PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1097 — 2005
Statystyka ekonomiczna

Stanisław Heilpern

NIETYPOWE REALIZACJE JEDNOWYMIAROWYCH ZMIENNYCH LOSOWYCH

1. Wstęp

Analizując dane statystyczne, możemy natrafić na obserwacje nietypowe, często różniące się dość wyraźnie od pozostałych. Przyczyny występowania tego typu obserwacji mogą być różne; możemy je z grubsza podzielić na dwie kategorie. Pierwsza kategoria obejmuje obserwacje, których wartości zostały błędnie podane. Dane mogą być też niewłaściwie zapisane, a także celowo zmienione. Przekłamania danych występują zarówno u samych źródeł ich zbierania, jak i podczas ich kodowania. Mogą też występować błędy obserwacji.

Drugim rodzajem obserwacji nietypowych są poprawne dane, wyraźnie odróżniające się, odstające od pozostałych. Pochodzą one wtedy zwykle z populacji innej niż zasadnicza część obserwacji, nazywana często *rdzeniem*. Zbiór obserwacji dzielimy wtedy na rdzeń oraz obserwacje, które będziemy nazywać odstającymi (ang. *outliers*). Obserwacje odstające można też traktować jako zaburzenia zniekształcające rozkład badanej populacji.

W praktyce obserwacje nietypowe pierwszej kategorii znajdujemy zwykle w sposób zdroworozsądkowy, wychwytując np. absurdalne wartości obserwacji, takie jak wzrost osoby wynoszący 348 cm czy gmina licząca 34,6 mln mieszkańców. W naszej pracy interesować nas będą jedynie obserwacje nietypowe drugiej kategorii, obserwacje odstające. Badać je będziemy metodami statystycznymi, wykorzystując wnioskowanie statystyczne. Założymy też, że rozpatrywane w pracy obserwacje będą jednowymiarowe, dotyczące pojedynczej cechy.

W przypadku obserwacji jednowymiarowych wykrywanie obserwacji odstających, czyli nietypowych, nie pasujących do ustalonego rozkładu opisującego badaną populację, sprowadza się do badania obserwacji ekstremalnych. Obserwacje

odstające pochodzące z innej populacji nie muszą być oczywiście obserwacjami ekstremalnymi, jednakże nie można ich wtedy wychwycić metodami statystycznymi. Również nie każda obserwacja ekstremalna jest odstająca.

Praca ma charakter przeglądowy. Składa się z dwóch części i dodatku. W pierwszej przedstawimy wybrane metody wykrywania obserwacji odstających, a w drugiej opiszemy *odporne* metody wyznaczania estymatorów charakterystyk rozkładu populacji uwzględniających obserwacje odstające, jednak słabo na nie reagujących. Ponadto dodatek obejmuje 4 tablice, zawierające wartości testów statystycznych umożliwiających wykrywanie obserwacji odstających.

Przykłady zamieszczone w pracy zawierają głównie dane fikcyjne, służące jedynie do ilustracji przedstawionych metod wykrywania obserwacji odstających. Podane są jednak też przykłady oparte na rzeczywistych danych, sygnalizujące możliwe, praktyczne zastosowania tych metod.

2. Wykrywanie obserwacji odstających

Intuicyjnie rzecz ujmując, obserwacja ekstremalna może być odstająca, gdy jest dostatecznie oddalona od pozostałych obserwacji, tzn. gdy jej wartość jest zbyt duża lub zbyt mała, nawet jeśli jest obserwacją ekstremalną.

Z tego też powodu do zagadnień związanych z wykrywaniem obserwacji odstających wykorzystuje się zwykle tzw. *testy niezgodności* (ang. *discordance tests*) [1], oparte na statystykach pozycyjnych. Załóżmy w tym celu, że $x_1, x_2, \ldots, x_n$ jest n-elementową próbą pochodzącą z populacji o ustalonym rozkładzie F. Mogą to być przykładowo odpowiedzi wylosowanych n respondentów na ustalone pytanie. W celu wykrycia obserwacji nietypowych, nie pochodzących z ustalonego rozkładu F, czyli odstających, należy uporządkować obserwacje, otrzymując ciąg

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)},$$

traktowany jako realizacja statystyk pozycyjnych $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. W przypadku danych jednowymiarowych obserwacje odstające, możliwe do zlokalizowania, są ekstremalne: dolne, czyli $x_{(1)}, x_{(2)}$ lub następne, albo górne, czyli $x_{(n)}, x_{(n-1)}$ i kolejne. Wartość górnych obserwacji ekstremalnych powinna być istotnie duża względem rozkładu statystyki pozycyjnej $X_{(n)}$, oczywiście przy założeniu, że populacja ma rozkład F. Podobne rozumowanie można przeprowadzić w odniesieniu do dolnych obserwacji odstających. Wtedy rozpatrujemy statystyki $X_{(1)}$. Powyższe rozważania stanowią istotę testów niezgodności.

Obserwacja $x_{(n)}$ uznana na podstawie takiego rozumowania za odstającą może pochodzić z innej populacji o rozkładzie G, czyli być zaburzeniem. W tej sytuacji

można wykorzystać testy jednorodności próby [2]. Testy te polegają na weryfikacji przeprowadzonej na podstawie wyników z próby hipotezy H_0, że próba ta pochodzi z ustalonej populacji generalnej o rozkładzie F, wobec hipotezy H_1, że jeden lub więcej elementów zostało wylosowanych z innej populacji o rozkładzie G. Najczęściej przyjmuje się, że populacja F ma rozkład normalny $N(\mu, \sigma)$, ale są też testy dotyczące rozkładu gamma, jednostajnego, logarytmiczno-normalnego, dwumianowego, Pareta czy Poissona [1].

Testy wychwytujące obserwacje odstające są zarówno dwustronne, czyli wykrywające dolne oraz górne obserwacje odstające, jak i jednostronne, wykrywające tylko dolne lub górne obserwacje. Ponadto jedne testy wykrywają pojedyncze obserwacje odstające, a inne całe grupy tych obserwacji.

Hipoteza alternatywna występująca przy testach jednorodności próby jest dość ogólna. Można ją zawęzić, dokładniej określając charakter zaburzenia. Na przykład w odniesieniu do populacji o rozkładzie normalnym $N(\mu, \sigma)$ hipoteza alternatywna może przybrać postać:

$$H_1: X_i \sim N(\mu + a, \sigma) \quad \text{dla pewnego } i.$$

Zakładamy wtedy, że jedna obserwacja jest zaburzeniem pochodzącym z innej populacji, której rozkład jest również normalny o tym samym odchyleniu standardowym, przesunięty o wektor długości a. Inna hipoteza alternatywna przyjmuje, że populacja, z której pochodzi zaburzenie, ma rozkład normalny o tym samej wartości oczekiwanej, lecz o większej wariancji, tzn.

$$H_1: X_i \sim N(\mu, b\sigma) \quad \text{dla pewnego } i,$$

gdzie $b > 1$.

