Charakterystyki liczbowe rozkładów cech statystycznych

Miary położenia, zmienności, asymetrii i koncentracji

(literatura: Ostasiewicz, Rusnak, Siedlecka. Statystyka elementy teorii i zadania. Wydawnictwo AE we Wrocławiu, Wrocław 1999 (2006 – wydanie 6)

Przykład 1. Szereg rozdzielczy szczegółowy

Na podstawie danych statystycznych (szereg rozdzielczy statystyczny szczegółowy n=20) wyznaczyć charakterystyki liczbowe rozkładu dla X – oceny infrastruktury drogowej na Podkarpaciu.

$X = \{1 - niedostateczna, 2 - mierna, 3 - dostateczna, 4 - dobra, 5 - bardzo\ dobra, 6 - celująca\}$.

Obliczyć podstawowe miary klasyczne i pozycyjne: położenia, zmienności, asymetrii i koncentracji – wyniki zinterpretować w praktyce.

Ocena infrastruktury drogowej - Xi	Dane posortowane Xi	$(X_i - \bar{x})^2$	$(X_i - \bar{x})^3$	$(X_i - \bar{x})^4$
4	1	$(4-4{,}1)^2 = 0{,}01$	$(4-4{,}1)^3 = -0{,}001$	$(4-4{,}1)^4 = 0{,}00$
5	2	$(5-4{,}1)^2 = 0{,}81$	$(5-4{,}1)^3 = 0{,}73$	$(5-4{,}1)^4 = 0{,}66$
5	2	$(5-4{,}1)^2 = 0{,}81$	$(5-4{,}1)^3 = 0{,}73$	$(5-4{,}1)^4 = 0{,}66$
3	3	$(3-4{,}1)^2 = 1{,}21$	$(3-4{,}1)^3 = -1{,}33$	$(3-4{,}1)^4 = 1{,}46$
6	3	$(6-4{,}1)^2 = 3{,}61$	$(6-4{,}1)^3 = 6{,}86$	$(6-4{,}1)^4 = 13{,}03$
4	3	$(4-4{,}1)^2 = 0{,}01$	$(4-4{,}1)^3 = -0{,}001$	$(4-4{,}1)^4 = 0{,}00$
6	3	$(6-4{,}1)^2 = 3{,}61$	$(6-4{,}1)^3 = 6{,}86$	$(6-4{,}1)^4 = 13{,}03$
5	3	$(5-4{,}1)^2 = 0{,}81$	$(5-4{,}1)^3 = 0{,}73$	$(5-4{,}1)^4 = 0{,}66$
6	4	$(6-4{,}1)^2 = 3{,}61$	$(6-4{,}1)^3 = 6{,}86$	$(6-4{,}1)^4 = 13{,}03$
3	4	$(3-4{,}1)^2 = 1{,}21$	$(3-4{,}1)^3 = -1{,}33$	$(3-4{,}1)^4 = 1{,}46$
3	5	$(3-4{,}1)^2 = 1{,}21$	$(3-4{,}1)^3 = -1{,}33$	$(3-4{,}1)^4 = 1{,}46$
2	5	$(2-4{,}1)^2 = 4{,}41$	$(2-4{,}1)^3 = -9{,}26$	$(2-4{,}1)^4 = 19{,}45$
5	5	$(5-4{,}1)^2 = 0{,}81$	$(5-4{,}1)^3 = 0{,}73$	$(5-4{,}1)^4 = 0{,}66$
1	5	$(1-4{,}1)^2 = 9{,}61$	$(1-4{,}1)^3 = -29{,}79$	$(1-4{,}1)^4 = 92{,}35$
5	5	$(5-4{,}1)^2 = 0{,}81$	$(5-4{,}1)^3 = 0{,}73$	$(5-4{,}1)^4 = 0{,}66$
5	5	$(5-4{,}1)^2 = 0{,}81$	$(5-4{,}1)^3 = 0{,}73$	$(5-4{,}1)^4 = 0{,}66$
6	6	$(6-4{,}1)^2 = 3{,}61$	$(6-4{,}1)^3 = 6{,}86$	$(6-4{,}1)^4 = 13{,}03$
3	6	$(3-4{,}1)^2 = 1{,}21$	$(3-4{,}1)^3 = -1{,}33$	$(3-4{,}1)^4 = 1{,}46$
2	6	$(2-4{,}1)^2 = 4{,}41$	$(2-4{,}1)^3 = -9{,}26$	$(2-4{,}1)^4 = 19{,}45$
3	6	$(3-4{,}1)^2 = 1{,}21$	$(3-4{,}1)^3 = -1{,}33$	$(3-4{,}1)^4 = 1{,}46$
suma	**82**	**= 43,8**	**= -23,16**	**= 194,63**

