WIERTNICTWO NAFTA GAZ • TOM 23/2 • 2006

Kazimierz Twardowski*, Jacek Traple*

UWAGI DOTYCZĄCE WĄTPLIWYCH WYNIKÓW POMIARÓW**

1. PROBLEM TZW. WĄTPLIWYCH WYNIKÓW POMIARÓW – UWAGI WSTĘPNE

Jednym z najbardziej kłopotliwych i kontrowersyjnych problemów w praktyce metrologicznej jest sposób postępowania z tzw. wątpliwymi wynikami pomiarów.

Zgodnie z normą PN-87 [10], przez wątpliwy rozumie się wynik różniący się od pozostałych wyników w stopniu przewyższającym różnice, których można się spodziewać w danej metodzie pomiarów. Tego typu sytuacja dotyczy opracowania wyników zestawionych w formie serii pomiarowej, traktowanej w kategoriach statystycznych jako empiryczna próba losowa reprezentująca badaną zbiorowość (populację) generalną (tj. będącą jej podzbiorem).

Jest to między innymi konsekwencja uznania samych pomiarów za, w gruncie rzeczy, formowanie próby statystycznej z nieskończonego zbioru pomiarów, które potencjalnie można przeprowadzić, czyli z populacji o nieskończonej liczbie elementów [18]. Z punktu widzenia rachunku prawdopodobieństwa i statystyki matematycznej oznacza to, że wyniki pomiarów traktowane są jako realizacje jednowymiarowej ciągłej zmiennej losowej [12, 15, 17, 21].

W praktyce problemy decyzyjne związane z serią pomiarową zawierającą wątpliwe wyniki pomiarów mogą być dwojakiego rodzaju:

1) W serii występują pojedyncze wyniki wątpliwe, określane w literaturze przedmiotu jako:
   - wartości odstające [13, 14],
   - wartości odbiegające [22],
   - wartości podejrzane (w znaczeniu mało prawdopodobne, mało wiarygodne) [14, 17].

   Wyniki tego typu ekstremalne (skrajne) w analizowanej serii, które po weryfikacji statystycznej są uznawane za nienależące do badanej populacji, utożsamiane są z wynikami obarczonymi błędem grubym [10, 14, 21].

* Wydział Wiertnictwa, Nafty i Gazu AGH, Kraków.

** Praca wykonana w ramach badań statutowych Wydziału Wiertnictwa, Nafty i Gazu AGH, Kraków

2) Seria zawiera dwa lub więcej wyników wątpliwych, co oznacza zwykle jej niejednorodność w sensie statystycznym – weryfikacja statystyczna wyników wątpliwych, w szczególności z wykorzystaniem testów służących do weryfikacji zespołu wyników w próbie [10], prowadzić może do ich wykluczenia z dalszej procedury opracowania danych pomiarowych w razie konieczności związanej np. z wymaganą liczebnością serii pomiarowej.

2. PRZYCZYNY WYSTĘPOWANIA WĄTPLIWYCH WYNIKÓW POMIARÓW

2.1. Pojedynczy wynik obarczony błędem grubym

Jedną z przyczyn występowania w serii pomiarowej wątpliwych wyników pomiarów są błędy grube, określane czasem jako omyłki lub błędy nadmierne [2]. Powstają one w efekcie jednorazowego wpływu istotnej przyczyny działającej przejściowo i występują tylko przy niektórych pomiarach (np. pomyłka przy odczycie lub zapisie wskazania przyrządu pomiarowego, niewłaściwy sposób pobrania, przechowywania lub przygotowania próbki do badań laboratoryjnych itp). Pojedynczy wynik pomiarowy obarczony błędem grubym jest, z definicji, skrajną wartością (minimalną lub maksymalną) w uporządkowanej wg rosnących lub malejących wartości serii pomiarowej. Szczególnie łatwy jest do wykrycia i identyfikacji w odniesieniu do serii pomiarowych zawierających wyniki pomiarów dokonywanych w spełnionych warunkach powtarzalności, tj. takich, w których niezależne wyniki badania takich samych jednostek są otrzymane tą samą metodą, w tym samym laboratorium, przez tego samego operatora z użyciem tego samego wyposażenia, w krótkich odstępach czasu [12–14].

W takich warunkach zmienność wyników pomiarów uzależniona jest tylko od błędów losowych (przypadkowych), których rozkład zgodnie z założeniem potwierdzonym praktycznie odpowiada teoretycznemu rozkładowi normalnemu (Gaussa) zmiennej losowej o zerowej wartości oczekiwanej i odchyleniu standardowym związanym z precyzją pomiarów w warunkach powtarzalności. W odniesieniu do poprawnej metody pomiarowej pojedynczy wynik i-tego pomiaru x_i można przedstawić równaniem:

$$x_i = \mu_x + e_i \qquad (2.1)$$

gdzie:

- μ_x – przyjęta wartość odniesienia mierzonej wielkości x rozumiana jako uzgodniona wartość odniesienia na cele porównawcze i oceniana najczęściej jako wartość średniej arytmetycznej serii pomiarowej,
- e_i – błąd losowy i-tego pomiaru o wartości oczekiwanej równej zeru oraz odchyleniu standardowym określanym jako odchylenie standardowe powtarzalności i będącym miarą precyzji metody.

W związku z powyższym stosowane w praktyce metrologicznej sposoby wykrywania wyników pomiarów obarczonych błędem grubym opierają się na znanych z rachunku prawdopodobieństwa i statystyki matematycznej regułach dotyczących zakresów (przedziałów) wartości zmiennej losowej o rozkładzie normalnym – w szczególności prakseologicznej reguły trzech sigm [2, 9, 15, 21]. W praktyce procedury te mogą być stosowane do analizy serii pomiarowych charakteryzujących się rozkładami w przybliżeniu normalnymi, pod warunkiem że są one rozkładami jednomodalnymi [13, 14].

