Analiza niepewności pomiarowych
Wstęp do analizy danych
Wykład 1

1 Uwagi wstępne

Prowadzony przeze mnie wykład ma na celu przygotowanie Państwa do analizowania wyników pomiarów. Podczas wykładu posługuję się prezentacją, która jest dostępna poprzez stronę:

http://pracownie1.fuw.edu.pl/anipw/400.html

jako **Slajdy/transparecje do wykładu**. Na tej samej stronie znajdują się też dodatkowe matriały związane z przedmiotem:

- Instrukcja jak pisać raport końcowy
- Konstrukcja histogramu
- Przykładowy raport

Materiały te mogą być pomocne przy opracowaniu danych zebranych podczas pomiarów i przygotowaniu raportu końcowego. Teksty kolejnych **Wykładów** są skrótami moich wykładów – można je traktować jako komentarze rozbudowujące i wyjaśniające treść **Slajdów/transparencji do wykładów**. Znajdą w nich Państwo także nieco dodatkowych uwag, na które podczas wykładu nie starczyło czasu. Wykłady 4 – 7 kończą przykładowe zadania do omawianej w nich części materiału.

Główna część materiału dotyczy zagadnień związanych z analizą danych: wyznaczaniem najlepszej oceny wartości mierzonej wielkości fizycznej i oceną dokładności uzyskanego wyniku – wyznaczaniem niepewności wyniku. Wyznaczanie najlepszej oceny wartości mierzonej wielkości i jej niepewności wymaga dokonania szczegółowej analizy metody pomiaru oraz przebiegu jej realizacji, tj. przebiegu pomiaru. Do głębszego zrozumienia procesu pomiaru i sformułowania metod analizy jego wyników posłużą nam modele matematyczne korzystające z pojęć rachunku pradopodobieństwa i metod statystyki matematycznej. Zaczniemy od najprostszych metod opisu i prezentacji dużych zbiorów danych liczbowych, czyli od **statystyki opisowej**.

2 Charakterystyki zbiorów danych liczbowych

Przyjrzyjmy się danym dotyczącym mas studentów I roku (anonimowa ankieta została przeprowadzona na Wydziale Fizyki w roku akademickim 2000/2001) przedstawionym w postaci histogramu – Figure 1, poniżej. Podczas wykładu prezentowałem tabelę zebranych danych, w której studentki i studenci, kolejno, wpisywali swoją masę i wzrost. Taka tabela zawiera pełną informację, ale ta informacja do wielu celów jest zbyt bogata – na przykład do porównywania różnych grup studenów bardziej przydatny jest mniej szczegółowy opis syntetycznie przedstawiający cały zbiór danych. Często używaną charakterystyką jest wartość **średnia**.

Średnia μ: dla zbioru danych $i \to x_i, i = 1, ..., N$ definiujemy wartość **średnią** μ:

$$\mu := \frac{1}{N} \sum_{i=1}^{i=N} x_i = \frac{1}{N}(x_1 + x_2 + \ldots + x_N).$$

Może nas także interesować szerokość rozkładu zbioru wartości x_i wokół ich średniej. Powszechnie stosowaną miarą rozrzutu danych jest **średnie odchylenie kwadratowe**.

Średnie odchylenie kwadratowe σ^2:

$$\sigma^2 := \frac{1}{N} \sum_{i=1}^{i=N} (x_i - \mu)^2$$

oraz pierwiastek z tej wielkości $\sigma = \sqrt{\sigma^2}$. Wielkość σ nazywana jest też **dyspersją rozkładu**.

Jak łatwo sprawdzić suma kwadratów odchyleń jest najmniejsza, gdy odchylenia liczone są od wartości średniej.

Uwaga: W powyższych definicjach dla oznaczenia średniej i dyspersji użyte zostały greckie litery μ i σ dla podkreślenia, że obliczenia dotyczą całej populacji. W przyszłości wielokrotnie będziemy konstruowali wyrażenia dające dobre przybliżenie wartości μ i σ obliczane na podstawie **próbki** danych tj. danych z losowo wybranego podzbioru pełnej populacji. Dla wielkości μ odpowiednim przybliżeniem jest wielkość równa śedniej danych próbki, którą często oznacza się jako $\bar{x}$ (kreska nad x wskazuje, że próbka została pobrana ze zbioru wielkości x).