- Miary położenia (klasyczne):

Średnia $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{82}{20} = 4{,}1$ – średnio infrastruktura drogowa na Podkarpaciu była oceniana przez ankietowanych na ocenę 4,1 (4,0 - dobrze).

- Miary położenia (pozycyjne):

Modalna (Dominanta) – wartość najczęstsza statystyki pojawiająca się w rozkładzie $D = M_o = 5$, liczność mody = 6. Najczęściej ankietowani oceniali infrastrukturę drogową na ocenę 5 – bardzo dobrze (6 – ankietowanych na 20 – 30%)

Kwartyle (Mediana, Kwartyl dolny, Kwartyl górny)

Mediana (M_e=Q_2) – wartość środkowa w rozkładzie cechy. W przypadku szeregu szczegółowego wyznaczamy pozycję mediany dla danych uporządkowanych. W przypadku parzystej liczby wyników są dwa takie wyniki $\frac{n}{2} = \frac{20}{2} = 10\ oraz\ \frac{n}{2} + 1 = 11$. Medianę obliczamy jako $Me = \frac{x_{10}+x_{11}}{2} = \frac{4+5}{2} = 4{,}5$. Połowa ankietowanych (50%) oceniła infrastrukturę drogową na ocenę 4,5 lub gorzej (co najwyżej 4,5: 1,2,3 lub 4), tym samym połowa ankietowanych oceniła infrastrukturę nie gorzej niż 4,5 (5 i 6).

Podobnie wyznaczamy:

Kwartyl dolny Q_1. W przypadku parzystej liczby wyników są dwa takie wyniki odpowiadające kwartylowi dolnemu $\frac{n}{4} = \frac{20}{4} = 5\ oraz\ \frac{n}{4} + 1 = 6$. Q_1 obliczamy zatem jako $Q_1 = \frac{x_5+x_5}{2} = \frac{3+3}{2} = 3$. Co czwarty ankietowany (25% ankietowanych) oceniał infrastrukturę drogową na ocenę 3 lub gorzej (co najwyżej 3: 1,2,3), tym samym ¾ (75%) ankietowanych oceniało infrastrukturę nie gorzej niż 3 (co najmniej 3: 3, 4, 5 i 6).

Kwartyl górny Q_3. W przypadku parzystej liczby wyników są dwa takie wyniki odpowiadające kwartylowi górnemu $\frac{3n}{4} = \frac{3*20}{4} = 15\ oraz\ \frac{3n}{4} + 1 = 16$. Q_3 obliczamy zatem jako $Q_3 = \frac{x_{15}+x_{16}}{2} = \frac{5+5}{2} = 5$. ¾ ankietowanych (75% ankietowanych) oceniało infrastrukturę drogową na ocenę 5 lub gorzej (co najwyżej 5: 1,2,3,4,5), tym samym co czwarty ankietowany (25%) ankietowanych oceniało infrastrukturę nie gorzej niż 5 (co najmniej 5: 5, 6).

- Miary zmienności (klasyczne):

Wariancja i odchylenie standardowe. Wariancję obciążoną wyznaczamy ze wzoru $S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{x})^2}{n} = \frac{43{,}8}{20} = 2{,}19\ (2{,}2)$. Wariancje nieobciążoną ${S_1}^2$ (wyznaczaną przez program Statistica), która daje lepsze oszacowanie dla nieznanej wartości wariancji w populacji w przypadku małych prób obliczamy ze wzoru: ${S_1}^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{x})^2}{n-1} = \frac{43{,}8}{20-1} = 2{,}31\ (2{,}3)$. Aby wyrazić zmienność w tych samych jednostkach co wartości statystyki obliczamy wartości odchylenia standardowego: $S = \sqrt{2{,}19}$=1,48 ($S_1 = \sqrt{2{,}31}$=1,52).