2.2 Seria pomiarowa niejednorodna statystycznie

Zgodnie z normami o statusie PN-ISO wydanymi przez Polski Komitet Normalizacyjny w roku 2002 i dotyczącymi metod oraz wyników pomiarów [12–14] istotnej zmianie uległy definicje oraz interpretacje niektórych podstawowych pojęć metrologicznych zawartych w wydanym przez Główny Urząd Miar przewodniku mającym charakter metodycznej instrukcji dotyczącej niepewności i błędów pomiarowych [23].

Najważniejsze wprowadzone zmiany omówiono w pracy *Próba oceny metody porometrii helowej do pomiarów laboratoryjnych porowatości skał* [20]. Związane są one w szczególności z pojęciami dokładności, poprawności i precyzji, a także warunków powtarzalności i warunków odtwarzalności badania. Jak wspomniano w rozdziale 2.1, wyniki pomiarów dokonywanych w spełnionych warunkach powtarzalności (2.1) zależne są tylko od błędów losowych. W związku z tym serie pomiarowe z tego typu wynikami z definicji cechować się powinny jednorodnością statystyczną i rozkładami zbliżonymi do rozkładu normalnego.

Zupełnie inaczej może się przedstawiać sytuacja w odniesieniu do serii pomiarowych zawierających wyniki pomiarów dokonywanych w spełnionych warunkach odtwarzalności, tj. warunkach, w których wyniki badań takich samych jednostek są otrzymywane za pomocą tej samej metody w różnych laboratoriach przez różnych operatorów z użyciem różnego wyposażenia [12–14]. Różnica między wartością oczekiwaną wyników badania, której estymatorem jest średnia arytmetyczna zbioru wyników występujących w serii a przyjętą wartością odniesienia określa poprawność wyników badania, której miarą jest obciążenie obrazujące ich całkowity błąd systematyczny. W razie zastosowania niepoprawnej metody pomiarowej pojedynczy wynik *i*-tego pomiaru (x_i) może być przedstawiony za pomocą relacji:

$$x_i = \mu_x + \Delta x_{syst} + e_i \tag{2.2}$$

gdzie:

μ_x – przyjęta wartość odniesienia mierzonej wielkości;

e_i – błąd losowy *i*-tego pomiaru, ale w warunkach odtwarzalności (tj. z zerową wartością oczekiwaną i odchyleniem standardowym odtwarzalności);

Δx_{syst} – całkowity błąd systematyczny wyników pomiarów uwzględniający składowe związane z metodą badania, konkretnym laboratorium wraz z jego wyposażeniem, a także z operatorem.

W odróżnieniu od nieprzewidywalnego błędu losowego e_i wyniku badania, a co za tym idzie, niemożliwego do skorygowania, błąd systematyczny pozostaje stały lub zmienia się w sposób przewidywalny, a więc i pozwalający na korektę.

Występowanie błędu systematycznego w warunkach pomiarów ujętych w formie serii pomiarowej przejawia się zwykle w jej niejednorodności statystycznej, czego skutkiem jest wielomodalny, najczęściej bimodalny (dwuwierzchołkowy), charakter rozkładu wartości pomiarowych.

Sytuacje tego rodzaju mogą wystąpić przede wszystkim w seriach pomiarowych zawierających wyniki pomiarów, które nie były prowadzone w warunkach powtarzalności. Może to dotyczyć w szczególności takich sprawczych czynników, jak różne laboratoria bądź różne metody badawcze. Możliwe są również tego typu sytuacje spowodowane niejednorodnością statystyczną jednostek badania, których dotyczą wyniki ujęte w serii pomiarowej. Przy tym można mówić o jednorodności statystycznej zarówno w sensie jakościowym (nieparametrycznym), jak i ilościowym (parametrycznym) [8]. Kryteria jednorodności jakościowej, które – najogólniej mówiąc – determinują wybór określonej populacji generalnej do badań, w zasadzie powinny być całkowicie zdefiniowane na podstawie merytorycznych przesłanek wynikających z postawionego zadania, w którym ta populacja występuje jako przedmiot badania.

Przykładowo jeżeli rozwiązujemy zadanie klasyfikacji skał karbonu produktywnego Górnośląskiego Zagłębia Węglowego na poszczególne typy litologiczne, populacja generalna obejmuje ogół warstw (próbek skalnych) niezależnie od ich charakteru litologicznego. Jeśli jednak postawimy zadanie klasyfikacji poszczególnych odmian węgla na typy technologiczne, wówczas do badanej populacji generalnej wchodzić będą tylko warstwy węgla. Populacja jakościowo jednorodna w pierwszym zadaniu – byłaby niejednorodna w drugim.

Próba statystyczna (seria pomiarowa) jest jakościowo niejednorodna, jeśli jej część lub części nie należą do badanej, merytorycznie zdefiniowanej populacji generalnej. W konkretnych zadaniach badawczych, związanych z naukami o Ziemi, ustalanie jednorodnych jakościowo prób statystycznych musi być zwykle poprzedzone również ustaleniem przestrzennego (powierzchniowo-głębokościowego) zakresu populacji generalnej.

Próby statystyczne niejednorodne w sensie jakościowym cechują się także z reguły niejednorodnością w sensie ilościowym, co w praktyce oznacza, że wykazują cechy podzielności na podzbiory (części prób) różniące się istotnie opisującymi je parametrami statystycznymi. W odniesieniu do serii pomiarowych będzie to oczywiście dotyczyło przede wszystkim miar położenia, w szczególności klasycznych średnich.