Mediana i średnie odchylenie bezwzględne

Zdefiniowane wyżej charakterystyki są stosowane powszechnie, mają jednak istotną wadę, gdy liczba przypadków (liczebność populacji) jest niewielka. Obie charakterystyki – średnia i średnie odchylenie kwadratowe są czułe na pojedyncze wartości znacznie odbiegające od pozostałych wartości x_i w populacji (w próbce). Taka pojedyncza nawet wartość – powiedzmy x_k – znacznie większa od pozostałych, gdy liczebość populacji N jest niewielka, znacząco wpływa na wartość tych charakterystyk obliczanych dla całej populacji. Z tego powodu używana jest często charakterystyka wolna od tej wady: **mediana** i związane z nią **średnie odchylenie bezwzględne**:

medianę m dla zbioru liczb x_i, $i = 1, \ldots, N$ uporządkowanych rosnąco (lub malejąco) definiujemy, jako $m = x_{n+1}$, gdy liczebność N zbioru jest liczbą nieparzystą, $N = 2n + 1$, a gdy N jest liczbą parzystą, $N = 2n$, to dla jednoznaczności przyjmujemy $m = (x_{n-1}+x_{n+1})/2$. Tak zdefiniowana mediana dzieli zbiór wartości x_i na dwa równoliczne podzbiory.

Mediana jest całkowicie „nieczuła“ na pojedyncze wartości x_i znacznie mniejsze lub zancznie większe od pozostałych wartości zbioru.

Z medianą związanae jest też **średnie odchylenie bezwzgędne** d**:**

średnie odchylenie bezwzględne d**:** jest miarą rozrzutu zbioru danych x_i, $i = 1, \ldots, N$ wokół **mediany** m tego zbioru:

$$d := \frac{1}{N} \sum_{i=1}^{i=N} |x_i - m|$$

Jak łatwo sprawdzić suma watrości bezwzględnych odchyleń jest najmniejsza, gdy odchylenia liczone są od mediany.

Dla mas studentów i studentek I roku odpowienie charakterystyki wynoszą:

średnia mas studentów $\mu_M = 73,06$ kg
mediana mas studentów $m_M = 72$ kg

średnia mas studentek $\mu_K = 56,70$ kg
mediana mas studentek $m_K = 55,5$ kg

Dla wszystkich, tj. studentów i studentek razem:

średnia mas $\mu = 69,17$ kg
mediana mas m =70 kg
.

Graficzna prezentacja danych – histogram

Wygodnym sposobem prezentacji danych jest **histogram**. Pełne dane dotyczące (deklarowanych) mas studentów przedstawia histogram (Figure 1) poniżej. Studenci podawali jako swoje masy całkowite liczby kilogramów, a więc histogram z krokiem 1 kg zawiera pełną informację – równoważną pełnej tabeli – o grupie studentów.
Brzegi przedziałów na osi poziomej zostały umieszczone w wartościach połówkowych, a każdy przedział ma szerokość 1 kg – każdą wartość podaną w tabeli można więc jednoznacznie przyporządkować do konkretnego przedziału na osi poziomej. Zwykle przyjmuje się, że przedziały histogramowania są prawostronnie domknięte – tzn. jeśli wśród danych pojawia się wartość równa wartości przyjętej jako granica przedziałów to doliczana jest jak należąca do przedziału po lewej stronie tej granicy.
Opis osi poziomej zawiera poza cyframi także nazwę wielkości i jej jednostkę (kg) – gdyby w tekście odwołującym się do tego histogramu był używany symbol na oznaczenie masy studenta, np. M, to i ten symbol powinien znaleźć się w opisie osi.
Opis osi pionowej zawiera nazwę wielkości i cyfry reprezentujące liczby studentów o masach mieszczących się w kolejnych przedziałach (wielkości niemianowane) – wysokoć „słupka“ odpowiada liczbie osób o masach zawartych wewnątrz przedziału stanowiącego podstawę słupka. Warto zwrócić uwagę, że między słupkami nie ma przerw – masa jest wielkością ciągłą, a więc podane przez studentów liczby całkowiete należy rozumieć jako reprezentujące wartości leżące wewnątrz całego przedziału.