Zmienność ocen infrastruktury drogowej wyrażona odchyleniem standardowym wynosi +/- 1,5.

Zatem typowy przedział zmienności wartości tej statystyki jest postaci: $(\bar{x} - S, \bar{x} + S) = (2{,}62 - 5{,}58)$. Jest to typowy przedział zmienności, w którym znajduje się 2/3 (67%)

wszystkich wyników. Typowy przedział zmienności dla ocen infrastruktury to przedział zmienności ocen (2,6 – 5,6), w którym znajduje się zmienność (rozrzut) ocen dla ok. 2/3 (67%) wszystkich badanych ankietowanych. Jest to dość duża zmienność. Aby ocenić jak jest to duża procentowo zmienność należy wyznaczyć klasyczny procentowy współczynnik zmienności (procentowe zróżnicowanie wartości statystyki) $V_s = \frac{S}{\bar{x}} * 100\% = \frac{1,48}{4,1} * 100 = 36,1\%$. Zmienność ocen infrastruktury drogowej jest na poziomie 36% (dość duża zmienność – znaczne istotne zróżnicowanie wartości badanej statystyki > 10%).

- Miary zmienności (pozycyjne):

Rozstęp jest pozycyjnym parametrem charakteryzującym zmienność rozkładu wartości badanych statystyk. Rozstęp obliczamy ze wzoru: $R = x_{max} - x_{min} = 6 - 1 = 5$. Maksymalna zmienność – rozrzut (rozstęp) dla ocen infrastruktury drogowej wynosi 5. Podobnie możemy policzyć rozstęp kwartylowy $R_{Q_1,Q_3} = Q_3 - Q_1 = 5 - 3 = 2$. Rozstęp kwartylowy charakteryzuje zmienność wyników po odrzuceniu 25% najmniejszych wyników (poniżej Q1=3 dostatecznie) oraz 25% największych wyników (powyżej Q3=5 bardzo dobrze). Tak charakteryzowana zmienność wyników dla ocen infrastruktury wynosi 2. Pozycyjnym współczynnikiem zmienności jest odchylenie ćwiartkowe (podobnie jak w przypadku miar klasycznych odchylenie standardowe). Wyznaczmy go ze wzoru: $Q = \frac{(Q_3 - M_e) + (M_e - Q_1)}{2} = \frac{Q_3 - Q_1}{2} = \frac{5-3}{2} = 1$. Procentowy współczynnik zmienności pozycyjny można wyznaczyć ze wzoru $V_Q = \frac{Q}{M_e} * 100\% = \frac{1}{4,5} * 100 = 22,2\%$.

- Miary asymetrii – skośności rozkładu:

Rozkłady różnią się miedzy sobą zarówno kierunkiem jak i siłą asymetrii. Do oceny asymetrii stosuje się porównanie $\bar{x}, M_e\ i\ M_o$. Jeżeli $\bar{x} = M_e = M_o$ mamy do czynienia z rozkładem w pełni symetrycznym. Jeżeli $\bar{x} > M_e > M_o$ mamy do czynienia z rozkładem o asymetrii prawostronnej. Natomiast jeżeli $\bar{x} < M_e < M_o$ mamy do czynienia z rozkładem o asymetrii lewostronnej. W naszym przypadku: $\bar{x} = 4,1$; $M_e = 4,5$; $M_0 = 5$, zatem mamy do czynienia z asymetrią lewostronną rozkładu ocen infrastruktury. Do określania siły i kierunku asymetrii można stosować współczynniki skośności oparte zarówno na parametrach klasycznych jak i pozycyjnych. W przypadku parametrów klasycznych można wyznaczyć np. współczynnik skośności wyrażony wzorem $A_S = \frac{\bar{x} - M_o}{S} = \frac{4,1-5}{1,48} = -0,61$ (asymetria lewostronna – dość duża). W przypadku parametrów pozycyjnych wyznacza się współczynnik skośności wyrażony wzorem $A_Q = \frac{Q_3 + Q_1 - 2M_e}{2Q} = \frac{5+3-2*4,5}{2*1} = -0,5$ (asymetria lewostronna). Można również wyznaczyć (podobnie jak w pakiecie Statistica) klasyczny współczynnik asymetrii (skośności) oparty na tzw. momencie centralnym rozkładu rzędu 3 ze wzoru: $A = \frac{m_3}{S^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^3} = \frac{-23,16 * \frac{1}{20}}{1,48^3} = -0,36$ (również świadczy to znaczącej asymetrii lewostronnej rozkładu ocen infrastruktury drogowej).

