LABORATORIUM METROLOGII

ĆWICZENIE 4

Statystyczna obróbka wyników pomiarowych

v. 1 / 11.2008

1. Cel ćwiczenia

Ćwiczenie ma na celu zapoznanie się z teoretycznymi i praktycznymi podstawami statystycznej obróbki wyników pomiarowych. W części doświadczalnej zwrócono uwagę na opanowanie umiejętności postępowania w przypadku badania zbiorów obiektów i wyznaczania ich statystycznych parametrów opisowych, takich jak wartość średnia, wartość minimalna i maksymalna, wartość modalna, mediana, wariancja i odchylenie standardowe. Ćwiczenia ma również na celu zaznajomienie z metodologią tworzenia podstawowych diagramów charakteryzujących rozrzut wyników pomiarowych, takich jak histogram i krzywa skumulowana.

2. Wprowadzenie

Dość częstym zjawiskiem w praktyce pomiarowej jest rozrzut wyników, który może mieć różne przyczyny i różną skalę. Ujawnianie tego rozrzutu nie zawsze jest potrzebne, ale również nie zawsze jest możliwe. Zależnie od rozdzielczości pomiaru oraz od skali rozrzutu w serii wyników, rozrzut może ale nie musi się ujawnić. Ujawnienie rozrzutu nie jest możliwe gdy nie dysponuje się układem o wystarczająco dużej rozdzielczości pomiarowej, a mierzona wartość zmienia się nieznacznie. Niemniej jednak wynik pomiaru ma zawsze wartość losową.

Istnieją dwie typowe sytuacje, w których uzyskuje się wyniki losowe pomiarów obarczone dyspersją:

- gdy wielokrotnie powtarza się pomiar tego samego parametru, dla jednego określonego obiektu, w nominalnie nie zmienionym podstawowym układzie warunków fizycznych, czyli przy nie zmieniających się znacząco wielkościach wpływowych (np.: temperatura, ciśnienie, wilgotność, itp.). W tej sytuacji dla zauważenia rozrzutu wymagana jest odpowiednio duża rozdzielczość pomiarów. Rozrzut wyników jest spowodowany przez szereg różnych wielkości wpływowych, które są kontrolowane w ograniczonym stopniu lub wcale nie są brane pod uwagę. Wszystkie wielkości wpływowe obciążają wyniki pomiarów błędami przypadkowymi (o nieznanej wartości i znaku).
- gdy dokonuje się pomiarów tego samego parametru dla serii obiektów tworzących klasę (np.: gdy mierzymy pojemność takich samych kondensatorów, czyli o tej samej pojemności znamionowej, tolerancji oraz technologii wykonania). Nawet gdy pomiary te wykonywane są niekoniecznie bardzo dokładnie, mierzone wartości mogą znacznie się różnić gdyż dotyczą różnych obiektów. W tej sytuacji błędy pomiaru nie mają istotnego znaczenia gdy są znacznie mniejsze od różnic parametru mierzonego dla poszczególnych obiektów.

Wyniki pomiarów wykonanych w seriach o dużej liczebności nie umożliwiają łatwego wyciągania wniosków na temat całej populacji, którą reprezentują serie pomiarowe. Dlatego dąży się do określenia minimalnej liczebności serii, która będzie reprezentatywna, czyli której parametry będą takie same jak całej populacji. Z drugiej strony, dąży się do obliczenia na podstawie serii takich parametrów, które będą najlepiej charakteryzować całą populację. Narzędzi do takiej kompresji wyników pomiarów dostarcza statystyka matematyczna, a zagadnienie poszukiwania parametrów charakteryzujących całą populację (zbiór pełny) na podstawie serii, nazywane jest estymacją.