Jak widać histogram mas studentów z krokiem $\Delta = 1$ kg ma dosyć bogatą strukturę (prawdopodobnie wynikającą z tendencji do podawania „okrągłych” wartości). Jeśli chcemy mieć bardziej syntetyczny obraz danych, to możemy wykonać histogram z większym krokiem. Dwa następne histogramy wykonano z krokiem $\Delta = 3$ kg.

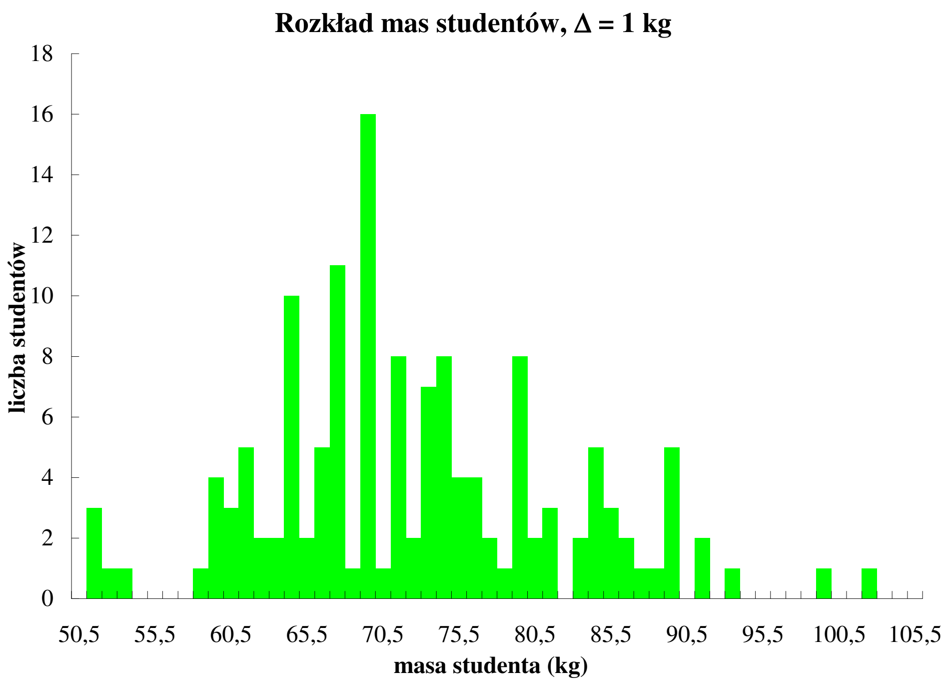

Figure 1: Masy studentów. Przedział histogramowania 1 kg.

Warto porównać oba histogramy dla przedziału histogramowania $\Delta = 3$ kg – ich kształty znacznie się różnią. Ten fakt należy potraktować jako wskazówkę, że przed wyciąganiem wniosków i formułowaniem hipotez dobrze jest sprawdzić, czy nie są one wynikiem przyjętych szerokości i położeń granic przedziałów histogramowania.

