PODSTAWY STATYSTYKI

Dr hab. inż. Piotr Konieczka

piotr.konieczka@pg.gda.pl

1

Wprowadzenie

Wynik analityczny to efekt przeprowadzonego pomiaru(ów).

Pomiar to zatem narzędzie wykorzystywane w celu uzyskania wymaganego wyniku.

Otrzymanie miarodajnego wyniku jest efektem zastosowania odpowiednio wybranego narzędzia.

Np.: pomiar objętości titrantu można zrealizować za pomocą:

- wiadra ::))!!!
- butelki :)!
- kroplomierza :)?
- biurety z podziałką co 1 ml :) - zależy jaką objętość mamy odmierzyć - gdy np.: 1000 ml to OK.
- odpowiednio dobranej biurety!!!

2

Wprowadzenie

Pomiar bezpośredni - to co oznaczam jest tym co mierzę.

Pomiar pośredni - to co oznaczam jest obliczane (wyznaczane) w oparciu o parametry, które mierzę.

Wprowadzenie

Postępowanie analityczne - PROCEDURA ANALITYCZNA

ETAPY PROCEDURY ANALITYCZNEJ:

- pobieranie próbki,
- przygotowanie próbki,
- kalibracja,
- oznaczenie końcowe,
- .....

Statystyka matematyczna

Część matematyki oparta na wykorzystaniu rachunku prawdopodobieństwa oraz ukierunkowana na badanie prawidłowości pojawiania się określonych cech w obiektach materialnych lub zjawiskach, występujących masowo, tzn. mogących pojawiać się dowolną ilość razy.

Statystyka przedstawia te prawidłowości za pomocą liczb.

Statystyka matematyczna

Statystyka pozwala znaleźć odpowiedź na wiele pytań np.:

- Jak dokładny jest wynik oznaczenia?
- Jak wiele oznaczeń powinno być przeprowadzonych aby zwiększyć precyzję pomiaru?
- Czy badany produkt spełnia stawiane mu wymogi, normy?

Statystyka to narzędzie, którego używanie musi być prowadzone w sposób rozsądny i zrozumiały.

Statystyka matematyczna

Parametry statystyczne

Wielkości liczbowe służące do opisu struktury zbiorowości statystycznej w sposób systematyczny.

Wśród tych parametrów wyróżnić można cztery podstawowe grupy:

- miary położenia
- miary rozproszenia
- miary asymetrii
- miary skupienia

Statystyka matematyczna

Miary położenia

Miary położenia charakteryzują za pomocą jednej wartości (w sposób syntetyczny) ogólny poziom wartości cechy w zbiorowości.

Do najczęściej stosowanych miar położenia należą

- średnia arytmetyczna
- modalna
- kwantyle
- mediana
- decyle

Statystyka matematyczna

Średnia arytmetyczna

suma wartości cechy mierzalnej podzielona przez liczbę jednostek skończonej zbiorowości statystycznej:

$$x_{śr} = \frac{\sum_{i=1}^{n} x_i}{n}$$

9

Statystyka matematyczna

Wybrane właściwości średniej arytmetycznej:

- suma wartości cechy jest równa iloczynowi średniej arytmetycznej i liczebności zbiorowości;
- średnia arytmetyczna spełnia warunek:

$$x_{\min} < x_{śr} < x_{\max}$$

- (suma odchyleń poszczególnych wartości cechy od średniej równa się zero:

$$\sum_{i=1}^{n} (x_i - x_{śr}) = 0$$

- suma kwadratów odchyleń poszczególnych wartości cechy od średniej jest minimalna:

$$\sum_{i=1}^{n} (x_i - x_{śr})^2 = \min$$

- średnia arytmetyczna jest wrażliwa na skrajne wartości cechy,
- średnia arytmetyczna z próby jest dobrym przybliżeniem (oszacowaniem, estymatorem) wartości oczekiwanej.

10

Statystyka matematyczna

Modalna

Modalna *Mo* (dominanta, moda, wartość najczęstsza) - jest to wartość cechy statystycznej, która występuje najczęściej. W zbiorze wyników może wystąpić kilka takich, które stanowią wartość modalną.

Statystyka matematyczna

Kwantyle

Kwantyle *Q* - definiuje się jako wartości cechy badanej zbiorowości, przedstawionej w postaci szeregu statystycznego, które dzielą zbiorowość na określone części pod względem liczby jednostek, części te pozostają do siebie w określonych proporcjach.