Z rachunku prawdopodobieństwa oraz z praktyki statystycznej wynika, że na ogół wystarczającym sprawdzianem jednorodności statystycznej próby (serii pomiarowej) może być weryfikacja hipotezy o jednomodalności (jednowierzchołkowości) jej empirycznej krzywej rozkładu [8, 10, 14]. Wynika to stąd, że geneza rozkładów wielomodalnych związana jest z mieszaniem się dwu lub więcej odrębnych statystycznie, jednorodnych zbiorów w jeden niejednorodny.

3. SPOSOBY POSTĘPOWANIA Z WĄTPLIWYMI WYNIKAMI POMIARÓW

3.1. Uwagi ogólne

Problem wątpliwych wyników pomiarów w praktyce metrologicznej, w szczególności w trakcie wszelkiego rodzaju badań, występuje stosunkowo często. W takim wypadku eksperymentator musi podjąć jedną z najtrudniejszych decyzji o włączeniu lub wykluczeniu tych wyników z dalszej analizy. W razie stwierdzenia, że przyczyną anomalnego wyniku była ewidentna pomyłka operatora lub zaburzenia w pracy aparatury, taki wynik powinien być natychmiast odrzucony.

Niestety zwykle niemożliwe jest ustalenie zewnętrznej przyczyny uzyskania anomalnego wyniku. Zachodzi wówczas konieczność podjęcia decyzji dotyczącej jego odrzucenia lub pozostawienia, opierając się jedynie na wynikach ujętych w całej serii pomiarowej. Wyszukiwanie argumentów przemawiających za odrzuceniem wyników wykonanych pomiarów jest wysoce ryzykowne, gdyż istnieje wówczas możliwość podjęcia decyzji tendencyjnych. Kwestia jest ważna i dyskusyjna; nawet wśród ekspertów spotkać można różne opinie na ten temat [17, 18, 21]. Decyzja o odrzuceniu danych nosi charakter mniej lub bardziej subiektywny i może narazić badacza, który ją podejmuje, na zarzut niezgodnej ze sztuką prezentacji swoich wyników. Sytuację pogarsza fakt, że taki nietypowy wynik może odzwierciedlać jakiś ważny rzeczywisty efekt fizyczny. Faktem jest, że wiele ważnych odkryć naukowych w przeszłości po raz pierwszy objawiło się jako wynik badań rażąco odbiegający od pozostałych.

W gruncie rzeczy jedynym rzetelnym sposobem postępowania w odniesieniu do wątpliwych wyników jest powtarzanie pomiaru, czasem nawet wielokrotne. Z różnych względów nie zawsze jest to możliwe, dlatego racjonalnym i w miarę obiektywnym sposobem podejmowania decyzji dotyczących wyników wątpliwych jest weryfikacja odpowiednich parametrycznych hipotez statystycznych z wykorzystaniem właściwych testów istotności.

W literaturze przedmiotu spotkać można całą gamę różnych kryteriów i testów statystycznych, od bardzo prostych do stosunkowo skomplikowanych. W nawiązaniu do przyczyn i sposobu występowania wątpliwych wyników w seriach pomiarowych wykorzystywane w praktyce kryteria i testy statystyczne podzielić można na dwie grupy, dotyczące odpowiednio pojedynczych wyników wątpliwych oraz zespołu (dwu lub więcej) wyników wątpliwych.

Przedstawiona poniżej krótka charakterystyka ważniejszych metod statystycznych weryfikacji pojedynczego wyniku wątpliwego w serii dotyczy następujących przyjętych oznaczeń i założeń:

- Wyniki n pomiarów mierzonej wielkości x przedstawione są w postaci uporządkowanego ciągu niemalejącego, w postaci

$$x_1 \leq x_2 \leq \ldots \leq x_n,$$

  który reprezentuje próbę statystyczną w formie empirycznego szeregu statystycznego szczegółowego (wyliczającego). W tej sytuacji wynikiem wątpliwym (x_w) może być jeden ze skrajnych wyników x_1 lub x_n.

- Badana wielkość x cechuje się w populacji generalnej rozkładem normalnym $N(\mu,\sigma)$ o nieznanych parametrach: wartości oczekiwanej μ oraz odchyleniu standardowym σ.
- Rozkład statystyczny wyników pomiarów w serii jest rozkładem jednomodalnym, zbliżonym do normalnego $N(\bar{x};S)$ o parametrach: średniej arytmetycznej $\bar{x}$ oraz odchyleniu standardowym z próby S w wersji oceny nieobciążonej, obliczanych według wzorów:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{3.2}$$

Podczas obliczania średniej i odchylenia standardowego z próby dla serii pomiarów, po odrzuceniu wyniku wątpliwego x_w, średnia i odchylenie oznaczane są odpowiednio jako $\bar{x}_{odrz}$ i S_{odrz}.

– W wypadku prób statystycznych bardzo małych – serii pomiarowych o liczebności $n \leq 10$ – odchylenie standardowe σ szacowane powinno być z rozstępu R według wzoru [9, 21]:

$$\hat{\sigma} = R\frac{1}{d_n} = S_R \tag{3.3}$$

gdzie:

$R = x_{max} - x_{min} = x_n - x_1$;

$\frac{1}{d_n}$ – współczynnik konwersji zależny od liczebności próby, odczytywany z tablic.

3.2. Metody weryfikacji pojedynczego wyniku wątpliwego

3.2.1. Metoda Dixona

Istotą metody Dixona [21] jest test wartości skrajnych Dixona, określany czasem jako test *Q*-Dixona [3, 4, 9] lub test *C*-Dixona [5, 22], określający maksymalny stosunek różnic, których można oczekiwać między wynikami o skrajnym zaszeregowaniu na różnych poziomach prawdopodobieństwa P i dla różnych liczebności próby n.