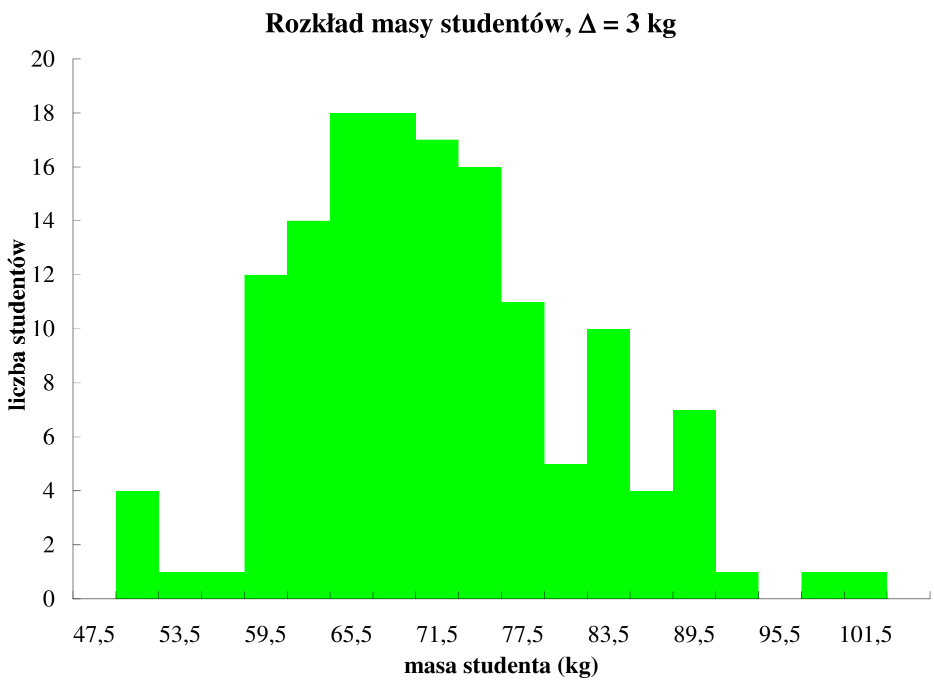

Figure 2: Masy studentów. Przedział histogramowania 3 kg.

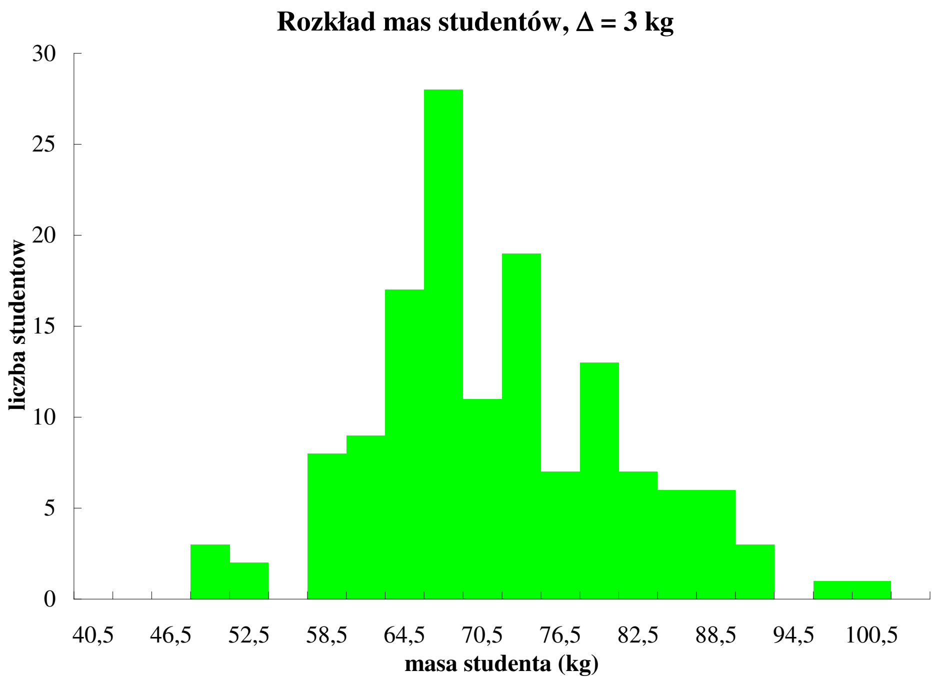

Figure 3: Masy studentów. Przedział histogramowania 3 kg – lewy brzeg przesunięty w stosunku do poprzedniego przykładu.