Kwantyl rzędu 1/2 to inaczej *mediana*.

Kwantyle rzędu 1/4, 1/2, 3/4 są inaczej nazywane *kwartylami*.

Kwantyle rzędu 1/10, 2/10,..., 9/10 to inaczej *decyle*.

Kwantyle rzędu 1/100, 2/100,..., 99/100 to inaczej *percentyle*.

Statystyka matematyczna

Kwartyle

Kwartyl pierwszy Q_1 dzieli zbiorowość na dwie części w ten sposób, że 25% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q_1, a 75% równe bądź wyższe od tego kwartyla.

Kwartyl drugi Q_2 to *mediana*.

Z kolei kwartyl trzeci Q_3 dzieli zbiorowość na dwie części w ten sposób, że 75% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q_3, a 25% równe bądź wyższe od tego kwartyla.

Statystyka matematyczna

Mediana

Mediana (wartość środkowa) *Me* – środkowa liczba w uporządkowanej niemalejąco próbce (dla próbki o liczności nieparzystej) lub średnia arytmetyczna dwóch liczb środkowych (dla próbki o liczności parzystej).

Mediana dzieli zbiorowość na dwie równe części; połowa jednostek ma wartości cechy mniejsze lub równe medianie, a połowa wartości cechy równe lub większe od *Me* - stąd nazwa wartość środkowa.

Statystyka matematyczna

Decyle

Decyle np. decyl pierwszy oznacza, że 10% jednostek ma wartości cechy mniejsze bądź równe od decyla pierwszego, a 90% jednostek wartości cechy równe lub większe od decyla pierwszego.

15

Statystyka matematyczna

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
16,3	16,3	16,4	16,5	16,6	16,6	16,8	16,8	17,0	17,1	17,3	17,3	17,5	17,7	17,7	17,7	17,8	17,9	17,9	17,9

Q_1 16,6

Q_2 Me 17,2

Q_3 17,7

$x_{śr} = 17,16$

$Mo = 17,7$ i $17,9$

16

Statystyka matematyczna

Miary rozproszenia (zmienności, dyspersji)

Z reguły miary rozproszenia odnoszą się do określania różnic pomiędzy obserwacjami a wartością średnią.

Do najczęściej stosowanych miar rozproszenia należą:

- rozstęp
- wariancja
- odchylenie standardowe
- współczynnik zmienności

Statystyka matematyczna

Rozstęp

Rozstęp to różnica pomiędzy wartością maksymalną, a minimalną cechy - jest miarą charakteryzującą empiryczny obszar zmienności badanej cechy, nie daje on jednak informacji o zróżnicowaniu poszczególnych wartości cechy w zbiorowości.

$$R = x_{\max} - x_{\min}$$

Statystyka matematyczna

Wariancja

Wariancja jest to średnia arytmetyczna kwadratów odchyleń poszczególnych wartości cechy od średniej arytmetycznej zbiorowości.

$$s^2 = \frac{1}{n-1}\sum_{i-1}^{n}\left(x_i - x_{śr}\right)^2$$

n>1!!!

dla n =2

$$s^2 = 2\left(\frac{R}{2}\right)^2$$

19

Statystyka matematyczna

Odchylenie standardowe

Definiowane jako miara rozproszenia uzyskanych poszczególnych wartości oznaczeń wokół wartości średniej:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - x_{śr}\right)^2}{n-1}}$$

gdzie:

x_i – wartość pojedynczego wyniku oznaczenia;
$x_{śr}$ – średnia arytmetyczna z uzyskanych wyników;
n – liczba uzyskanych wyników;

20

Statystyka matematyczna

Odchylenie standardowe jest równe zeru wtedy i tylko wtedy, gdy wszystkie wyniki są identyczne. W każdym innym przypadku wielkość ta jest dodatnia. Zatem im większe rozproszenie wyników, tym wartość *s* jest większa.

Rozrzut wyników związany jest z każdym postępowaniem analitycznym. Możliwe jest natomiast, że zjawiska tego nie udało się zaobserwować ze względu na np. zbyt niską rozdzielczość stosowanego przyrządu pomiarowego.

Statystyka matematyczna

Serie wyników pomiarów uzyskane z wykorzystaniem przyrządów kontrolno-pomiarowych o różnej rozdzielczości.