Do najczęściej obliczanych statystyk z serii należą:

- **Wartość średnia**, obliczana jako średnia arytmetyczna jest estymatorem zgodnym, nieobciążonym i najefektywniejszym wartości oczekiwanej:

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

  gdzie: n – liczba wszystkich wyników pomiarów w serii, i – kolejny numer wyniku pomiaru.
- **Wartość modalna** (moda, dominanta), która jest wartością najczęściej powtarzającego się w serii wyniku pomiaru. Jeśli niektóre wyniki w serii powtarzają się równie często mamy do czynienia z rozkładem wielomodalnym zmiennej losowej jaką jest wynik pomiaru. Wartość modalna jest oznaczana przez Mo(x).
- **Mediana** (wartość środkowa) jest środkową wartością uporządkowanych rosnąco (szereg rozdzielczy) wyników w serii. Gdy liczebność serii jest wyrażona liczbą nieparzystą medianę można określić bezpośrednio, natomiast dla serii o parzystej liczbie elementów medianę wylicza się jako wartość średnią z dwóch elementów środkowych:

$$\mathrm{Me}(x) = x_{(n+1)/2} \ \text{ dla } n \text{ nieparzystych,}$$

$$\mathrm{Me}(x) = \frac{1}{2}\left(x_{n/2} + x_{(n/2)+1}\right) \text{ dla } n \text{ parzystych.}$$

- **Statystyki pozycyjne**, które są określane jako minimalna i maksymalna wartość wyników w danej serii, czyli występują jako pierwszy i ostatni element szeregu rozdzielczego. Statystyki te są oznaczane jako $x_{\min}$ i $x_{\max}$.
- **Wariancja empiryczna** jest obliczana dla serii długich (n > 30) jako suma kwadratów odchyleń poszczególnych wyników w serii od wartości średniej, podzielona przez liczbę wyników pomiarowych:

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

  Dla serii krótkich (n < 30) oblicza się wariancję empiryczną skorygowaną:

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

- Odchylenie standardowe (średniokwadratowe), które podobnie jak wariancja jest miarą rozproszenia wyników pomiarów w serii, oblicza się jako pierwiastek kwadratowy z wariancji empirycznej nieskorygowanej:

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

Przedstawione powyżej statystyki są parametrami, które syntetycznie charakteryzują serie pomiarowe i ułatwiają porównanie serii o różnej liczebności co pozwala na wnioskowanie o ich reprezentatywności.

Jak już wspomniano metodami wyznaczania parametrów populacji za pomocą parametrów próby (serii) zajmuje się teoria estymacji. W ramach tej teorii opracowana jest:

- **estymacja punktowa**, która polega na określeniu na podstawie wyników z serii pomiarowej jednej wartości (estymatora punktowego), która jest oszacowaniem odpowiedniego parametru populacji. Przykładowo, wartość średnia $\bar{x}$ wyliczona na podstawie serii wyników jest estymatorem punktowym wartości oczekiwanej całej populacji.
- **estymacja przedziałowa**, która ma na celu określenie przedziału, w którym z określonym prawdopodobieństwem (poziomem ufności) znajduje się badany parametr populacji. Jako środek wspomnianego przedziału przyjmuje się wartość średnią $\bar{x}$, a jako granice pewną krotność odchylenia standardowego, zależną od poziomu ufności. W ten sposób możemy określić niepewność $\pm\Delta_q$ dla serii długich wyników o rozkładzie normalnym:

$\Delta_q = 3\sigma \cong 3s$ dla poziomu ufności $1 - q = 0{,}9973$,

$\Delta_q = 2\sigma \cong 2s$ dla poziomu ufności $1 - q = 0{,}9544$,

$\Delta_q = \sigma \cong s$ dla poziomu ufności $1 - q = 0{,}6826$,

gdzie: Δ_q - połowa szerokości przedziału niepewności, a $\bar{x} \pm \Delta_q$ jest estymatorem przedziałowym wartości oczekiwanej, na poziomie ufności $1 - q$.

Serię wyników pomiarów obarczonych rozrzutem można przedstawić również graficznie w formie histogramu lub krzywej skumulowanej. W tym celu należy:

- uporządkować wyniki pomiarów dla danej serii według rosnących wartości, tworząc w ten sposób tzw. szereg rozdzielczy:

$$x_{\min} < x_2 < x_3 < \ldots < \mathrm{Me}(x) < \ldots < x_{n-1} < x_{\max}$$

- podzielić cały otrzymany przedział $x_{min} ... x_{max}$ (gdzie: $x_{min} = x_1$, $x_{max} = x_n$) na k podprzedziałów o równej szerokości $\Delta_i x$ (gdzie: i = 1, 2, …,k):

$$\Delta_i x = \frac{x_{max} - x_{min}}{k} = \frac{x_n - x_1}{k},$$

  tak aby w każdym przedziale $\Delta_i x$ znajdowało się co najmniej kilka wyników z serii. Dla serii długich (n > 30) liczbę przedziałów k można wyznaczyć w przybliżeniu na podstawie empirycznego wzoru Sturgesa:

$$k \cong 1 + 3{,}3 \lg(n).$$

- określić wysokość słupka histogramu w każdym przedziale $\Delta_1 x$, $\Delta_2 x$, …, $\Delta_k x$, która jest równa ilości wyników pomiarowych m_i o wartościach z danego przedziału lub częstości w_i występowania wyniku w tym przedziale:

$$w_i = \frac{m_i}{n}$$

- określić wartości rzędnych krzywej skumulowanej jako częstości skumulowane ν_j, wyznaczane dla kolejnych przedziałów $\Delta_1 x$, $\Delta_2 x$, …, $\Delta_k x$, jako sumy wcześniej obliczonych częstości w_i we wszystkich przedziałach znajdujących się na lewo od przedziału dla którego jest obliczana częstość skumulowana:

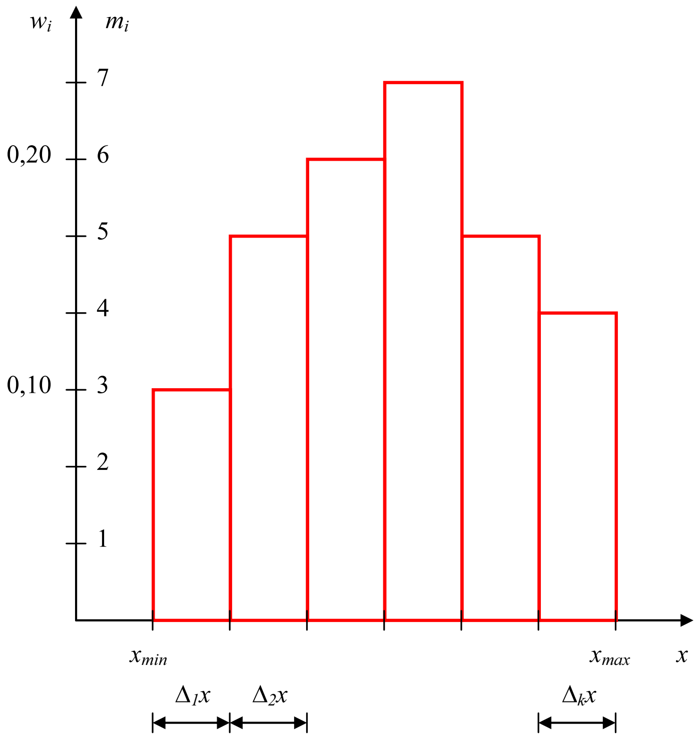
$$\nu_j = \sum_{i=1}^{j} w_i ,$$

  gdzie: i = 1, 2, …, j dla j = 1, 2, …, k.

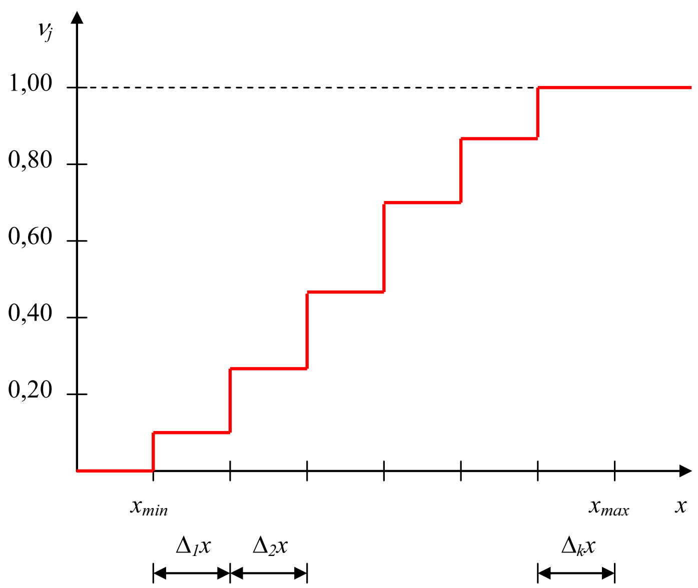
Histogram empirycznego rozkładu prawdopodobieństwa jest oszacowaniem funkcji gęstości prawdopodobieństwa. Podczas konstruowania tego diagramu przyporządkowuje się każdemu przedziałowi $\Delta_i x$ słupek o wysokości proporcjonalnej do częstości w_i. Przy zwiększaniu liczebności serii n i liczby przedziałów k histogram ten wygładza się, dążąc w granicy do funkcji gęstości prawdopodobieństwa.