W literaturze spotkać można nieco odmienne warianty testu Dixona, różniące się zalecanymi stosunkami testowymi w zależności od liczebności próby.

W wersji źródłowej [21], zalecanej do prób o liczebności $3 \leq n \leq 20$, testem weryfikującym zaliczenie (lub nie) danego wątpliwego wyniku do próby jest zmienna losowa Q, zależna od liczebności n próby, określana:

– dla $n \leq 8$

$$Q = \frac{x_2 - x_1}{x_n - x_1} \tag{3.4a}$$

– dla $8 \leq n \leq 15$ $n \leq 8$

$$Q = \frac{x_3 - x_1}{x_{n-1} - x_1} \tag{3.4b}$$

– dla $n \geq 8$

$$Q = \frac{x_3 - x_1}{x_{n-2} - x_1} \tag{3.4c}$$

Wyliczoną wartość statystyki testowej Q na podstawie próby, przy użyciu odpowiedniego wzoru (3.4a, b lub c) porównuje się z wartością krytyczną $Q_{n,\alpha}$ danego poziomu istotności α oraz liczebności próby n. Ta wartość krytyczna wyznacza prawostronny obszar krytyczny testu i dlatego:

– Jeżeli $Q \geq Q_{n,\alpha}$, to wątpliwy wynik należy odrzucić.
– Jeżeli $Q < Q_{n,\alpha}$, to nie ma podstaw do odrzucenia wątpliwego wyniku.

Najczęściej wykorzystywany w praktyce poziom istotności testowania α wynosi 0,01; 0,05; lub 0,10. Wartości krytyczne testu Q-Dixona są zestawione np. w *Statystyce stosowanej dla inżynierów* [21].

Nieco inny wariant testu C-Dixona prezentowany jest w poradniku wydanym przez amerykańska Environmental Protection Agency [5] i opisany m.in. w pracy Danuty Wiechuły, Krzysztofa Laski i Jerzego Kwapulińskiego [22]. Statystyka testowa C, zależna od liczebności prób, jest określana:

– dla $3 \leq n \leq 7$,

$$\text{jako: } C = \frac{x_n - x_{n-1}}{x_n - x_1} \tag{3.5a}$$

– dla $8 \leq n \leq 10$,

$$\text{jako: } C = \frac{x_n - x_{n-1}}{x_n - x_2} \tag{3.5b}$$

– dla $11 \leq n \leq 13$,

$$\text{jako: } C = \frac{x_n - x_{n-2}}{x_n - x_2} \tag{3.5c}$$

– dla $14 \leq n \leq 25$,

$$\text{jako: } C = \frac{x_n - x_{n-2}}{x_n - x_3} \tag{3.5d}$$

Wartości krytyczne testu C-Dixona zestawiono w *Practical methods for Data Analysis* [5].

3.2.2. Reguła trzysigmowa

Regułę tę można stosować w odniesieniu do prób wyników pomiarów w wypadku dużych liczebności prób – $n > 30$ (tzw. próby duże) [9, 15, 18, 21].

Zgodnie z tym kryterium prawdopodobieństwo, że jakiś wynik pomiaru znajduje się poza przedziałem $(\bar{x} - 3S, \bar{x} + 3S)$, wynosi 0,0027 ($\bar{x}$ – średnia z próby, S – odchylenie standardowe z próby). Wobec wyników pomiarów po odrzuceniu ich wątpliwej wartości wyznacza się wartość średnią x_{odrz} oraz odchylenie standardowe (S_{odrz}) i na ich podstawie określa się przedział trzysigmowy $\bar{x}_{odrz} \pm S_{odrz}$. Jeżeli wątpliwy wynik pomiaru x_w znajduje się poza tym przedziałem, będzie to wstępna wskazówka, że nie należy on do danej próby. Informację tę można potwierdzić za pomocą przedziału ufności średniej.

3.2.3. Przedział ufności średniej

Przedział ten w sposób bardziej ścisły określa, czy wątpliwy wynik pomiaru należy uwzględnić, czy też odrzucić z dalszej analizy. Rozmiary tego przedziału zależą od przyjętego poziomu istotności α, określającego prawdopodobieństwo niezaliczenia danego wyniku wątpliwego do próby i wiążącego się z poziomem (współczynnikiem) ufności $(1 - \alpha)$, dla którego konstruowany jest przedział ufności średniej.

Z rozważań o estymacji przedziałowej średniej [15, 21] wynika, że w praktyce, przy niezbyt rygorystycznym podejściu, przedział ufności dla danej dużej próby ($n > 30$) z wyłączeniem pomiaru wątpliwego można określić ze wzoru:

$$\bar{x}_{odrz} - u_\alpha \frac{S_{odrz}}{\sqrt{n}} < \mu < \bar{x}_{odrz} + u_\alpha \frac{S_{odrz}}{\sqrt{n}} \tag{3.6}$$

gdzie u_α – wartość krytyczna odczytana z tablic dystrybuanty rozkładu $N(0; 1)$ dla poziomu istotności α.

W praktyce metrologicznej wykonuje się z reguły mniejszą niż 30 liczbę pomiarów w serii. Decyduje o tym, niekiedy, nadmierny koszt lub pracochłonność badań bądź też wydłużony czas ich trwania, co z reguły powoduje zmianę warunków pomiaru.

Przy opracowywaniu małej liczby wyników pomiarów – tzw. małych prób o liczebnościach $10 < n \le 30$ – korzysta się, jak ogólnie wiadomo, z rozkładu t-Studenta, w którym jednym z parametrów jest liczba stopni swobody $f = n - 1$. Wówczas przedział ufności wobec poziomu istotności α konstruuje się wg zależności:

$$\bar{x}_{odrz} - t_{\alpha,n-1} \frac{S_{odrz}}{\sqrt{n}} < \mu < \bar{x}_{odrz} + t_{\alpha,n-1} \frac{S_{odrz}}{\sqrt{n}} \tag{3.7}$$

gdzie $t_{\alpha,n-1}$ – wartość krytyczna rozkładu t-Studenta przy poziomie istotności testowania α i liczbie stopni swobody $(n - 1)$.