Barnett w [1] podzielił sprawdziany testów niezgodności wykrywających obserwacje odstające na siedem klas:

I. *Statystyki nadwyżka/rozproszenie.* Są to statystyki oparte na różnicy sąsiednich obserwacji ekstremalnych podzielonej przez jedną z miar rozproszenia. Na przykład mogą to być statystyki wykrywające górne obserwacje odstające, takie jak

$$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$$

lub

$$\frac{x_{(n)} - x_{(n-1)}}{\sigma},$$

gdzie σ jest odchyleniem standardowym bazowego rozkładu. W analogiczny sposób konstruujemy sprawdziany testu wykrywającego dolne obserwacje odstające. Gdy odchylenie to nie jest znane, a ten przypadek najczęściej występuje w praktyce, zastępujemy je estymatorem wyznaczonym z próby, np. s. Miarę rozproszenia występującą w mianowniku powyższych wzorów można też obliczyć, usuwając z próby wartości ekstremalne. Sprawdziany te wykorzystujemy w testach jednostronnych.

II. *Statystyka rozstęp/rozproszenie.* Statystyki te otrzymujemy przez wstawienie w liczniku rozstępu wyznaczonego z próby. Przykładem jest sprawdzian postaci

$$\frac{x_{(n)} - x_{(1)}}{\sigma}.$$

Stosując ten sprawdzian, nie można niestety ocenić, czy mamy do czynienia z dolną, górną, czy jedną i drugą obserwacją odstającą.

III. *Statystyka odchylenie/rozproszenie.* W liczniku tej statystyki wstawiamy odchylenie od wartości centralnej. Na przykład dla górnych obserwacji odstających przyjmujemy

$$\frac{x_{(n)} - \bar{x}}{\sigma}.$$

Wspomniane wady statystyki opartej na rozstępie są tu częściowo zrekompensowane. Zamiast średniej można też wziąć inną miarę położenia. Inną modyfikacją może być statystyka

$$\max_i \frac{|x_i - \bar{x}|}{s}.$$

IV. *Sumy kwadratów.* Statystyki tego typu są ilorazami sum kwadratów odchyleń od wartości średniej dla próby z usuniętymi elementami ekstremalnymi i sum kwadratów wyznaczonych dla pełnej próby. Na przykład w odniesieniu do testu umożliwiającego wykrycie dwóch dolnych obserwacji odstających dla populacji o rozkładzie normalnym stosujemy statystykę

$$\frac{\sum_{i=3}^{n} (x_{(i)} - \bar{x}_{1,2})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

gdzie $\bar{x}_{1,2} = \frac{1}{n-2}\sum_{i=3}^{n} x_{(i)}$.

V. *Statystyki ekstremum/położenie.* Są to statystyki oparte na ilorazach wartości ekstremalnych przez miary położenia, takich jak prosty sprawdzian wykrywający górne obserwacje odstające:

$$\frac{x_{(n)}}{\bar{x}}.$$

Sprawdziany tego typu są wykorzystywane w odniesieniu do populacji o rozkładzie gamma.

VI. *Statystyki momentów wyższego stopnia.* Wykorzystuje się w tym przypadku statystyki oparte na miarach skośności i spłaszczenia:

$$\frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^{3/2}} \quad \text{lub} \quad \frac{n\sum_{i=1}^{n}(x_i-\bar{x})^4}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2}.$$

Miary tego typu są jednak rzadko stosowane w praktyce.

VII. *W-statystyki.* Są to statystyki oparte na kombinacjach liniowych statystyk pozycyjnych podzielonych przez sumy kwadratów odchyleń obserwacji od wartości średniej. Przykładem jest W-statystyka Shapiro-Wilka [2]

$$\frac{\sum_{i=1}^{[n/2]} a_{n,i}(x_{(n-i+1)}-x_{(i)})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

gdzie współczynniki $a_{n,i}$ spełniają warunki $\sum_{i=1}^{n} a_{n,i} = 0$ oraz $\sum_{i=1}^{n} a_{n,i}^2 = 1$.

Przeprowadzając testy wykrywające obserwacje odstające, możemy natrafić na dwie niepożądane sytuacje. Są to tzw. efekty maskowania i przeciągania. Efekt *maskowania* występuje, gdy jedna obserwacja „zasłania" drugą, tzn. w testach dotyczących pojedynczych obserwacji odstających. Podczas wykrywania np. górnej obserwacji odstającej może się zdarzyć, że wcześniejsza obserwacja $x_{(n-1)}$ leży blisko obserwacji ostatniej $x_{(n)}$, tak że ją „zasłania", i że test jej nie wykryje, mimo że jest faktycznie odstająca. Efekt maskowania można zniwelować, stosując np. testy wykrywające grupy obserwacji odstających. Jednakże należy pamiętać, że np.

obserwacja $x_{(n-2)}$ może „zasłaniać" parę obserwacji odstających $x_{(n-1)}$ i $x_{(n)}$. Możemy też najpierw sprawdzić, czy obserwacja $x_{(n-1)}$ w relacji do zbioru obserwacji $x_{(1)}$, $x_{(2)}$, ..., $x_{(n-2)}$ jest odstająca. Jeśli tak, to para obserwacji $x_{(n-1)}$ oraz $x_{(n)}$ jest odstająca. W przeciwnym razie sprawdzamy nietypowość obserwacji $x_{(n)}$. Procedurę tą możemy oczywiście rozszerzyć na większą liczbę k obserwacji. Wtedy zaczynamy od obserwacji $x_{(n-k+1)}$.

Efekt *przyciągania* zaś występuje, gdy obserwacja nie będąca odstającą będzie „przyciągana" przez jedną obserwację lub przez grupę obserwacji odstających. Sytuacja taka może wystąpić np. wówczas, gdy stosujemy testy badające, czy wybrana grupa obserwacji nie narusza jednorodności (zob. przykład 4), lub testy typu „odchylenie/rozproszenie". W drugim przypadku górna obserwacja odstająca może tak „przyciągnąć" w swoją stronę wartość średnią, że dolna obserwacja nie będąca odstającą może zostać za nią uznana. Sytuacja taka występuje w przykładzie 1 w [4] w przypadku obserwacji wielowymiarowych.

Przedstawimy teraz wybrane testy statystyczne umożliwiające wykrywanie obserwacji odstających [1; 2]. Na początku założymy, że populacja ma rozkład normalny $N(\mu, \sigma)$. Test istotności Dixona jest jednym z najczęściej stosowanych. Polega on na obliczeniu wartości następujących statystyk:

$$d_n = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}, \qquad d_1 = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}},$$

$$d_{1,n} = \max\left(d_1, d_n\right), \qquad d_{n,n-1} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}}, \qquad d_{1,2} = \frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}}.$$

Jak widać, są to statystyki typu nadwyżka/rozproszenie. Jeśli dla ustalonego poziomu istotności α [2]:

a) $d_n \geq d_{1,\alpha}$ to element $x_{(n)}$ narusza statystyczną jednorodność próby, czyli można go uznać za obserwację odstającą,

b) $d_1 \geq d_{1,\alpha}$ to element $x_{(1)}$ narusza statystyczną jednorodność próby,

c) $d_{1,n} \geq d_{2,\alpha}$, to elementy $x_{(1)}$ i $x_{(n)}$ naruszają statystyczną jednorodność próby,

d) $d_{n,n-1} \geq d_{3,\alpha}$, to elementy $x_{(n)}$ i $x_{(n-1)}$ naruszają statystyczną jednorodność próby,

e) $d_{1,2} \geq d_{3,\alpha}$, to elementy $x_{(1)}$ i $x_{(2)}$ naruszają statystyczną jednorodność próby,

gdzie wartości krytyczne $d_{1,\alpha}$, $d_{2,\alpha}$ oraz $d_{3,\alpha}$ odczytujemy z tabl. I, zawierającej wartości krytyczne dla testu Dixona (por. [2]).

