykwartylowego czy rozstępu. W analizach rynku nieruchomości najczęściej wykorzystuje się odchylenie standardowe jako miarę rozproszenia wyników w bazie danych. Odchylenie standardowe jest miarą bardziej naturalną niż wariancja, ponieważ jest wielkością o takim samym wymiarze jak dane wyjściowe czy średnia. Czyli dla zbioru cen odchylenie standardowe jest wyznaczone w złotych, natomiast wariancja w złotych podniesionych do kwadratu. Oczywiście im mniejsze odchylenie standardowe tym lepiej, ponieważ dane są bardziej skupione wokół wartości średniej, co jest sytuacją bardzo pożądaną w procesie wyceny nieruchomości. Współczynnik zmienności określa jakim procentem wartości średniej jest odchylenie standardowe. Mierzy on zatem rozproszenie danych w stosunku do średniej. Jest to wielkość niemianowana, dlatego jest dobrą miarą do porównywania rozproszenia cen wokół średniej w różnych bazach danych.

Przykład
Obliczmy miary rozproszenia dla przykładowych dwudziestu danych
$Q = 17{,}5$
$R = 46$
$s^2 = 159{,}84$
$s = 12{,}64$
$V = 38{,}89$

Oprócz miar centralnej tendencji oraz rozproszenia w opisie zbioru danych pomocne są dwa kolejne typy miar. Są nimi miary asymetrii i spłaszczenia, które określają w jaki sposób rozproszone są dane w zbiorze.

Miary asymetrii (skośności)

- *Współczynnik asymetrii Pearsona*

$$W_{A_P} = \frac{\bar{x}-D}{s} \quad (8)$$

$W_{A_P} < 0$ dla danych rozproszonych bardziej lewostronnie
$W_{A_P} > 0$ dla danych rozproszonych bardziej prawostronnie

- *Współczynnik asymetrii*

$$W_A = \frac{n}{(n-1)(n-2)}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3} \quad (9)$$

$W_A < 0$ dla danych rozproszonych bardziej lewostronnie

$W_A > 0$ dla danych rozproszonych bardziej prawostronnie

Miara koncentracji (spłaszczenia)

- *Współczynnik spłaszczenia (kurtoza)* – określa czy dane są skupione wokół średniej, czy też ich koncentracja wokół średniej jest bardzo mała

$$W_K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \qquad (10)$$

Uwaga: pojęcia asymetrii i spłaszczenia łatwiej jest określić na podstawie analizy kształtu rozkładu częstości. Rozkłady częstości będą przedmiotem drugiej części artykułu i wówczas miary asymetrii i spłaszczenia zostaną szczegółowo omówione.

Przykład

Miary asymetrii i spłaszczenia dla rozpatrywanych wcześniej dwudziestu danych wynoszą:

W_{A_P} = 0,6723

W_A = 0,2368

W_K = – 0,4889

Wykorzystanie arkusza kalkulacyjnego do wyznaczenia miar

Wszystkie wymienione wyżej miary można obliczyć z wykorzystaniem arkusza kalkulacyjnego Excel, stosując dwa sposoby:

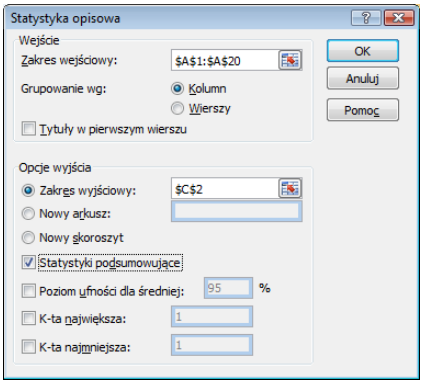
1) wpisując każdą funkcję oddzielnie, korzystając z opcji
Formuły/Wstaw funkcję/Statystyczne

Nazwy funkcji służących do obliczenia poszczególnych miar są napisane wielkimi literami. Tablica oznacza analizowany zbiór danych, w rozpatrywanym przykładzie, zbiór dwudziestu danych.

min: MIN(tablica)
max: MAX(tablica)
q_i: KWARTYL(tablica; nr kwartyla)
Me: KWARTYL(tablica; 2)
$\bar{x}$: ŚREDNIA(tablica)
D: WYST.NAJCZĘŚCIEJ(tablica)
s^2: WARIANCJA(tablica)
s: ODCH.STANDARDOWE(tablica)
W_A: SKOŚNOŚĆ(tablica)
W_K: KURTOZA(tablica)

2) wyliczając jednocześnie wszystkie podstawowe miary, korzystając z opcji
Dane/Analiza danych/Statystyka opisowa

Po wybraniu tej opcji ukaże się okno, które po wpisaniu odpowiednich zakresów danych wygląda następująco (rys. 1):



Rys. 1.