Jeżeli wątpliwy wynik pomiaru x_w lokuje się poza określonym wg relacji (3.6) lub (3.7) przedziałem ufności, należy go odrzucić.

Jak już wspomniano wcześniej, w wypadku prób statystycznych bardzo małych ($n \le 10$), odchylenie standardowe od próby określane powinno być wg relacji (3.3) i na podstawie rozstępu R.

3.2.4. Test Grafa

Zalecany jest w badaniach analitycznych dotyczących jakości wód [3, 16] w wypadku długich serii pomiarowych ($n > 25$).

Wartość wątpliwa x_w jest odrzucana jako obarczona grubym błędem, jeżeli znajduje się poza przedziałem

$$\bar{x}_{odrz} \pm 4 S_{odrz},$$

gdzie:
$\overline{x}_{odrz}$ – średnia arytmetyczna z próby po wyłączeniu wyniku wątpliwego x_w,
S_{odrz} – odchylenie standardowe po wyłączeniu wyniku wątpliwego.

3.2.5. Kryterium Chauveneta

Jego istota polega na ocenie prawdopodobieństwa uzyskania wątpliwego wyniku pomiaru x_w przy założeniu, że wszystkie wyniki pomiarów są prawidłowe i podlegają rozkładowi normalnemu [17].

Po obliczeniu średniej $\overline{x}$ i odchylenia standardowego z próby S wg wzorów (3.1) i (3.2) wylicza się wartość, czyli liczbę określającą, o ile odchyleń standardowych x_w różni się od $\overline{x}$ (jest to inaczej unormowana lub standaryzowana wartość x_w)

$$t_w = \frac{|x_w - \overline{x}|}{S} \tag{3.8}$$

Następnie odczytuje się z tabeli zawartej we *Wstępie do analizy błędu pomiarowego* J.R. Taylora [17] prawdopodobieństwo $P\,(t_w S)$ tego, że właściwy wynik pomiaru będzie się różnił od $\overline{x}$ o t_w lub więcej odchyleń standardowych. Po pomnożeniu liczby wykonanych pomiarów n przez $P\,(t_w S)$ otrzymuje się wartość n_{gor} określającą spodziewaną liczbę pomiarów dających wyniki co najmniej tak złe, jak x_w (lub gorsze):

$$n \cdot P\,(t_w S) = n_{gor} \tag{3.9}$$

Jeśli n_{gor} jest mniejsze niż 0,5, to wynik x_w nie spełnia kryterium Chauveneta i powinien zostać odrzucony.

3.2.6. Metoda statystyki *U* wg normy PN-87/N-01052/13

Testowanie wyniku wątpliwego x_1 lub x_n polega w tym wypadku na obliczeniu wartości statystyki U_1 lub U_n według wzoru [10]

$$U_1 = \frac{\overline{x} - x_1}{S} \tag{3.10a}$$

lub

$$U_n = \frac{x_n - \overline{x}}{S} \tag{3.10b}$$

W zależności od przyjętego poziomu istotności α i liczebności próby n odczytuje się z właściwej tablicy [10] wartość krytyczną h, którą porównuje się z obliczonymi statystykami U_1 i U_n. Jeśli $U_1 > h$ lub $U_n > h$, należy uznać, że wynik wątpliwy x_1 (lub x_n) jest obarczony błędem grubym i wyeliminować go z dalszych obliczeń.

Jeśli $U_1 \leq h$ lub $U_n \leq h$, należy uznać, że wynik wątpliwy należy do badanej populacji.

3.2.7. Test Grubbsa w wypadku jednej wartości odstającej

Jego wykorzystanie polega na obliczeniu statystyki testowej Grubbsa G dla zaobserwowanej wartości odstającej największej lub najmniejszej w serii [14] – odpowiednio x_n lub x_1 – według wzorów:

$$G_n = \frac{x_n - \overline{x}}{S} \tag{3.11a}$$

lub

$$G_n = \frac{\overline{x} - x_1}{S} \tag{3.11b}$$

Obliczone wartości statystyki G porównuje się z krytycznymi górnymi wartościami testu Grubbsa $G_{\alpha,n}$, odpowiadającymi liczebności próby n oraz poziomu istotności $\alpha = 0{,}05$ (5%) i $\alpha = 0{,}01$ (1%), zestawionymi w odnośnej tablicy w normie [14].

Jeśli G_1 (lub G_n) $\leq G_{0,05;n}$, to badany wynik – odpowiednio x_1 (lub x_n) uznaje się za poprawny.

Jeśli $G_{0,05;n} < G_1$ (lub G_n) $\leq G_{0,01;n}$, to badany wynik uznaje się za niepewny (i oznacza jedną gwiazdką).

Jeśli G_1 (lub G_n) $> G_{0,01;n}$, to badany wynik uznaje się za wartość odstającą (i oznacza dwiema gwiazdkami).

Nieco inaczej prezentowany jest test Grubbsa w celu eliminacji wyników pomiarów obarczonych błędami grubymi w pracy *Podstawy badań inżynierskich* [7].

3.2.8. Inne sposoby weryfikacji wyników wątpliwych

Jednym z ciekawszych sposobów wykorzystywanych w badaniach analitycznych, w szczególności w odniesieniu do wód podziemnych w celu wykrycia i eliminacji błędów grubych jest procedura kart kontrolnych [16, 19].