Przykład 1

Poniższe dane przedstawiają wzrost losowo wybranych 13 osób mierzony w cm:

152; 172; 173; 174; 174; 175; 176; 176; 177; 178; 179; 195; 203.

Chcąc przeprowadzić test Dixona, powinniśmy najpierw sprawdzić założenie normalności rozkładu badanej zmiennej [2]. Posługując się testem Shapiro-Wilka, stwierdzamy, że nie ma podstaw do odrzucenia tezy o normalności rozkładu.

Możemy teraz przejść do wykrywania obserwacji odstających. Sprawdzimy przykładowo dwie hipotezy: pierwszą, że obserwacja x_1 nie narusza jednorodności próby, oraz drugą, że obserwacje x_{13} i x_{12} również nie naruszają jednorodności próby. W tym celu obliczamy

$$d_1 = 20/51 = 0{,}392, \qquad d_{13,12} = 24/51 = 0{,}471.$$

Po przyjęciu poziomu istotności $\alpha = 0{,}05$ otrzymujemy następujące wartości krytyczne: $d_{1,\alpha} = 0{,}361$ oraz $d_{3,\alpha} = 0{,}461$. Wartości sprawdzianów obydwu hipotez są większe od wartości krytycznych, tzn. $d_{1,\alpha} < d_1$, $d_{3,\alpha} < d_{13,12}$, więc hipotezy te należy odrzucić. Badane obserwacje naruszają jednorodność próby, czyli mamy podstawy do uznania x_1, x_{12}, x_{13} za obserwacje odstające.

Test Dixona jest oparty na różnicach między największymi bądź najmniejszymi wartościami z próby. Z tego też powodu nie jest on odporny na efekt maskowania. Zawodzi bowiem, gdy są dwie obserwacje odstające, między którymi są małe odległości.

Przykład 2

W przykładzie 1 efekt maskujący występuje podczas testowania hipotezy sprawdzającej, czy obserwacja x_{13}, rozpatrywana pojedynczo, a nie w parze z x_{12} jak w tym przykładzie, nie narusza jednorodności próby. Wtedy $d_{13} = 0{,}157 < 0{,}361 = d_{1,\alpha}$. W tym przypadku nie mamy podstaw do odrzucenia tej hipotezy i uznania obserwacji x_{13} za odstającą.

Efekt maskujący nie odgrywa większej roli w teście Grubbsa. Test ten jest oparty na odległościach między skrajnymi elementami a wartością średnią, czyli na statystykach

$$G_{(n)} = \frac{x_{(n)} - \bar{x}}{s}, \qquad G_{(1)} = \frac{\bar{x} - x_{(1)}}{s},$$

gdzie $\bar{x}$ jest średnią, a s odchyleniem standardowym wyznaczonym z próby. Jeśli $G_{(i)} \geq G_\alpha$, to obserwacja $i = 1$ lub n może być uważana za odstającą. Wartość kry-

tyczną odczytujemy z tabl. II. Jest to sprawdzian testu typu rozproszenie/odchylenie. p-wartość testu (p-value) możemy też oszacować z góry nierównością

$$p \le nP\left(T_{n-2} > \sqrt{\frac{n(n-2)G_{(n)}^{\ 2}}{(n-2)^2 - nG_{(n)}^{\ 2}}} \right), \tag{1}$$

gdzie T_k jest zmienna losową o rozkładzie Studenta z k-stopniami swobody. Równość w (1) zachodzi, gdy $G_{(n)} \ge \sqrt{\frac{(n-1)(n-2)}{2n}}$ [1].

Przykład 3

Przeprowadzimy test Grubbsa w odniesieniu do danych przedstawiających wydatki w tys. zł wybranych losowo 8 rodzin:

1,22; 1,24; 1,32; 1,39; 1,49; 1,68; 1,87; 3,02.

Wartość sprawdzianu testu $G_{(8)} = 2{,}295$. Wartość krytyczna dla poziomu istotności $\alpha = 0{,}05$ jest równa 2,03, więc obserwację $x_8 = 3{,}02$ można uznać za odstającą. Prawa strona nierówności (1) jest mniejsza niż 10^{-5}, a tym bardziej p-wartość testu, co potwierdza wcześniejsze rozważania.

Na danych z przykładu 3 możemy zaobserwować efekt przyciągania – drugą obok maskowania niebezpieczną sytuację towarzysząca wykrywaniu obserwacji odstających.

Przykład 4

Po zastosowaniu testu Dixona do pary obserwacji $x_{(7)}$ i $x_{(8)}$ z przykładu 3 otrzymujemy $d_{8,7} = 0{,}74$ oraz $d_{3,\alpha} = 0{,}545$ dla $\alpha = 0{,}05$. Można więc na podstawie testu Dixona uważać parę $x_{(7)}$ i $x_{(8)}$ za obserwacje odstające. Jednakże obserwacja $x_{(7)}$ nie jest odstająca. Została ona jedynie „przyciągnięta" przez obserwację $x_{(8)}$. Stosując ten test do obserwacji $x_{(7)}$ rozpatrywanej względem pozostałych, po usunięciu obserwacji $x_{(8)}$ mamy $d_7 = 0{,}29$ i $d_{1,\alpha} = 0{,}507$. Hipotezy o jednorodności nie możemy więc odrzucić. Obserwacja $x_{(7)}$ nie jest zatem odstająca. Potwierdza to też test Grubbsa. Wartość sprawdzianu testu dla 7-elementowej próby wynosi $G_{(7)} = 1{,}709$, a wartość krytyczna testu jest w tym przypadku równa 1,94.

Testy Dixona i Grubbsa mogą też służyć do sprawdzenia jednorodności próby, do weryfikacji hipotezy, że próba pochodzi z populacji o rozkładzie normalnym, wobec hipotezy, że jedna lub para obserwacji należą do innej populacji [2].

Przykład 5

Podana poniżej, uporządkowana próba przedstawia wartości sumy ubezpieczeniowej AC 8 klientów pewnej firmy ubezpieczeniowej z 2001 r.:

6 000; 11 500; 12 000; 15 000; 19 500; 20 000; 28 000; 38 000.

Można sprawdzić, że test Shapiro-Wilka nie odrzuca na poziomie istotności $\alpha = 0{,}05$ hipotezy dotyczącej normalności rozkładu populacji generalnej. Po zastosowaniu testu Dixona do najmniejszej i największej obserwacji otrzymujemy wartości testu równe odpowiednio 0,172 oraz 0,313. Test ten nie odrzuca hipotezy o jednorodności próby, ponieważ wartość krytyczna wynosi w tym przypadku $d_{1,\alpha} = 0{,}468$. Również test Grubbsa potwierdza ten wynik. Wartości tego testu są odpowiednio równe dla skrajnych obserwacji 1,246 i 1,881, a wartość krytyczna wynosi 2,03. Nie ma więc podstaw do odrzucenia hipotezy o jednorodności próby. Otrzymane dane pochodzą więc z jednorodnej populacji. Ekstremalne wartości sumy ubezpieczeniowej wylosowanych klientów nie odbiegają istotnie od wartości tej sumy dla pozostałych klientów. Można ich zaliczyć do tej samej kategorii.