- Miary koncentracji rozkładu:

Miarą skupienia (koncentracji) poszczególnych wartości wokół średniej jest współczynnik skupienia Kurtoza $K = \frac{m_4}{S^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{S^4} = \frac{194,63*\frac{1}{20}}{1,48^4} = 2,03$. Porównanie koncentracji badanego rozkładu z rozkładem Normalnym (Gaussa), który jest symetrycznym rozkładem o koncentracji wyrażonej Kurtozą K=3 odbywa się za pomocą analizy współczynnika koncentracji $K' = K - 3 = 2,03 - 3 = -0,97$. Zatem rozkład naszej badanej statystyki ma mniejszą koncentrację (jest bardziej spłaszczony) w porównaniu do analogicznego (o tych samych parametrach) rozkładu normalnego (jest rozkładem tzw. platokurtycznym). Gdy K>3 mamy do czynienia z rozkładami bardziej wysmukłymi (tzw. leptokurtycznymi).

Przykład 2. Szereg rozdzielczy punktowy – dane uporządkowane (dane statystyczne z przykładu 1)

Na podstawie danych z przykładu 1 utworzono szereg rozdzielczy punktowy:

Klasa (i) k=6 liczba klas	X_i	n_i	Liczność sk. $n_{i,sk}$	Procent w_i	Procent sk. $w_{i,sk}$	$x_i * n_i$	$(x_i - \bar{x})^2 * n_i$	$(x_i - \bar{x})^3 * n_i$	$(x_i - \bar{x})^4 * n_i$
ocena 1	1	1	1	0,05	0,05	1*1=1	9,61	-29,791	92,3521
ocena 2	2	2	3	0,1	0,15	2*2=4	8,82	-18,522	38,8962
ocena 3	3	5	8	0,25	0,4	3*5=15	6,05	-6,655	7,3205
ocena 4	4	2	10	0,1	0,5	4*2=8	0,02	-0,002	0,0002
ocena 5	5	6	16	0,3	0,8	5*6=30	4,86	4,374	3,9366
ocena 6	6	4	20	0,2	1	6*4=24	14,44	27,436	52,1284
suma		**20**		1		**82**	43,8	**-23,16**	**194,634**

Wyznaczyć dla tak zgrupowanych danych jeszcze raz: średnią, wariancję i odchylenie standardowe oraz współczynnik skośności (A oparty o moment centralny m_3) oraz współczynnik koncentracji ($K' = K - 3$) – zauważmy, że otrzymamy te same wyniki co w przykładzie 1.

Średnia: $\bar{x} = \frac{\sum_{i=1}^{k} x_i*n_i}{n} = \frac{82}{20} = 4,1$. Wariancja $S^2 = \frac{\sum_{i=1}^{k}(X_i-\bar{x})^2*n_i}{n} = \frac{43,8}{20} = 2,19\ (2,2)$, odchylenie standardowe $S = \sqrt{2,19} = 1,48$. Współczynnik asymetrii (skośności) $A = \frac{m_3}{S^3} = \frac{\frac{1}{n}\sum_{i=1}^{k}(x_i-\bar{x})^3*n_i}{S^3} = \frac{-23,16*\frac{1}{20}}{1,48^3} = -0,36$. Współczynnik koncentracji $K' = K - 3 = \frac{m_4}{S^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{k}(x_i-\bar{x})^4*n_i}{S^4} - 3 = \frac{194,63*\frac{1}{20}}{1,48^4} - 3 = 2,03 - 3 = -0,97$.