Ta metoda statystycznej kontroli jakości należy do najbardziej efektywnych, a jej istota polega na rejestracji wyników pomiarów badanych parametrów (zwykle stężeń analizowanego składnika). Wyniki pomiarów zestawia się w postaci punktów na odpowiednich prostokątnych wykresach, nazywanych kartami kontrolnymi. Zasadnicze linie kontrolne ustalane są zwykle na takim poziomie ograniczeń wartości parametru, że prawdopodobieństwo ich przekroczenia przy poprawnych pomiarach wynosi $P = 0{,}99$ (granice kontrolne 3σ). Często wykreślane są również tzw. pomocnicze lub wewnętrzne linie kontrolne, odpowiadające granicom 2σ i prawdopodobieństwu $P = 0{,}95$.

Innym ciekawym podejściem do odstających wyników pomiarów jest metoda *robust statistics* [1] – metoda statystyk *robust* [16]. Polega ona nie na odrzucaniu odstających obserwacji, lecz na ich przystosowaniu poprzez oszacowanie drogą iteracyjną, tzw. *robust* ($\overline{x}_r$ i σ_r), wartości średniej i odchylenia standardowego. Powyższa metoda wykorzystuje tzw. elastyczne postępowanie statystyczne bez konieczności kontrowersyjnego odrzucania wyników pomiarów obarczonych błędami grubymi. W szczególności cechuje się tym, że nie jest zbyt czuła na odstępstwa od założeń, na których się opiera [6].

3.3. Metody weryfikacji zespołu wyników wątpliwych

3.3.1. Metoda statystyki U^* wg normy PN-87/N-01052/13

Testowanie zespołu (dwóch lub więcej) wyników wątpliwych w serii pomiarowej, w tym w szczególności wyników skrajnych x_1 i x_n, polega na krokowym obliczaniu wartości statystyki U_n^* wg wzoru:

$$U_n^* = \max_{1 \le k < n} \left| \frac{x_k - \bar{x}}{S} \right| \tag{3.12}$$

gdzie x_k – k-ty wynik pomiaru w serii ($k = 1, 2, ..., n$).

Zalecane poziomy istotności testowania α^* w tym wypadku stanowią podwojoną wartość poziomów istotności α stosowanych w odniesieniu do pojedynczych wartości wątpliwych, tj. $\alpha^* = 0{,}20$ lub 0,10.

W zależności od przyjętego poziomu istotności α^* i liczebności próby n odczytuje się z właściwej tablicy [10] wartość krytyczną h, którą porównuje się z obliczaną statystyką U_n^*:

- Jeśli $U_n^* \le h$, należy uznać, że próba nie zawiera wyników obarczonych błędami grubymi.
- Jeśli $U_n^* > h$, należy uznać, że wynik x_k odpowiadający wartości statystyki U_n^* obarczony jest błędem grubym. Po jego wyeliminowaniu z próby należy powtórzyć wszystkie czynności w odniesieniu do kolejnego wątpliwego wyniku pomiaru dla próby o liczebności zmniejszonej z n do $(n - 1)$.

3.3.2. Test Grubbsa w odniesieniu do dwóch wartości odstających

W celu zbadania, czy dwie największe wartości zaobserwowane mogą być wartościami odstającymi, należy obliczyć statystykę testową Grubbsa G według wzoru [14]:

$$G = \frac{s_{n-1,n}^2}{s_o^2} \tag{3.13}$$

gdzie:

$$s_o^2 = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \tag{3.13a}$$

$$s_{n-1,n}^2 = \sum_{i=1}^{n-2} \left(x_i - \bar{x}_{n-1,n} \right)^2 \tag{3.13b}$$

$$\bar{x}_{n-1,n} = \frac{1}{n-2} \sum_{i=1}^{n-2} x_i \tag{3.13c}$$

W odniesieniu do dwóch najmniejszych wartości zaobserwowanych należy obliczyć statystykę testową Grubbsa G według wzoru:

$$G = \frac{s_{1,2}^2}{s_o^2} \tag{3.14}$$

gdzie:

$$s_{1,2}^2 = \sum_{i=3}^{n} \left(x_i - \bar{x}_{1,2} \right)^2 \tag{3.14a}$$

$$\bar{x}_{1,2} = \frac{1}{n-2} \sum_{i=3}^{n} x_i \tag{3.14b}$$

Obliczone wartości statystyki G porównuje się z krytycznymi dolnymi wartościami testu Grubbsa $G_{\alpha,n}$, odpowiadającymi liczebności próby n oraz poziomowi istotności $\alpha = 0{,}05$ (5%) i $\alpha = 0{,}01$ (1%), zestawionymi w tablicy w normie [14]:

- Jeśli $G \geq G_{0,05;n}$, to badane wartości traktowane są jako poprawne.
- Jeśli $G_{0,05;n} > G \geq G_{0,01;n}$, to badane wartości określane są jako niepewne (i oznaczane jedną gwiazdką).
- Jeśli $G < G_{0,01;n}$, to badane wyniki nazywane są wartościami odstającymi (i oznaczane dwiema gwiazdkami).

3.3.3. Test Rosnera

Stosowany jest wobec prób o liczebnościach $n \geq 25$ obserwacji, w których notuje się do 10 wartości odbiegających [5, 22]. Zakłada się, że analizowany zbiór ma rozkład normalny. W trakcie weryfikacji i rozkładu odbiegającego od rozkładu normalnego należy go odpowiednio przekształcić lub proces ustalania odbiegających wyników prowadzić, wykorzystując inny test.

Stosowanie testu wymaga wskazania maksymalnej liczby wartości odbiegających r.