Test Grubbsa wykrywa jedynie pojedyncze obserwacje odstające, a korzystając z testu Dixona, można odkryć pary tego typu obserwacji. Są jednak testy znajdujące większą liczbę obserwacji odstających. Testy te możemy podzielić na dwie kategorie. Do pierwszej należą testy sprawdzające, czy zestaw k obserwacji stanowi zbiór obserwacji odstających. Mogą to być obserwacje najmniejsze, największe lub zarówno jedne, jak i drugie. Druga grupa obejmuje testy sekwencyjne, sprawdzające po kolei, czy dana obserwacja jest odstająca. Jeśli jest, to odrzucamy ją z zestawu i wybieramy kolejną, którą następnie testujemy.

Jako przykład podamy test będący uogólnieniem testu Grubbsa, sprawdzający zestaw k największych wartości, przy założeniu normalności populacji oraz nieznanej wartości oczekiwanej i wariancji. Test ten jest oparty na statystyce

$$U = \frac{x_{(n-k+1)} + \ldots + x_{(n-1)} + x_{(n)} - k\bar{x}}{s}.$$

Jest on typu rozproszenie/odchylenie. Wartości krytyczne testu podane są w tabl. III. Ponadto p-wartość możemy oszacować z góry nierównością

$$p \le \binom{n}{k} P\left(T_{n-2} > \sqrt{\frac{n(n-2)U^2}{k(n-k)(n-1) - nU^2}} \right). \tag{2}$$

Równość zachodzi, gdy $U \geq \sqrt{\frac{k^2(n-1)(n-k-1)}{n(k+1)}}$. Gdy $k = 1$, otrzymujemy test Grubbsa.

Przykład 6

Rozpatrzmy następujące dane dotyczące dochodów, podanych w tys. zł, badanych 21 gospodarstw domowych:

1; 1,1; 1,2; 1,3; 1,3; 1,4; 1,5; 1,5; 1,5; 1,6; 1,6; 1,7; 1,8; 1,8; 2; 2,3; 2,3; 2,4; 4; 6,3; 10.

Średnie dochody wynoszą $\bar{x} = 2{,}36$, a odchylenie standardowe $s = 2{,}11$. Sprawdzimy, czy trzy największe obserwacje są odstające ($k = 3$) na poziomie istotności $\alpha = 0{,}05$. Sprawdzian hipotezy jest w tym przypadku równy $U = \frac{x_{(19)} + x_{(20)} + x_{(21)} - 3\bar{x}}{s} = 6{,}26$. Wartości krytyczne testu dla $n = 20$ oraz n = 30 wynoszą odpowiednio 4,11 i 4,56. Te trzy obserwacje możemy więc uznać za odstające. Również prawa strona nierówności (2) jest równa 0,0002, co potwierdza, że obserwacje te możemy uważać za odstające.

Do testów sekwencyjnych zaliczamy test oparty na kurtozie z próby

$$K_n = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}.$$

Pozwala on sprawdzić zarówno najmniejsze, jak i największe wartości, przy założeniach takich jak w poprzednim teście. Wartości krytyczne testu podano w tabl. IV.

Przykład 7

Dane z przykładu 6, $\alpha = 0{,}05$. Sprawdzian hipotezy dla pełnej próby jest równy $K_{21} = 9{,}693$, a po odrzuceniu kolejnych największych obserwacji kurtoza wynosi odpowiednio: $K_{20} = 9{,}478$, $K_{19} = 7{,}496$ oraz $K_{18} = 2{,}346$. Wartości krytyczne są równe $K_\alpha = 4{,}13$ dla $n = 15$ oraz $K_\alpha = 4{,}17$ dla $n = 20$. Trzy największe obserwacje można uznać za odstające, czwartą już nie.

Stosując powyższe metody, zakładaliśmy, że badana cecha w populacji ma rozkład normalny. W przypadku innych rozkładów również istnieją podobne metody znajdowania obserwacji odstających. Na przykład gdy populacja ma rozkład wykładniczy, możemy wykorzystać statystykę postaci [1]

$$Ex = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)}}.$$

Statystyka ta jest bardzo prosta, jednak jest podatna na efekt maskowania. Duża wartość obserwacji $x_{(n-1)}$ może „przysłonić" największą obserwację $x_{(n)}$, podejrzewaną o to, że pochodzi z innej populacji. Można dla statystyki Ex określić dystrybuantę jej rozkładu. Wynosi ona

$$F_n(t) = 1 - n(n-1)B\left(\frac{2-t}{1-t}, n-1\right),$$

gdzie $0 \le t \le 1$, $B(r,s) = \dfrac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$ jest funkcją beta, a $\Gamma(r)$ funkcją gamma.

Przykład 8

Wynotowano czasy obsługi w minutach wybranych losowo 8 klientów:

0,2; 0,9; 1,7; 2,1; 4,1; 6,8; 7,9; 19,7.

Stosując test Kołmogorowa-Smirnowa [2], możemy sprawdzić, że dane pochodzą z rozkładu wykładniczego. Chcąc zbadać, czy obserwacja $x_{(8)} = 19,7$ jest odstająca, wyznaczamy wartość statystyki testowej $Ex = (19,7 - 7,9)/19,7 = 0,599$. p-wartość (*p-value*) jest równa

$$1 - F_t(0,599) = 0,119.$$

Na poziomie istotności $\alpha = 0,05$ nie możemy więc uważać obserwacji $x_{(8)}$ za odstającą.

Gdy dane pochodzą z rozkładu logarytmiczno-normalnego, to możemy skorzystać ze znanego faktu, że ich logarytmy mają rozkład normalny, i stosować wcześniej przedstawione metody dotyczące rozkładu normalnego.

Przykład 9

W tab. 1 przedstawiono uporządkowane dane dotyczące wydatków X w tys. zł (starych) 75 wybranych losowo gospodarstw domowych w Polsce w 1993 roku na transport i komunikację. Dane pochodzą z raportu RAD Project „Powerty and Targeting od Social Assistance in Eastern Europe and Former Soviet Union" Banku Światowego. Pełna próba liczyła 16 051 gospodarstw domowych.