	Przyrząd 1	Przyrząd 2	Przyrząd 3
Uzyskane wyniki	17	16,8	16,83
	17	17,1	17,14
	17	16,9	16,88
	17	17,4	17,43
	17	17,3	17,27
	17	17,2	17,24
	17	17,0	16,96
Wartość odchylenia standardowego	**0**	**0,22**	**0,223**

Odchylenie standardowe:

a. dla znanej wartości rzeczywistej μ_x

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

b. dla nieznanej wartości rzeczywistej (oszacowanie $x_{śr}$)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_{śr})^2}{n-1}}$$

c. względne odchylenie standardowe

$$RSD\ (s_R) = \frac{s}{x_{śr}}$$

d. odchylenie standardowe średniej arytmetycznej

$$\overline{s} = \frac{s}{\sqrt{n}}$$

e. odchylenie standardowe metody (ogólne)

$$s_g = \sqrt{\frac{1}{n-k}\sum_{i=1}^{k} s_i^2 (n_i - 1)}$$

gdzie:

n - ogólna liczba oznaczeń
k - liczba serii

dla równolicznych serii wzór upraszcza się do postaci:

$$s_g = \sqrt{\frac{1}{k} \cdot \sum_{i=1}^{k} s_i^2}$$

25

Statystyka matematyczna

Współczynnik zmienności

Współczynnik zmienności (*CV*) powstaje przez pomnożenie wartości *RSD* przez 100%:

$$CV = RSD \cdot 100\%$$

Współczynnik zmienności - jest ilorazem bezwzględnej miary zmienności cechy i średniej wartości tej cechy, jest wielkością niemianowaną, najczęściej podawaną w procentach.

Współczynnik zmienności stosuje się w porównaniach zróżnicowania:

- kilku zbiorowości pod względem tej samej cechy,
- tej samej zbiorowości pod względem kilku różnych cech.

26

Statystyka matematyczna

Miary asymetrii

Wskaźnik skośności - wielkość bezwzględna wyrażona jako różnica między średnią arytmetyczną a modalną.

Współczynniki skośności (asymetrii) - są stosowane w porównaniach, do określenia siły oraz kierunku asymetrii, są to liczby niemianowane, im większa ich wartość tym silniejsza asymetria.

Pozycyjny współczynnik asymetrii określa kierunek i siłę asymetrii jednostek znajdujących się miedzy pierwszym z trzecim kwartylem.

Statystyka matematyczna

Miary koncentracji

Współczynnik skupienia (koncentracji) (kurtoza) *K* - jest miarą skupienia poszczególnych obserwacji wokół średniej. Im wyższa wartość współczynnika tym bardziej wysmukła krzywa liczebności, większa koncentracja wartości cech wokół średniej.

Statystyka matematyczna

Rozkłady zmiennych losowych

Zastosowanie określonej procedury analitycznej sprawia, że w sposób jednoznaczny wyznaczony jest rozkład wyników pomiarów (*cechy*), które można traktować jako niezależne zmienne losowe. Wynik jest konsekwencją przeprowadzenia pomiaru. Zbiór otrzymanych wyników oznaczeń tworzy rozkład (*empiryczny*).

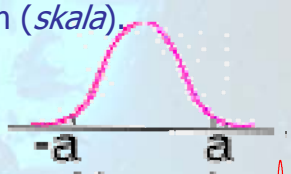
Rozkład normalny

niezwykle ważny rozkład prawdopodobieństwa w wielu dziedzinach (rozkład Gaussa)

definiowany dwoma parametrami: średnią (odpowiada za *położenie* rozkładu) i odchyleniem standardowym (*skala*).

Własności rozkładu normalnego $N(\mu_x, s)$:

- wartość oczekiwana: μ_x
- mediana: μ_x
- wariancja: s^2



29

Statystyka matematyczna

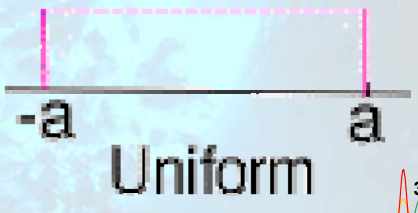
Rozkład jednostajny (prostokątny)

ciągły rozkład prawdopodobieństwa, dla którego gęstość prawdopodobieństwa w przedziale $\langle -a, +a \rangle$ jest stała i różna od zera, a poza nim równa zeru.