Zakres wejściowy to zbiór danych umieszczony w kolumnie, w przykładzie zakres A1:A20 (patrz rys. 2). Zakres wyjściowy to miejsce w arkuszu, w którym chcemy, żeby została umieszczona tabelka z obliczonymi miarami. Standardowo tabelka z miarami jest umieszczana w nowym arkuszu. Jeśli chcemy ją umieścić w bieżącym arkuszu, obok zbioru danych, wystarczy wpisać w zakresie wyjściowym adres pierwszej komórki bloku komórek, w którym chcemy umieścić tabelkę, czyli adres komórki C2. Następnie zaznaczamy opcję *Statystyki podsumowujące* i wybieramy *OK*, w wyniku otrzymamy poniższą tabelkę

	A	B	C	D	E
1	12				
2	13		*Kolumna1*		
3	17				
4	21		Średnia	32,5	
5	24		Błąd standardowy	2,827031	
6	24		Mediana	31,5	
7	26		Tryb	24	
8	27		Odchylenie standardowe	12,64287	
9	28		Wariancja próbki	159,8421	
10	31		Kurtoza	-0,48886	
11	32		Skośność	0,236828	
12	35		Zakres	46	
13	37		Minimum	12	
14	38		Maksimum	58	
15	41		Suma	650	
16	43		Licznik	20	
17	44				
18	46				
19	53				
20	58				
21					

Rys. 2.

W tabelce przedstawionej na rysunku 2 znajduje się pojęcie, o którym nie było dotychczas mowy – błąd standardowy. Jest to niepewność (błąd) oszacowania średniej i jest on w rozpatrywanym przykładzie określony następującym wzorem:

$$s_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n(n-1)}} \tag{11}$$

W tym miejscu, warto zwrócić uwagę na różnice między odchyleniem standardowym liczonym ze wzoru (6), a niepewnością oszacowania średniej $\bar{x}$ liczoną ze wzoru (11). Odchylenie standardowe s charakteryzuje rozrzut danych wokół średniej $\bar{x}$. Błąd $s_{\bar{x}}$ dotyczy natomiast dokładności oszacowania samej średniej $\bar{x}$. Niepewność oszacowania średniej jest wielkością, którą powinno się wyliczać ilekroć korzystamy ze średniej. Niestety, częstą praktyką jest podawanie średniej bez niepewności jej oszacowania. Średnia podana bez błędu jest pozbawiona bardzo ważnej informacji o tej mierze w analizowanym zbiorze danych.

Wyjaśnienia wymagają również pojęcia: tryb, zakres, licznik. Tryb oznacza dominantę, zakres – rozstęp, licznik – liczebność bazy danych. Jeśli w tabelce obok określenia tryb otrzymamy symbol *#N/D!* oznacza to, że zbiór danych nie ma dominanty. Graficzna prezentacja danych, przede wszystkim wykorzystanie histogramu w analizach rynku nieruchomości, zostanie przedstawiona w drugiej części artykułu.

PODSUMOWANIE

W artykule omówione zostały podstawowe miary służące do opisu bazy danych nieruchomości transakcyjnych. Są nimi miary położenia, zmienności, asymetrii i koncentracji. Do opisu bazy danych skupionych i rozmieszczonych w przybliżeniu symetrycznie stosuje się średnią i wariancję czy odchylenie standardowe – miary wykorzystujące wszystkie wartości badanej cechy. Dla baz danych silnie asymetrycznych, dwu lub wielomodalnych czy baz zawierających kilka wartości cechy bardzo odbiegających od pozostałych, właściwszymi do opisu są kwartyle czy rozstęp. Są to miary wykorzystujące nie wszystkie lecz wybrane wartości badanej cechy. W przypadku wyceny, baza nieruchomości reprezentatywnych powinna spełniać warunki typowego zbioru danych, czyli zbioru o niewielkiej asymetrii i rozproszeniu danych. Do opisu tego typu zbiorów najlepszymi miarami są: średnia obliczona wraz z błędem oszacowana, wariancja, odchylenie standardowe, współczynnik zmienności, współczynnik asymetrii Pearsona, współczynnik asymetrii i kurtoza.

Literatura:

Aczel A.D. Statystyka w zarządzaniu. PWN, Warszawa 2000

Jóźwiak J., Podgórski J. Statystyka od podstaw. PWE, Warszawa 2000

Parlińska M., Parliński J. Statystyczna analiza danych z Excelem. Wydawnictwo SGGW, Warszawa 2011

Pułaska-Turyna B. Wabrane zagadnienia statystyki. WSBFiZ, Warszawa 2002