Tabela 1. Wydatki na transport i komunikację w tys. zł (starych)

Lp.	X	Lp.	X	Lp.	X	Lp.	X
1	**9,992**	20	258,760	39	498,427	58	1030,000
2	**80,481**	21	276,477	40	502,651	59	1037,197
3	130,781	22	280,000	41	556,381	60	1162,000
4	134,928	23	302,173	42	564,000	61	1178,483
5	138,423	24	304,201	43	601,913	62	1238,697
6	140,162	25	306,000	44	602,318	63	1264,000
7	145,423	26	309,271	45	602,832	64	1363,434
8	149,315	27	330,997	46	662,703	65	1438,256
9	156,076	28	338,907	47	737,025	66	1486,832
10	160,961	29	343,196	48	794,103	67	1575,202
11	176,143	30	365,372	49	832,000	68	1602,007
12	181,132	31	375,181	50	834,231	69	1661,952
13	202,801	32	420,000	51	854,987	70	1784,646
14	215,633	33	425,286	52	860,000	71	2741,000
15	221,758	34	426,318	53	887,733	72	3092,000
16	222,965	35	436,021	54	899,191	73	3384,615
17	232,884	36	452,124	55	914,158	74	3946,092
18	243,361	37	458,330	56	988,405	75	5526,653
19	253,066	38	466,442	57	1012,989		

Źródło: RAD Project „Powerty and Targeting od Social Assistance in Eastern Europe and Former Soviet Union".

Można sprawdzić, że wartości logarytmów danych mają rozkład normalny. Po zastosowaniu testu Dixona do pary najmniejszych wartości logarytmów danych otrzymujemy wartość testu równą $d_{1,2} = 0{,}407$. Parę obserwacji x_1, x_2 możemy traktować jako obserwacje odstające, ponieważ wartość krytyczna na poziomie istotności $\alpha = 0{,}05$ jest mniejsza niż 0,322. Podobny wynik otrzymujemy, gdy stosujemy uogólniony test Grubbsa. Wartość sprawdzianu testu wynosi 6,986, a wartość krytyczna jest mniejsza niż 5,62. Natomiast obserwacji o największej wartości x_{75} nie można uznać za odstającą. Obydwa testy nie odrzucają wtedy hipotezy zerowej.

Gospodarstwa domowe odpowiadające obserwacjom x_1 i x_2 wydają więc na transport i komunikację istotnie mniej niż pozostałe rozpatrywane gospodarstwa. Możemy je, przeprowadzając całościową analizę wydatków tego typu, zaliczyć do innej kategorii gospodarstw domowych. Nie wykluczone są również w tym przy-

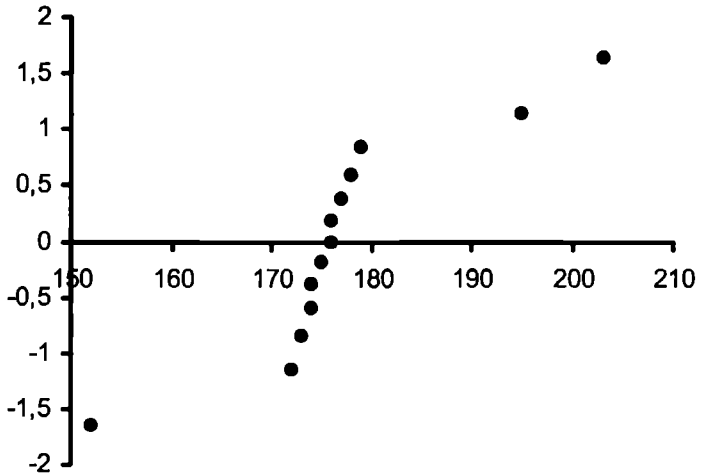
padku pomyłki bądź przekłamania występujące podczas wypełniania ankiety lub jej przetwarzania.

Gdy do wykrywania obserwacji odstających nie można stosować metod wnioskowania statystycznego, nie można np. określić rozkładów prawdopodobieństwa, wówczas możemy skorzystać z metod graficznych. Są one w przypadku jednowymiarowych obserwacji bardzo proste. Wystarczy zaznaczyć na osi liczbowej wartości obserwacji i wyróżnić wartości wyraźnie mniejsze lub większe od pozostałych. Inna metoda graficzna jest oparta na wykresie pudełkowym. Długość wąsów nie powinna przekraczać półtora długości całego pudełka. Obserwacje oddalone na większą odległość od boków pudełka zaznaczamy jako punkty izolowane. Możemy je interpretować jako obserwacje odstające. Można też szczególnie wyróżnić punkty oddalone o odcinek trzykrotnie dłuższy niż bok pudełka. Na obserwacje te należy zwrócić szczególną uwagę podczas dalszej analizy. Mogą być one rzeczywiście obserwacjami odstającymi.

Przykład 10

Dane z przykładu 6. Na rys. 1 przedstawiono wykres pudełkowy dotyczący powyższych danych. Wynika z niego, że możemy w tym przypadku wyróżnić trzy obserwacje odstające $x_{19} = 4$, $x_{20} = 6{,}3$ oraz $x_{21} = 10$, a dwie ostatnie możemy szczególnie wyróżnić.

Rys. 1. Wykres pudełkowy z zaznaczonymi obserwacjami odstającymi
Źródło: opracowanie własne.



Rys. 2. Wykres prawdopodobieństwa normalnego dotyczący danych z przykładu 1
Źródło: opracowanie własne.

Metody graficzne możemy też stosować przed przystąpieniem do wykrywania obserwacji odstających wcześniej przedstawionymi w tym punkcie metodami statystycznymi. Może to być np. wykres prawdopodobieństwa. W przykładzie 1 hipoteza zakładająca normalność rozkładu wzrostu badanych osób nie została odrzucona. Na rys. 2 przedstawiono wykres prawdopodobieństwa normalnego dla obserwacji z tego przykładu. Na osi poziomej odkładane są wartości obserwacji, a na pionowej kwantyle rozkładu normalnego. Na podstawie tego wykresu obserwacje $x_{(1)}$, $x_{(12)}$ i $x_{(13)}$ możemy wstępnie uznać za odstające. Potwierdzono to podczas dokładnej analizy metodami statystycznymi.

3. Metody odporne

Analizując dane, możemy stosować też metody słabo reagujące na różnego rodzaju obserwacje nietypowe. Są to tzw. *metody odporne*. Obserwacje nietypowe nie są tym razem automatycznie usuwane, ale podlegają analizie razem z pozostałymi obserwacjami. Należy w związku z tym zaznaczyć, że zastosowanie metod odpornych nie obejmuje jedynie zagadnień związanych z występowaniem obserwacji nietypowych. Odporność tych metod może też dotyczyć występowania rozkładu innego niż założony, np. zarówno różnych odstępstw od rozkładu normalne-

go, jak i braku spełnienia założonego wcześniej warunku niezależności. Jednak w naszej pracy będziemy dalej omawiać przedstawione metody z punktu widzenia ich odporności na występowanie obserwacji nietypowych, w tym przypadku odstających, głównie zaburzeń.

Wiadomo, że klasyczne estymatory, takie jak miara położenia – średnia, czy rozproszenia – odchylenie standardowe, nie są odporne na obserwacje nietypowe. Nawet odrzucenie jednej obserwacji może istotnie zmienić ich wartości, zwłaszcza odchylenia standardowego.

Przykład 11

Rozpatrzmy obroty w tys. zł 10 spółek: Trastychy, Atlantis, MNI, Kopex, Amica, Indykpol, Kruszwica, BRE, Swarzędz i Forte. Przedstawione w tab. 2 uporządkowane dane pochodzą z 10 września 2005 roku.