Kryterium sprawdzającym odbieganie wartości parametru spośród danych analizowanego zbioru stanowi wyrażenie:

$$R_r = \frac{\left| y^{(r-1)} - \bar{x}^{(r-1)} \right|}{s^{(r-1)}} \tag{3.15}$$

gdzie:

$$\bar{x}^{(i)} = \frac{1}{(n-1)} \sum_{j=1}^{n-i} x_j \; ;$$

$$s^{(i)} = \left[\frac{1}{(n-1)} \sum_{j=1}^{n-i} (x_j - \bar{x}^{(i)})^2 \right]^{1/2} \; ;$$

$y^{(i)}$ – najbardziej odległa wartość z pomiaru od $\bar{x}^{(i)}$.

Technika obliczeń przy stosowaniu testu Rosnera jest następująca. Obliczone charakterystyki statystyczne (parametry) dla całego zbioru n wartości, oznaczone jako $\overline{x}^{(0)}$ oraz $s^{(0)}$ porównujemy kolejno z poszczególnymi wynikami, ustalając wartość najbardziej od niej odbiegającą $y^{(0)}$. Obliczamy wartość kryterium porównawczego (R_r), którą w wypadku odbiegającego parametru porównujemy z krytyczną wartością tablicową (λ_r) odczytaną dla liczebności zbioru n, ilości wartości odbiegających r oraz ustalonego poziomu istotności α [5, 22].

Jeśli $R_r \geq (\lambda_r)$, badana wartość x_r jest wartością odbiegającą w zbiorze danych. Można ją uznać za wynik wątpliwy i odrzucić z dalszych analiz.

Jeśli wynik porównania jest odwrotny, wówczas prowadzone są obliczenia dla kolejnej wartości odbiegającej i powtarzana jest procedura testowania.

Odrzucamy tę odbiegającą wartość $y^{(0)}$ z analizowanego zbioru i prowadzimy dalsze obliczenia, ustalając wartość średnią $\overline{x}^{(1)}$ oraz $s^{(1)}$ dla podzbioru złożonego z (n – 1) obserwacji. Ustalamy ponownie najbardziej odległą wartość próby od $\overline{x}^{(1)}$, oznaczając ją $y^{(1)}$. Obliczamy wartość R_r dla odbiegającego wyniku $y^{(1)}$ i porównujemy z krytyczną wartością tablicową λ_r dla (n–1; α).

W razie odrzucenia tej wartości procedura obliczeń, porównywania i odrzucania odbiegających wartości kontynuowana jest aż do chwili, gdy ze zbioru danych usunięte zostaną wszystkie odbiegające wartości.

3.3.4. Test Walsha

Test Walsha stosowany jest do dużych prób o liczebności przewyższającej kilkadziesiąt obserwacji i nie wymaga spełnienia warunku normalności rozkładu analizowanych obserwacji w próbie [5, 22].

Wobec prób o liczebności $n > 220$ przyjmuje się poziom istotności testowania $\alpha = 0{,}05$, natomiast dla prób o mniejszych liczebnościach, z przedziału $220 > n > 60$, analizę prowadzi się odnośnie do poziomu istotności $\alpha = 0{,}10$. Do zbiorów o mniejszych liczebnościach ($n < 60$) nie stosuje się tego testu.

Wobec badanego uporządkowanego zbioru danych określamy ilość r wartości wątpliwych, prawdopodobnie istotnie odbiegających od wartości wyników pomiarów w zbiorze. W tym teście r – w skrajnym wypadku – może być równe 1.

Dla przyjętego poziomu istotności α wartości pomiarów są uznane za odbiegające i odrzucane, jeśli spełniony jest warunek:

- w odniesieniu do r wartości największych:

$$x_{(n-1-r)} - (1+a)x_{(n-r)} + ax_{(n-1-k)} \ > \ 0 \tag{3.16}$$

- w odniesieniu do r wartości najmniejszych:

$$x_{(r)} - (1+a)x_{(r-1)} + ax_{(k)} \ < \ 0 \tag{3.17a}$$

gdzie:

$$c = (2n)^{\frac{1}{2}}$$ (obliczona wartość ulega zaokrągleniu w górę do wartości całkowitej);

$$k = r + c; \tag{3.17b}$$

$$b^2 = 1/\alpha;$$

$$a = \frac{1 + b\sqrt{(c - b^2)/(c-1)}}{c - b^2 - 1}.$$

W razie równoczesnego spełnienia obydwu warunków, odrzuca się wartości odbiegające zarówno w górę, jak i w dół.

Wartości odbiegające najczęściej są wynikami pomiarów nieporównywalnie małych bądź nieporównywalnie dużych w stosunku do reszty danych. Wartości te najczęściej są wynikiem pomyłek manualnych.

4. UWAGI KOŃCOWE

Problem traktowania źródłowych wyników pomiarów, które znacznie odbiegają od innych wyników i jako wątpliwe mogą być uznane za odstające, w szczególności obarczone błędem grubym, jest trudny do jednoznacznego rozstrzygnięcia. Największe kontrowersje wzbudza kwestia odrzucania danych. Jak zauważa J.R. Taylor [17]: „wielu naukowców ma poczucie, że odrzucenie danych pomiarowych nigdy nie jest usprawiedliwione, chyba że istnieje zewnętrzny dowód na niepoprawność kwestionowanych wyników. Bardziej umiarkowane stanowisko głosi, że kryterium Chauveneta powinno być wykorzystywane do zidentyfikowania danych, które należy co najmniej wziąć pod uwagę jako kandydatów do odrzucenia".

Nie ma wątpliwości, że podobnie należy traktować także inne metody weryfikacji wątpliwych wyników pomiarów. Generalnie słuszne jest zalecenie prezentowane m.in. w polskich normach [13, 14], że dane obarczone oczywistymi błędami powinny być przeanalizowane i albo poprawione, albo odrzucone.