Przykład 3. Szereg rozdzielczy przedziałowy – dane uporządkowane

Na podstawie danych zgrupowanych w postaci szeregu rozdzielczego przedziałowego dla czasów dojazdu na uczelnię 15 studentów (min):

Przedziały klasowe k=4	Środki $\dot{x}_i$	n_i	Liczność sk. $n_{i,sk}$	Procent w_i	Procent sk. $w_{i,sk}$	$\dot{x}_i * n_i$	$(\dot{x}_i - \bar{x})^2 * n_i$	$(\dot{x}_i - \bar{x})^3 * n_i$	$(\dot{x}_i - \bar{x})^4 * n_i$
[5-13]	9	3	3	0,2	0,2	27	522,72	-6899,904	91078,7328
[14-22]	18	5	8	0,333333	0,53	90	88,2	-370,44	1555,848
[23-31]	27	4	12	0,266667	0,8	108	92,16	442,368	2123,3664
[32-40]	36	3	15	0,2	1	108	571,32	7884,216	108802,1808
suma		**15**		1		**333**	**1274,4**	**1056,24**	**203560,128**

Wyznaczyć statystyki opisowe parametrów rozkładu badanej statystyki (**Interpretację praktyczną podać samodzielnie !!!**), takie jak:

Średnia= $\bar{x} = \frac{\sum_{i=1}^{k} \dot{x}_i * n_i}{n} = \frac{333}{15} = 22{,}2$ (min)

Modalna - dla danych zgrupowanych wartość modalnej można oszacować następująco. Określić przedział zawierający modalną (klasa druga [14-22] – liczność 5). Wyliczyć jej oszacowanie ze wzoru (3.12 – podręcznik Ostasiewicz): $M_o = 14 + \frac{5-3}{(5-3)+(5-4)} * 9 = 20$ (min).

Mediana - dla danych zgrupowanych wartość kwartyli (w tym wypadku wartości środkowej mediany można oszacować następująco. Określić przedział zmienności, w którym znajduje się mediana (klasa druga – procent skumulowany powyżej 0,5: [14-22] – liczność 5). Wyliczyć jej oszacowanie ze wzoru (3.14 – podręcznik Ostasiewicz): $M_e = 14 + \frac{7{,}5-3}{5} * 9 = 22{,}1$ (min).

Podobnie mamy dla kwartyla dolnego. Przedział zmienności, w którym znajduje się Q_1 (to dalej klasa druga – procent skumulowany powyżej 0,25: [14-22] – liczność 5). Zatem $Q_1 = 14 + \frac{3{,}75-3}{5} * 9 = 15{,}35$ (min).

Oraz dla kwartyla górnego. Przedział zmienności, w którym znajduje się Q_3 (to klasa trzecia – procent skumulowany powyżej 0,75: [23-31] – liczność 4). Zatem $Q_3 = 23 + \frac{11{,}25-8}{4} * 9 = 30{,}31$ (min).

Wariancja i odchylenie standardowe: $S^2 = \frac{\sum_{i=1}^{k} (\dot{x}_i - \bar{x})^2 * n_i}{n} = \frac{1274{,}4}{15} = 84{,}96, S = \sqrt{84{,}96} = 9{,}2.$

Skośność i Kurtoza: Współczynnik asymetrii (skośności) $A = \frac{m_3}{S^3} = \frac{\frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^3 * n_i}{S^3} = \frac{1056,24 * \frac{1}{15}}{9,2^3} = 0,09$. Współczynnik koncentracji $K' = K - 3 = \frac{m_4}{S^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^4 * n_i}{S^4} - 3 = \frac{203560,13 * \frac{1}{15}}{9,2^4} - 3 = 1,9 - 3 = -1,1$.

- **Zadania do samodzielnego rozwiązania (z książki: Ostasiewicz, Rusnak, Siedlecka. Statystyka elementy teorii i zadania. Wydawnictwo AE we Wrocławiu, Wrocław 1999 (2006)).**

- przeanalizować przykłady dla średniej geometrycznej i harmonicznej – str. 51 – 53. Interpretacja i zastosowanie tego typu średnich.

- Zad. 3.3, Zad. 3.10, Zad. 3.12, Zad. 3.14, Zad. 3.30, Zad. 3.31, Zad. 3.34.