Tabela 2. Obroty spółek w tys. zł

spółka	Tra	Atl	MNI	Kop	Ami	Ind	Kru	BRE	Swa	For
obroty	312	379	415	430	437	625	694	798	801	1320

Źródło: „Gazeta Wyborcza" z 10.09.2005.

Na podstawie zarówno testu Dixona, jak i testu Grubbsa spółkę Forte możemy uważać za obserwację odstającą. W pierwszym przypadku wartość testu wynosi 0,515, przy wartości krytycznej równej 0,412 na poziomie istotności $\alpha = 0{,}05$. Dla testu Grubbsa otrzymujemy natomiast wartość testu 2,309 i wartość krytyczną 2,18. Można więc uważać, że spółka Forte ma obroty istotnie większe od pozostałych rozpatrywanych spółek, i zaliczyć ją do innej kategorii.

Powyższe rozważania mają istotne znaczenie przy wyznaczaniu podstawowych parametrów charakteryzujących badaną grupę spółek. Średnie obroty wyznaczone dla 9 spółek, po odrzuceniu Forte, są równe 543,44 tys. zł, a odchylenie standardowe 187,70. Są one istotnie różne od wartości uzyskanych dla całej grupy spółek. Średnia wynosi wtedy 621,10 tys. zł, a odchylenie 302,67. Różnica jest duża zwłaszcza w przypadku odchylenia standardowego. Odrzucenie zaś w przykładzie 9 obserwacji uznanych za odstające nie spowoduje powstania aż tak dużej względnej różnicy wartości tych charakterystyk. Średnie wydatki na transport i komunikację wyznaczone dla całej próby wynoszą 810,54 tys. starych zł, a po odrzuceniu dwóch obserwacji odstających otrzymujemy 831,51 tys. zł. Dla odchylenia standardowego różnica ta jest znikoma: przed odrzuceniem obserwacji odstających mamy 935,86 tys. zł, a po odrzuceniu – 939,9 tys. zł – jest to spowodowane m.in.

dość oczywistym faktem, że dla większych prób odrzucenie małej liczby obserwacji ma dużo mniejszy wpływ na wartości charakterystyk.

Przedstawimy teraz najczęściej stosowane w zagadnieniach związanych z analizą odpornościową estymatory parametrów położenia i rozproszenia stosowane w praktyce. Niech $x_1, ..., x_n$ będzie n-elementową próbką pochodzącą z badanej populacji. Na jej podstawie wyznaczymy wielkości charakteryzujące naszą populację. Dla każdego estymatora podamy też wartość *punktu załamania* (ang. *breakdown point*) ε^*, czyli najmniejszego udziału zaburzeń w próbie, powyżej którego estymator się załamuje. Dokładną definicję i więcej informacji na ten temat czytelnik znajdzie w [3; 5; 6].

Zaczniemy od *miar położenia*. Będą one oparte głównie na statystykach pozycyjnych (por. [5]).

I. *Statystyki pozycyjne*

Niech $x_{(i)}$ będzie i-tą z kolei wartością obserwacji po ich uprzednim uporządkowaniu w porządku niemalejącym. Tak określoną wartość będziemy nazywać *i-tą statystyką pozycyjną*. Środkową wartość nazywamy *medianą* $Me = x_{([0,5n])}$. Z teoretycznego punktu widzenia dowolna liczba z przedziału $[x_{(m)}, x_{(m+1)})$ może być uważana za medianę. Uogólnieniem mediany jest *q-ty kwanty*:

$$Q_q = x_{([qn])},$$

gdzie $0 \le q \le 1$, a dla $q = 0$ zakładamy, że $[qn] = 1$. Gdy $q = 0,5$, otrzymujemy medianę, gdy $q = 0$ – najmniejszą wartość obserwacji $x_{(1)}$, a gdy $q = 1$ – największą $x_{(n)}$. Punkt załamania dla q-kwantyla wynosi $\varepsilon^* = 1 - q$. Dla mediany otrzymujemy wartość $\varepsilon^* = 0,5$. Połowę obserwacji możemy wtedy w miarę bezpiecznie odrzucić [3].

II. *($\alpha;\beta$)-obcięta średnia*:

$$\bar{x}_{\alpha;\beta} = \frac{1}{r-m} \sum_{j=m+1}^{r} x_{(j)},$$

gdzie $0 \le \alpha, \beta \le 0,5$, $m = [\alpha n]$ oraz $r = n - [\beta n]$. W tym przypadku odrzucamy (α100)% najmniejszych i (β100)% największych obserwacji, a następnie obliczamy zwykłą średnią arytmetyczną. Gdy $\alpha = 0$ średnia jest obcięta prawostronnie, a $\beta = 0$ lewostronnie oraz gdy $\alpha = \beta$, wtedy mamy do czynienia z symetrycznym 2β-obcięciem. Gdy $\alpha = \beta$, wówczas punkt załamania $\varepsilon^* = \beta$. Dla średniej jest więc równy zeru. Jedna obserwacja może w tym przypadku istotnie zmienić wartość tego estymatora.

III. *Średnia typu* ($\alpha;\beta$)*-Winsora*

Tego typu średnią obliczmy, zastępując (α100)% najmniejszych i (β100)% największych obserwacji odpowiednio przez $m+1$-szą i r-tą obserwację, w której $0 \le \alpha,\beta \le 0{,}5$, $m = [\alpha n]$ oraz $r = n - [\beta n]$. Przybiera ona postać

$$W_{\alpha;\beta} = \frac{1}{n}\left(\sum_{j=m+1}^{r} x_{(j)} + m x_{(m+1)} + (n-r) x_{(r)} \right).$$

Średnie obcięte oraz typu Winsor są prostymi estymatorami reagującymi na obserwacje nietypowe. Polegają one na usunięciu skrajnych obserwacji lub ich zastąpieniu przez bardziej typowe wielkości. W zastosowaniach praktycznych pojawiają się problemy dotyczące symetrii ucięcia oraz wartości parametrów α i β. Rozwiązanie ich nie jest proste ani jednoznaczne. Na przykład w odniesieniu do symetrycznych ucięć możemy się spotkać z propozycjami $\alpha = 0{,}05$; 0,1; 0,15 oraz 0,25 i oczywiście z medianą ($\alpha = 0{,}5$).

Przy wyznaczaniu miary położenia danych pochodzących z rozkładu normalnego $N(\mu, \sigma)$ często stosuje się prostą metodę zaproponowaną przez Anscombe'a. Wartość parametru μ wyznaczamy wtedy, stosując formułę:

$$\begin{cases} \overline{x}, & \text{gdy } |z_i| < c\sigma \text{ dla wszystkich } i, \\ \overline{x}_{1/n;0}, & \text{gdy } |z_{(1)}| \ge c\sigma \text{ i } |z_{(1)}| > |z_{(n)}|, \\ \overline{x}_{0;1/n}, & \text{gdy } |z_{(n)}| \ge c\sigma \text{ i } |z_{(1)}| < |z_{(n)}|, \end{cases}$$

gdzie $z_i = x_i - \overline{x}$, a stała c jest ustalana arbitralnie, np. $c = 2$. W przypadku nieznanego odchylenia standardowego podstawiamy odchylenie standardowe z próby, tzn. $\sigma = s$. Podobnie postępujemy w odniesieniu do średniej Winsora. Estymator Anscombe'a możemy interpretować jako średnią wyznaczoną dla grupy obserwacji powstałej po odrzuceniu obserwacji pochodzących z zaburzenia o rozkładzie $N(\mu + \Delta, \sigma)$ [1].