Znaczne kłopoty w praktyce mogą wiązać się z wyborem właściwej metody weryfikacji wątpliwych wyników pomiarów, adekwatnej do konkretnej sytuacji, tym bardziej że możliwości wyboru, a szczególnie w wypadku pojedynczych wartości wątpliwych są znaczne, a wyniki weryfikacji z wykorzystaniem różnych metod często w mniejszym lub większym stopniu różnią się [9, 21, 22]. Należy sądzić, że najbardziej uzasadnione jest korzystanie z metod weryfikacji wykorzystujących teorię testowania hipotez statystycznych, kiedy można kontrolować ryzyko popełnienia błędu 1. rodzaju (zakładamy *á priori* poziom istotności testowania α) w trakcie wnioskowania statystycznego.

Zagadnienie zasługuje na bliższą analizę, także z ewentualnym uwzględnieniem mocy poszczególnych testów i ryzyka błędów 2. rodzaju.

Podczas weryfikacji zespołów (wielu) wyników wątpliwych w analizowanych seriach pomiarowych, kluczowym zagadnieniem do wstępnego rozstrzygnięcia jest ewentualna niejednorodność statystyczna, która może się wiązać m.in. z błędami systematycznymi w odniesieniu do pomiarów prowadzonych w warunkach odtwarzalności.

LITERATURA

[1] Analytical Methods Committee: *Robust statistics – How not to Reject Outliers.* Part 2, Inter – laboratory trials. „Analyst", vol. 114, Dec. 1989.

[2] *Podstawy metrologii technicznej. Laboratorium,* red. J. Bednarczyk, Kraków 2000 (Skrypty Uczelniane AGH, nr 1591).

[3] Dancer K.: *Significance of statistics in Quality Assurance*. [W:] *Accreditation and quality assurance in analytical chemistry.* Ed. H. Günzler, Springer-Verlag, Berlin--Heidelberg 1996.

[4] Gibbons R.D.: *Statistical methods for groundwater monitoring.* John Wiley & Sons, New York 1994.

[5] *Guidance for Data Quality Assessment*. *Practical methods for Data Analysis*. U.S. Enviromental Protection Agency, Washington 1998.

[6] Kendall G.M., Auckland R.W.: *Słownik terminów statystycznych.* PWE, Warszawa 1986.

[7] Kukiełka L.: *Podstawy badań inżynierskich.* Wyd. Nauk. PWN, Warszawa 2002.

[8] *Ocena metanonośności węgli kamiennych Górnośląskiego Zagłębia Węglowego na podstawie wyników pomiarów otworowych.* Red. K. Twardowski, Wyd. Centrum PPGSMiE PAN, Kraków 1997.

[9] Ostachowicz J.: *Technika opracowywania danych pomiarowych w ćwiczeniach laboratoryjnych z fizyki.* Wyd. Ośrod. Eduk. Niestac. AGH, Kraków 1998.

[10] PN-87/N-01052/13: *Badania statystyczne – Zasady wykrywania w próbce wyników obarczonych błędami grubymi*. Polski Komitet Normalizacji, Miar i Jakości, Warszawa 1987.

[11] PN-ISO 3301: *Statystyczna interpretacja danych. Porównywanie dwóch wartości średnich w przypadku obserwacji parami*. Polski Komitet Normalizacyjny, Warszawa 1994.

[12] PN-ISO 3534-1: *Statystyka – Terminologia i symbole*, Sekcja 3: *Ogólne terminy dotyczące obserwacji i wyników badań*. Polski Komitet Normalizacyjny, Warszawa 2002.

[13] PN-ISO 5725-1: *Dokładność (poprawność i precyzja) metod pomiarów i wyników pomiarów*, Cz. 1, *Ogólne zasady i definicje.* Polski Komitet Normalizacyjny, Warszawa 2002.

[14] PN-ISO 5725-2: *Dokładność (poprawność i precyzja) metod pomiarów i wyników pomiarów*, Cz. 2, *Podstawowa metoda określania powtarzalności i odtwarzalności standardowej metody pomiarowej.* Polski Komitet Normalizacyjny, Warszawa 2002.

[15] Skwarczyński A.: *Statystyczne metody oceny wyników pomiarów*, Cz. 1, *Zmienna losowa jednowymiarowa*. Kraków 1980 (Skrypty Uczelniane AGH, nr 724).

[16] Szczepańska J., Kmiecik E.: *Statystyczna kontrola jakości danych w monitoringu wód podziemnych.* Wyd. AGH, Kraków 1998.

[17] Taylor J.R.: *Wstęp do analizy błędu pomiarowego.* Wyd. Nauk. PWN, Warszawa 1995.

[18] *Teoria pomiarów.* Red. H. Szydłowski, PWN, Warszawa 1974.

[19] Thompson J.R., Koronacki J.: *Statystyczne sterowanie procesem. Metoda Deminga etapowej optymalizacji jakości.* Wyd. AOW PLJ, Warszawa 1994.

[20] Twardowski K., Traple J., Rychlicki S.: *Próba oceny metody porometrii helowej do pomiarów laboratoryjnych porowatości skał.* „Wiertnictwo-Nafta-Gaz" 2005, t. 21, z. 2.

[21] Volk W.: *Statystyka stosowana dla inżynierów.* Wyd. Nauk.-Techn., Warszawa 1973.

[22] Wiechuła D., Loska K., Kwapuliński J.: *Zastosowanie procedur statystycznych w interpretacji wyników badań środowiskowych*. Gliwice 2005, (Zeszyty Naukowe Politechniki Śląskiej, Górnictwo, z. 267).

[23] *Wyrażanie niepewności pomiaru. Przewodnik*. Główny Urząd Miar, Warszawa 1999.

[24] Zieliński R.: *Tablice statystyczne*. PWN, Warszawa 1972.