Ponieważ rozkład jest ciągły, nie ma większego znaczenia czy punkty *-a* i *+a* włączy się do przedziału czy nie. Rozkład jest określony parą parametrów *-a* i *+a.*

Własności rozkładu jednostajnego:

- wartość oczekiwana: *0*
- mediana: *0*
- wariancja: $a^2/3$



30

Statystyka matematyczna

Rozkład trójkątny

Własności rozkładu trójkątnego w przedziale $\langle -a, +a \rangle$:

- wartość oczekiwana: *0*
- mediana: *0*
- wariancja: $a^2/6$

31

Statystyka matematyczna

Znajomość rozkładu zmiennej losowej - pełna informacja na temat badanej cechy (może to być np.: stężenie, zawartość, właściwość fizykochemiczna).

Rzadko istnieje możliwość dysponowania taką pełną informacją.

Wnioskowanie na temat cechy oparte o analizę pewnej ograniczonej liczby elementów (*próbek*) reprezentujących fragment całego zbioru opisywanego rozkładem.

Wtedy należy wnioskować o badanej cesze na podstawie oszacowania niektórych jej parametrów (*parametry statystyczne*) lub na podstawie rozkładu empirycznego.

32

Statystyka matematyczna

Weryfikacja hipotez statystycznych

Hipoteza to sąd o populacji oparty na prawdopodobieństwie, przyjęty w celu wyjaśnienia jakiegoś zjawiska, prawa lub faktu i wymagający sprawdzenia; przypuszczenie.

Weryfikacją hipotez nazywamy sprawdzanie sądów o populacji, sformułowanych bez zbadania jej całości. Przebieg procedury weryfikacyjnej wygląda następująco:

1. Sformułowanie hipotezy zerowej i alternatywnej

Hipoteza zerowa – H_o - prosta postać hipotezy poddana testom

Hipoteza alternatywna – H_1 - przeciwstawiona hipotezie zerowej

2. Wybór odpowiedniego testu

Test służy do sprawdzania hipotezy.

Statystyka matematyczna

3. Określenie poziomu istotności α.

4. Wyznaczenie obszaru krytycznego testu

Wielkość obszaru krytycznego wyznacza dowolnie mały poziom istotności α, natomiast jego położenie określane jest przez hipotezę alternatywną.

5. Obliczenie parametru testu na podstawie próby

Wyniki próby opracowuje się w odpowiedni sposób, zgodnie z procedurą wybranego testu i są one podstawą do obliczenia statystyki testowej.

Statystyka matematyczna

6. Podjęcie decyzji

Wyznaczona na podstawie próby wartość statystyki porównywana jest z wartością krytyczną testu.

· Jeżeli wartość ta znajdzie się w obszarze krytycznym to hipotezę zerową należy odrzucić jako nieprawdziwą.

· Jeżeli natomiast wartość ta znajdzie się poza obszarem krytycznym oznacza to, że brak jest podstaw do odrzucenia hipotezy zerowej. Stąd wniosek, że hipoteza zerowa może być prawdziwa.

Statystyka matematyczna

Rodzaje błędów popełnianych przy weryfikacji:

błędy I-go rodzaju - odrzucenie hipotezy zerowej H_o gdy jest ona prawdziwa

błędy II-go rodzaju - przyjęcie H_o gdy jest ona fałszywa

		hipoteza	
		prawdziwa	fałszywa
hipoteza	przyjęta	✓	błąd II-go rodzaju
	odrzucona	błąd I-go rodzaju	✓

Statystyka matematyczna

Coraz częściej do weryfikowania hipotez statystycznych wykorzystywane są różnego rodzaju programy komputerowe (np. program *Statistica*).

W takim przypadku sposób postępowania ogranicza się do wyliczenia, dla badanego zbioru danych, parametru p – po wybraniu odpowiedniego testu statystycznego.

Tak obliczona wartość porównuje się następnie z przyjętą wartością poziomu istotności α.

Jeżeli obliczona wartość p jest mniejsza od wartości α: $p < \alpha$ odrzuca się hipotezę zerową H_o. W przeciwnym przypadku hipotezy zerowej nie odrzuca się.

Statystyka matematyczna

Cyfry znaczące, reguły zaokrąglania liczb

Problem poprawnego zapisywania wyników pomiarów polega najczęściej na zrozumieniu czym są cyfry znaczące i na poznaniu reguł towarzyszących zaokrąglaniu liczb.