Innym diagramem charakteryzującym rozrzut wyników pomiarów jest krzywa skumulowana, która jest wykresem częstości skumulowanych ν_j dla takich samych przedziałów $\Delta_i x$ jak w przypadku histogramu. W granicy dla $n \to \infty$ i $k \to \infty$ wykres częstości skumulowanej dąży do dystrybuanty.

Przykładowe diagramy charakteryzujące rozrzut 30-elementowej serii wyników pomiarów przedstawiono na rys. 1 i rys. 2.



Rys. 1. Przykładowy histogram empirycznego rozkładu prawdopodobieństwa dla serii 30-elementowej



Rys. 2. Przykładowa krzywa skumulowana dla serii 30-elementowej

3. Program ćwiczenia

1. Dla populacji obiektów wskazanych przez prowadzącego ćwiczenie, określić parametry znamionowe oraz parametr, który będzie mierzony seryjnie.
2. Określić jaki przyrząd pomiarowy i na jakim zakresie będzie odpowiedni do wykonania pomiarów seryjnych wcześniej wskazanych obiektów.
3. Kolejno wybrać losowo z całej populacji próby (serie): 50-, 30- i 10-elementową i dla każdej z serii wykonać pomiary wybranego parametru obiektów. Obiekty do każdej serii powinny być losowane z całej populacji i bez powtórzeń.
4. Dla każdej serii pomiarowej wyznaczyć podstawowe statystyki oraz dokonać estymacji punktowej i przedziałowej wartości oczekiwanej.
5. Porównać wyniki estymacji otrzymane dla poszczególnych serii i dokonać analizy reprezentatywności wybranych serii pomiarowych.
6. Przedstawić graficznie rozrzut wyników w otrzymanych seriach pomiarowych a następnie porównać wykreślone histogramy i krzywe skumulowane.
7. Określić podobieństwa kształtu otrzymanych histogramów do kształtu funkcji gęstości prawdopodobieństwa znanych rozkładów.

4. Pytania kontrolne

1. W jakich typowych sytuacjach rozrzut wyników pomiarów można opisywać stosując statystykę matematyczną?
2. W jaki sposób rozdzielczość pomiaru wpływa na ujawnianie się dyspersji wyników?
3. Podać definicje i wyjaśnić znaczenie najczęściej stosowanych statystyk z próby (serii).
4. W jaki sposób oblicza się medianę dla serii o parzystej liczbie elementów?
5. Na czym polega reprezentatywność serii?
6. Wymienić rodzaje estymacji i wyjaśnić na czym polegają.
7. Co to jest poziom ufności i jaki ma wpływ na szerokość przedziału niepewności dla wyników o rozkładzie normalnym?
8. Omówić metodologię tworzenia histogramu i krzywej skumulowanej.
9. Czym jest szereg rozdzielczy dla serii wyników pomiarów i jakimi wartościami jest ograniczony?
10. Jak wyznacza się częstość skumulowaną dla poszczególnych przedziałów zmiennej losowej oraz do jakiej wartości dąży zawsze krzywa skumulowana?
11. Jaki jest związek między histogramem a funkcją gęstości prawdopodobieństwa oraz między krzywą skumulowaną a dystrybuantą?

5. Literatura

1. Tylor J.R.: Wstęp do analizy błędu pomiarowego. Wydawnictwo Naukowe PWN, Warszawa 1995.
2. Turzeniecka D.: Ocena niepewności wyniku pomiarów. Wydawnictwo Politechniki Poznańskiej, Poznań 1997.
3. Skubis T.: Podstawy metrologicznej interpretacji wyników pomiarów. Wydawnictwo Politechniki Śląskiej, Gliwice 2004.