Inna metoda stosowana w praktyce polega na usunięciu ze zbioru obserwacji pewnej liczby danych ekstremalnych, spełniających określony warunek, a następnie obliczeniu średniej dla pozostałych danych. Warunkiem tym może być spełnienie nierówności

$$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)}} \ge c,$$

gdzie c jest ustaloną stałą.

IV. *L-estymatory*

Są to kombinacje liniowe statystyk pozycyjnych:

$$L = \sum_{i=1}^{n} c_i x_{(i)},$$

gdzie współczynniki c_i przybierają mniejsze wartości względem obserwacji ekstremalnych, zmniejszając ich wpływ na wartość estymatora, czyli zwiększając jego odporność. Współczynniki te powinny być unormowane, tzn. spełniać warunek $\sum_{i=1}^{n} c_i = 1$. W niektórych, skrajnych przypadkach dopuszcza się nawet wartości ujemne tych wag [3].

L-estymatory są uogólnieniem wcześniej rozpatrywanych estymatorów. Po przyjęciu jednakowych wartości wag $c_i = 1/n$ otrzymujemy średnią; mediana odpowiada sytuacji, gdy wagi przybierają wartości zerowe dla wszystkich obserwacji poza środkowymi, czyli $c_m = 1$ dla nieparzystego n oraz $c_m = c_{m+1} = 0{,}5$ dla parzystego. W podobny sposób otrzymujemy kwantyle Q_q. Obcięta średnia powstaje, gdy $c_1 = c_2 = \ldots = c_m = c_{r+1} = c_{r+2} = \ldots = c_n = 0$ oraz $c_i = 1/(r - m)$ dla $m + 1 \le i \le r$, a Winsora, gdy $c_i = 1/n$ dla $m+1 \le i \le r - 1$, $c_m = (m+ 1)/n$, $c_r = (n - r + 1)/n$ oraz $c_i = 0$ dla pozostałych. W przypadku symetrycznego rozkładu, np. normalnego, zakłada się zwykle, że wagi spełniają warunek $c_i = c_{n+1-i}$.

Innymi przykładami *L*-estymatora są: tzw. *trimean*, oparty na medianie i kwartylach,

$$Tm = 0{,}25(Q_{0{,}25} + 2Me + Q_{0{,}75}),$$

estymator Gastwirtha

$$G = 0{,}3x_{([n/3]+1)} + 0{,}4Me + 0{,}3x_{(n-[n/3])}$$

oraz liniowa kombinacja statystyk pozycyjnych postaci

$$L = (x_{(2)} + 3x_{(3)} + 5x_{(4)} + \ldots + (2m - 3)x_{(m)} + (2m - 3)x_{(m+1)} + (2m - 5)x_{(m+3)} + \ldots + 3x_{(n-2)} + x_{(n-1)})/a,$$

gdzie $a = 2(m - 1)^2$ (suma wszystkich współczynników znajdujących się przy $x_{(i)}$) dla n parzystego [1].

Przykład 12

W tab. 3 zawarte są wartości następujących wybranych miar położenia: średniej ($\bar{x}$), mediany (*Me*), 0,5%- i 0,10%-obciętych średnich ($\bar{x}_{0;5\%}, \bar{x}_{0;10\%}$) oraz średnich typu 0,5%- i 0,10%-Winsora ($W_{0;5\%}, W_{0;10\%}$), trimen (T) i L dla danych z przykładu 6.

Tabela 3. Wartości wybranych miar położenia

Miara położenia	$\bar{x}$	Me	$\bar{x}_{0;5\%}$	$\bar{x}_{0;10\%}$	$W_{0;5\%}$	$W_{0;10\%}$	T	L
Wartość	2,36	1,60	1,98	1,75	2,19	1,97	1,73	1,75

Źródło: opracowanie własne.

Widać, że w przykładzie tym występują istotne różnice między wartościami poszczególnych miar położenia. Największą wartość przybiera średnia, a najmniejszą mediana. Chcąc więc wyznaczyć przeciętny stan dochodów za pomocą różnych miar położenia, otrzymamy wartości leżące między 1,60 tys. zł a 2,36 tys. zł. Różnice te wynikają głównie z prawostronnej symetrii badanego rozkładu oraz z występowania dwóch wyraźnie odstających obserwacji. Inna sytuacja wystąpi, gdy rozkład będzie symetryczny, i to nawet jeśli będą odstające obserwacje. Wtedy wartości różnych miar położenia mogą być prawie identyczne.

Przedstawmy teraz najczęściej stosowane odporne *miary rozproszenia* [5].

1. *Mediana odchylenia bezwzględnego* (ang. *median absolut deviation* – MAD). Jest to środkowe odchylenie od mediany

$$MAD = Me\{|x_1 - Me|, |x_2 - Me|, ..., |x_n - Me|\},$$

gdzie *Me* oznacza medianę zbioru danych $x_1, ..., xn$.

2. *Liniowa kombinacja kwantyli*:

$$Q_{\alpha,\beta} = c(Q(\beta) - Q(\alpha)),$$

gdzie $0 \le \alpha \le \beta \le 1$. Zwykle zakłada się, że kwantyle są symetryczne, czyli $\alpha = 1 - \beta$, a stała c jest tak dobrana, aby spełnione były warunki regularności otrzymanych estymatorów. W przypadku symetrycznych kwantyli, gdy $c = 0,5$, $\alpha = 0,25$, otrzymamy odchylenie ćwiartkowe, a gdy $c = 1$, $\alpha = 0$, najprostszą miarę rozproszenia – rozstęp.

3. *(α; β)-obcięte standardowe odchylenie*:

$$s_{\alpha;\beta} = \sqrt{\frac{1}{r-m}\sum_{j=m}^{r}\left(x_{(j)} - \overline{x}_{\alpha;\beta}\right)^2},$$

gdzie $0 \le \alpha, \beta \le 0{,}5$, $m = [\alpha n]$ oraz $r = n - [\beta n]$. W podobny sposób można określić standardowe odchylenie (α; β)-typu Winsora.

4. *L-estymatory*. Są to kombinacje liniowe statystyk pozycyjnych

$$s_L = \sum_{i=1}^{n} b_i x_{(i)},$$

dla których wagi, mogące być ujemne, sumują się do zera. Po przyjęciu, że $b_1 = -1$, $b_2 = b_3 = \ldots = bn_{-1} = 0$ i $bn = 1$, otrzymujemy rozstęp. Liniowa kombinacja kwantyli jest oczywiście również szczególnym przypadkiem L-estymatora.

Przykład 13

W tab. 4 zawarte są wartości następujących wybranych miar rozproszenia: odchylenia standardowego (s), MAD, odchylenia ćwiartkowego (Q), 0,5%- i 0,10% -obciętych odchyleń ($s_{0;5\%}, s_{0;10\%}$) oraz odchyleń typu 0,5%- i 0,10%-Winsora ($s^W_{0;5\%}, s^W_{0;10\%}$) dla danych z przykładu 6.