Cyfry znaczące to w zapisie dziesiętnym danej liczby wszystkie jej cyfry bez początkowych zer. Aby określić ilość cyfr znaczących w liczbie należy „czytać" liczbę od lewej strony aż do napotkania pierwszej cyfry różnej od zera. Ta cyfra i każda następna to są właśnie cyfry znaczące.

Statystyka matematyczna

Dla przykładu w poniżej przedstawionych liczbach podkreślono te cyfry, które są cyframi znaczącymi.

11,23

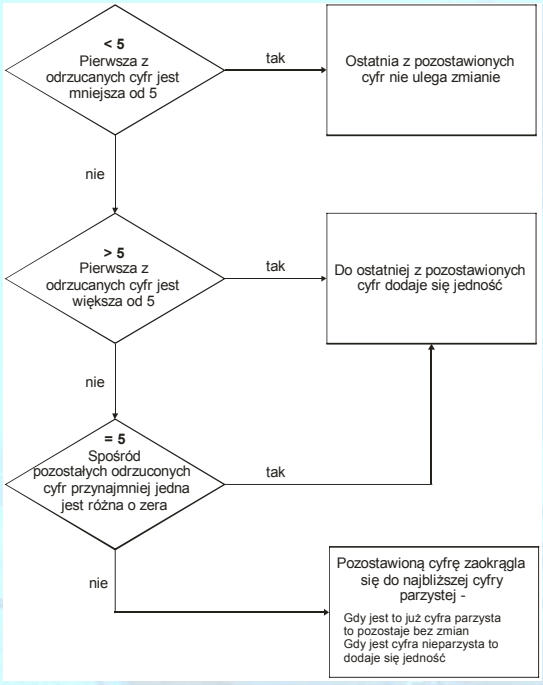
0,00123

0,1200

203,20

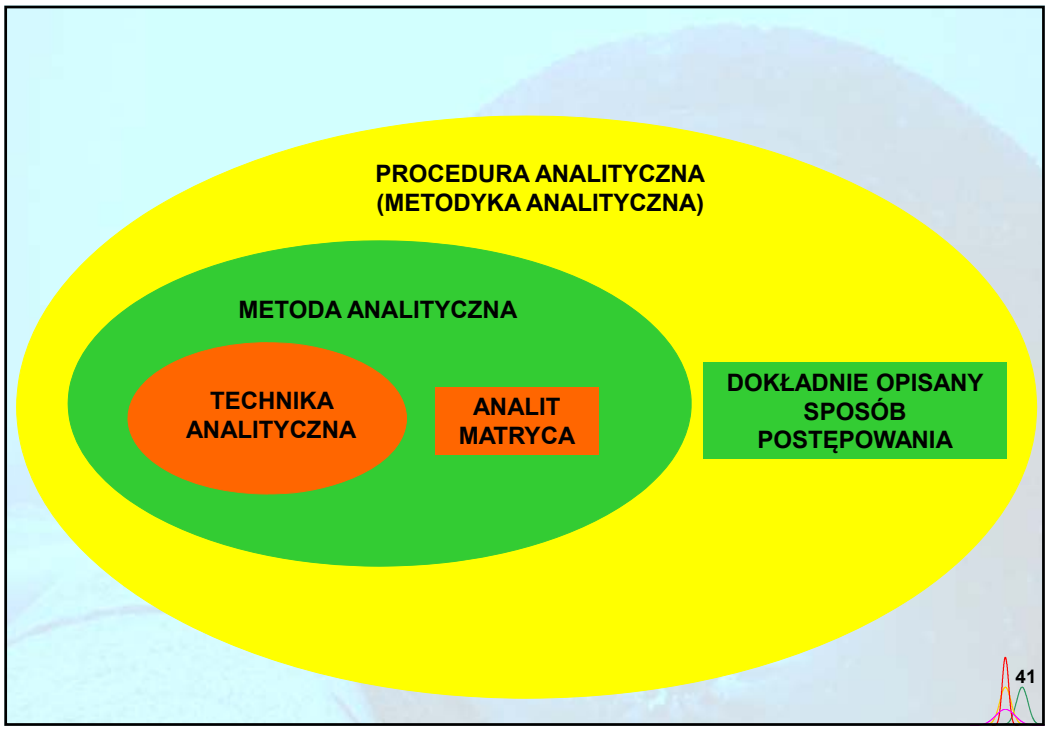
$3{,}3 \cdot 10^3$

39



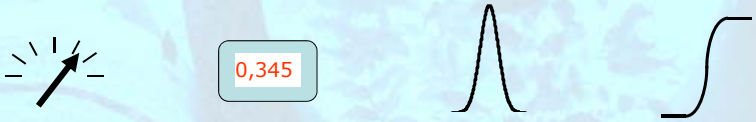
40

40



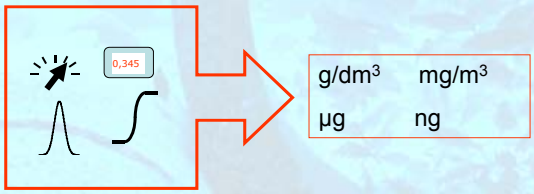
Sygnał - następstwo i konsekwencja przeprowadzonego pomiaru – główny obiekt zainteresowań analityka.

Cel pracy analityka - uzyskanie informacji analitycznej o badanym obiekcie na podstawie otrzymanego w wyniku zastosowania odpowiedniej procedury pomiarowej sygnału wyjściowego.



Pod postacią sygnału jest kodowana informacja na temat badanej próbki. Rola analityka polega, zatem na „rozkodowaniu” otrzymanego sygnału i to w taki sposób, aby uzyskana informacja była jak najbardziej miarodajna.

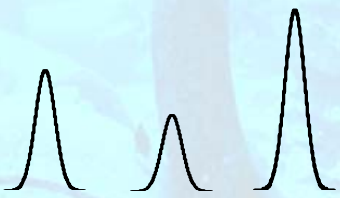
Narzędziem służącym do tego rozkodowania jest proces analityczny i stosowane w nim procedury analityczne.



43

Cechą każdego sygnału jest jego wielkość. Przy niektórych pomiarach sygnałowi można nadać także parametr pozycji (umiejscowienia).

Parametry metody analitycznej wyznaczane są w oparciu o analizę otrzymanych wartości sygnałów i o tym należy pamiętać w trakcie ich określania.



44

Precyzja (ang. *precision*) – zgodność pomiędzy niezależnymi wynikami uzyskanymi w trakcie analizy danej próbki z zastosowaniem danej procedury analitycznej.

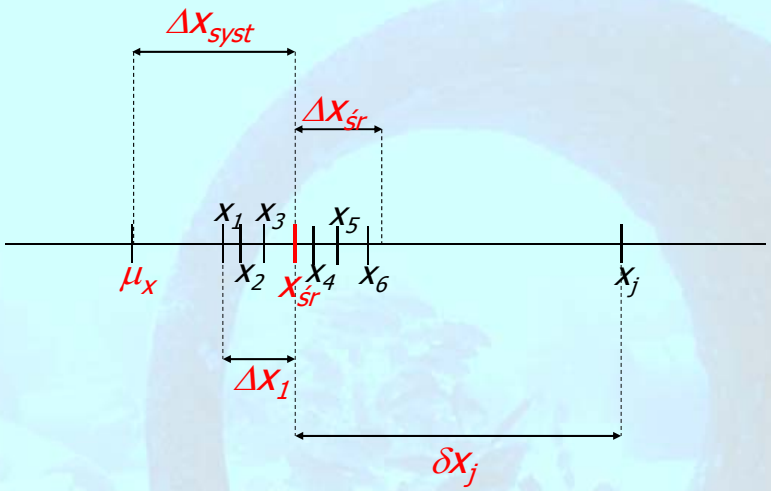
Powtarzalność (ang. *repeatability*) – precyzja wyników uzyskanych w tych samych warunkach pomiarowych (dane laboratorium, analityk, instrument pomiarowy, odczynniki).

Precyzja pośrednia (ang. *intermediate precision*) – długoterminowe odchylenie procesu pomiarowego, do którego wyznaczenia wykorzystuje się odchylenie standardowe serii pomiarów uzyskanych w danym laboratorium w kilkutygodniowym okresie czasu. Precyzja pośrednia jest pojęciem szerszym od powtarzalności.

Odtwarzalność (ang. *reproducibility*) – precyzja wyników uzyskanych w różnych laboratoriach z zastosowaniem danej metody pomiarowej.

Dokładność *(*ang. *accuracy)* – zgodność pomiędzy uzyskanym wynikiem pomiaru z wartością rzeczywistą (oczekiwaną).