Tabela 4. Wartości wybranych miar rozproszenia

Miara rozproszenia	s	MAD	Q	$s_{0;5\%}$	$s_{0;10\%}$	$s^W_{0;5\%}$	$s^W_{0;10\%}$
Wartość	2,11	0,3	0,45	1,18	0,66	2,17	0,83

Źródło: opracowanie własne.

Widać, że najmniejszą wartość ma mediana odchylenia bezwzględnego MAD = 0,3, a największą odchylenie Winsora $s^W_{0;5\%} = 2{,}17$. Ponadto odchylenia Winsora przybierają wartości większe niż ucięte odchylenia. Wartości obydwu estymatorów są większe dla mniejszej wartości granic α i β.

4. Zakończenie

W pracy przedstawiono wybrane metody wykrywania obserwacji odstających oraz odporne metody estymacji. Nie są to oczywiście wszystkie tego typu metody.

Obszerne omówienie tych metod znajduje się w pracach [1; 3]. Należy też podkreślić, że nie ma w literaturze jednoznacznych i precyzyjnych definicji obserwacji nietypowych. Dalsza część rozważań dotycząca nietypowych obserwacji, omawiająca obserwacje wielowymiarowe, znajduje się w [4].

Dodatek – tablice statystyczne

Tablica I. Wartości krytyczne dla testu Dixona

n	$d_{1:0,90}$	$d_{1:0,95}$	$d_{2:0,90}$	$d_{2:0,95}$	$d_{3:0,90}$	$d_{3:0,95}$
3	0,886	0,941				
4	0,679	0,765	0,910	0,955	0,935	0,967
5	0,557	0,642	0,728	0,807	0,782	0,845
6	0,482	0,560	0,609	0,689	0,670	0,736
7	0,434	0,507	0,530	0,610	0,596	0,661
8	0,399	0,468	0,479	0,554	0,545	0,607
9	0,370	0,437	0,441	0,512	0,505	0,565
10	0,349	0,412	0,409	0,477	0,474	0,531
11	0,332	0,392	0,385	0,450	0,449	0,504
12	0,318	0,376	0,367	0,428	0,429	0,481
13	0,305	0,361	0,350	0,410	0,411	0,461
14	0,294	0,349	0,336	0,395	0,395	0,445
15	0,285	0,338	0,323	0,381	0,382	0,430
16	0,277	0,329	0,313	0,369	0,370	0,418
17	0,269	0,320	0,303	0,359	0,359	0,406
18	0,263	0,313	0,295	0,349	0,350	0,397
19	0,258	0,306	0,288	0,341	0,341	0,379
20	0,251	0,300	0,282	0,334	0,333	0,372
21	0,247	0,295	0,276	0,327	0,326	0,365
22	0,242	0,290	0,270	0,320	0,320	0,358
23	0,238	0,285	0,265	0,314	0,314	0,352
24	0,234	0,281	0,260	0,309	0,309	0,347
25	0,230	0,277	0,255	0,304	0,304	0,343
26	0,227	0,273	0,250	0,299	0,300	0,338
27	0,224	0,269	0,246	0,295	0,296	0,334
28	0,220	0,266	0,243	0,291	0,292	0,330
29	0,218	0,263	0,239	0,287	0,288	0,326
30	0,215	0,260	0,236	0,283	0,285	0,322

Źródło: [2].

Tablica II. Wartości krytyczne dla testu Grubbsa

n	0,05	0,01	n	0,05	0,01	n	0,05	0,01	n	0,05	0,01
3	1,15	1,15	8	2,03	2,22	15	2,41	2,71	40	2,87	3,24
4	1,46	1,49	9	2,11	2,32	16	2,44	2,75	50	2,96	3,34
5	1,67	1,75	10	2,18	2,41	18	2,50	2,82	60	3,03	3,41
6	1,82	1,94	12	2,29	2,55	20	2,56	2,88	100	3,21	3,60
7	1,94	2,10	14	2,37	2,66	30	2,74	3,10	120	3,27	3,66

Źródło: [1].

Tablica III. Wartości krytyczne dla testu U

n	$k = 2$		$k = 3$		$k = 4$	
	0,05	0,01	0,05	0,01	0,05	0,01
5	2,10	2,16				
6	2,41	2,50				
7	2,66	2,79	2,97	3,08		
8	2,87	3,02	3,39	3,42		
9	3,04	3,22	3,58	3,73	3,82	3,98
10	3,18	3,40	3,82	4,00	4,17	4,34
12	3,44	3,70	4,24	4,44	4,72	4,92
14	3,66	3,92	4,57	4,83	5,20	5,42
16	3,83	4,10	4,85	5,14	5,60	5,85
18	3,96	4,25	5,08	5,38	5,91	6,20
20	4,11	4,41	5,30	5,60	6,22	6,54
30	4,56	4,92	6,03	6,41	7,29	7,64
40	4,84	5,29	6,49	6,98	7,93	8,38
50	5,06	5,51	6,82	7,34	8,38	8,88
100	5,62	6,06	7,77	8,27	9,71	10,30

Źródło: [1].

Tablica IV. Wartości krytyczne dla testu K

n	0,05	0,01	n	0,05	0,01	n	0,05	0,01	n	0,05	0,01
5	2,90	3,10	12	4,05	5,20	40	4,06	5,04	500	3,37	3,60
7	3,55	4,23	15	4,13	5,30	50	3,99	4,88	1000	3,26	3,41
8	3,70	4,53	20	4,17	5,36	75	3,87	4,59			
9	3,86	4,82	25	4,16	5,30	100	3,77	4,39			
10	3,95	5,00	30	4,11	5,21	200	3,57	3,98			

Źródło: [1].

Literatura

[1] Barnett V., Levis T., *Outliers in Statistic Data* (wyd. 3), J. Wiley & Sons, Chichester 1995.
[2] Domański C., *Testy statystyczne*, PWE, Warszawa 1990.
[3] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A., *Robust Statistics*, J. Wiley & Sons, New York 1986.
[4] Heilpern S., *Obserwacje nietypowe – przypadek wielowymiarowy*, Prace Naukowe Akademii Ekonomicznej nr 1097, AE, Wrocław 2005.
[5] *Statystyczne metody analizy danych*, red. W. Ostasiewicz, AE, Wrocław 1998.
[6] Staude R.G., Sheather S.J., *Robust Estimation and Testing Robust Statistics*, J. Wiley & Sons, New York 1990.

UNTYPICAL REALIZATIONS OF THE ONE-DIMENSIONAL RANDOM VARIABLES

Summary

The paper is devoted to the untypical realizations of the one-dimensional random variables. This is a review, which presents selected methods of the detection of the outliers in univariate data. These are the methods based on the statistical inference or on the graphical methods. They detect the single outliers or the group of them. The methods used in the normal population case, such as: Dixon's, Grubbs' – ordinary and generalized and based on the kurtosis are presented. The method used in the exponential case studied is also. The second part of the paper is devoted to the methods of the robust estimation.

The paper contains the examples illustrating presented methods and indicates their potential applications. There are four tables with the critical values of the statistical tests used to the detection of the outliers in the appendix.