Poprawność *(*ang. *trueness)* – zgodność wyniku oznaczenia (obliczonego na podstawie serii pomiarów) z wartością oczekiwaną.



Δx_{syst} - błąd systematyczny metody analitycznej
Δx_i - błąd przypadkowy pojedynczego wyniku
$\Delta x_{śr}$ - błąd przypadkowy średniej arytmetycznej
δx_j - błąd gruby

47

Ze względu na sposób podawania wartości błędu wyniku oznaczenia można wyróżnić:

błąd bezwzględny - d_x, który opisać można następującą zależnością:

$$d_{x_i} = x_i - \mu_x$$

błąd względny - ε_x, opisywany za pomocą równania:

$$\varepsilon_{x_i} = \frac{d_{x_i}}{\mu_x}$$

48

Z kolei biorąc pod uwagę źródła błędów, wyróżnić można:

- błędy metodyczne
- błędy instrumentalne
- błędy osobowe

49

Dokładność i miary niedokładności

1. dokładność wyniku pojedynczego oznaczenia (DOKŁADNOŚĆ):

$$d_{x_i} = x_i - \mu_x = \Delta x_{syst} + \Delta x_i + \delta x_i$$

2. dokładność wyniku analizy (POPRAWNOŚĆ):

$$d_{x_{śr}} = x_{śr} - \mu_x = \Delta x_{syst} + \Delta x_{śr}$$

3. dokładność procedury analitycznej:

$$d_{x_{met}} = E(x_{met}) - \mu_x = \Delta x_{syst}$$

50

BŁĄD GRUBY

- wynik jednorazowego wpływu przyczyny działającej przejściowo,
- występuje przy niektórych pomiarach,
- przyczyny to np.: pomyłka przy odczycie wskazań przyrządu pomiarowego, pomyłka w obliczeniach,
- zmienna losowa - jednak o nieznanym rozkładzie i nieznanej wartości oczekiwanej,
- najłatwiejszy do wykrycia i usunięcia,
- bywa zarówno dodatni jak i ujemny (inaczej niż w przypadku błędu systematycznego).

51

BŁĄD SYSTEMATYCZNY

- błąd systematyczny stały - wartość nie zależy od poziomu zawartości analitu – a_{syst}
- błąd systematyczny zmienny - wartość błędu zależy (liniowo) od poziomu zawartości analitu - $b_{syst} \cdot \mu_x$

$$\Delta x_{syst} = a_{syst} + b_{syst} \cdot \mu_x$$

$$x_{śr} = \mu_x + \Delta x_{syst} = \mu_x + a_{syst} + b_{syst} \cdot \mu_x = a_{syst} + (1 + b_{syst})\, \mu_x$$

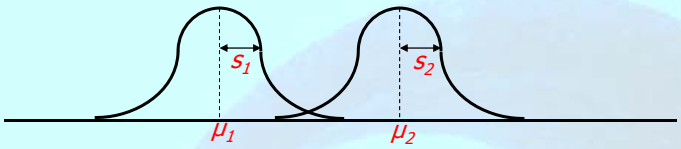
52

Rozrzut wyników

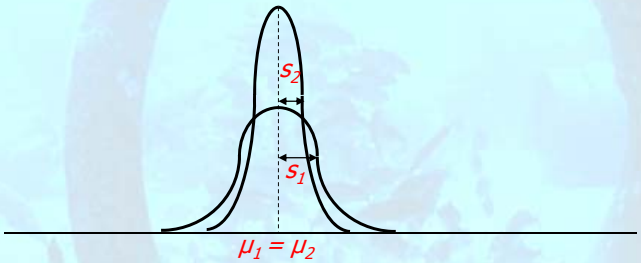
Błędy przypadkowe:

- występują zawsze,
- są zazwyczaj małe i powodują, że wynik nieznacznie różni się od wartości rzeczywistej,
- przyczyna powstawania - zespół czynników przypadkowych,
- wielkość błędu - zmienna losowa,
- zmniejszanie wielkości błędu przez zwiększanie ilości pomiarów,
- nie można ich wyeliminować stosując poprawki,
- rozkład Gaussa - opis rozkładu błędów przypadkowych.

53



np.: wykonanie daną metodą pomiarową (stałe odchylenie standardowe) analiz dla próbek o różnej zawartości analitu



np.: wykonanie analiz dla tej samej próbki (taka sama wartość oczekiwana) dwiema niezależnymi metodami (różne wartości odchyleń standardowych)

54