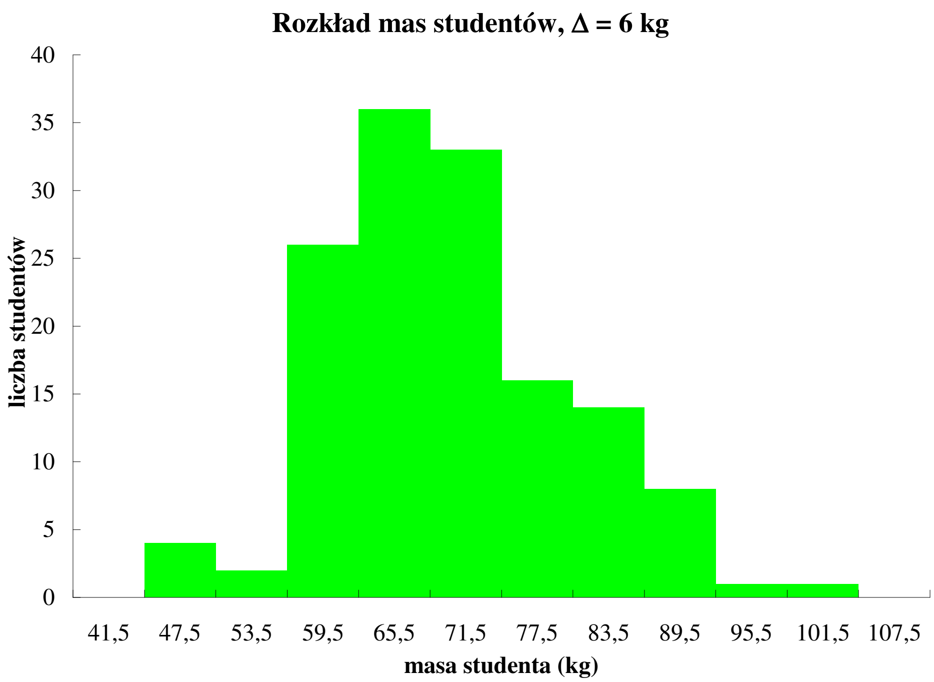

Figure 4: Masy studentów. Przedział histogramowania 6 kg.

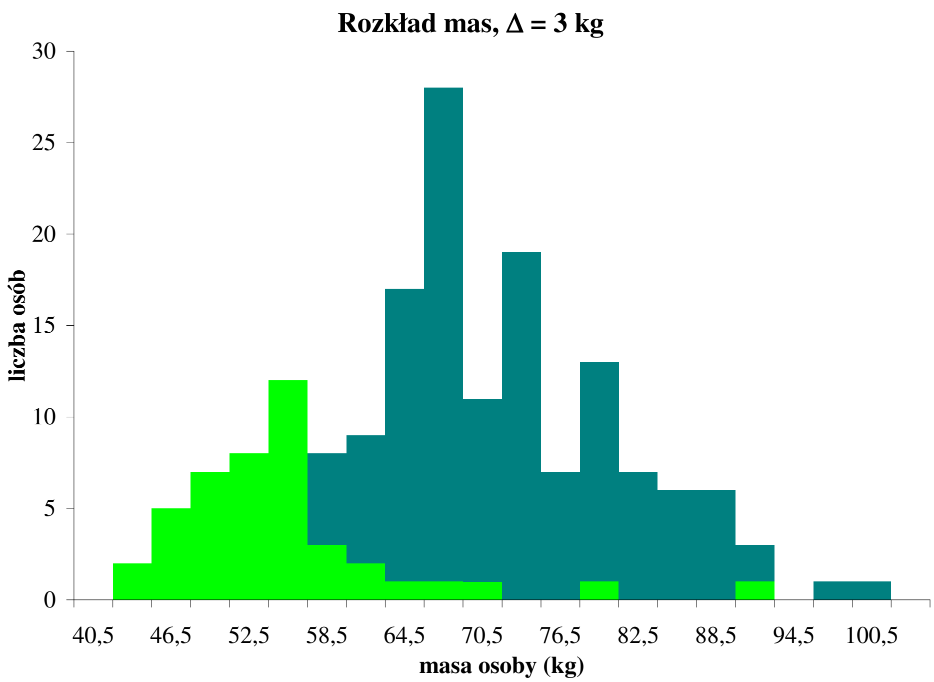

Figure 5: Masy studentów (ciemny kolor) i studentek (jasny kolor). Przedział histogramowania 3 kg – „słupki” studentów naniesione od górnych krawędzi historgrmu dla studentek.

Na koniec jeszcze histogram ostatnich cyfr liczb pojawiających się w *tabliczce mnożenia* liczb od 1 do 10 – wielkości dyskretnych.

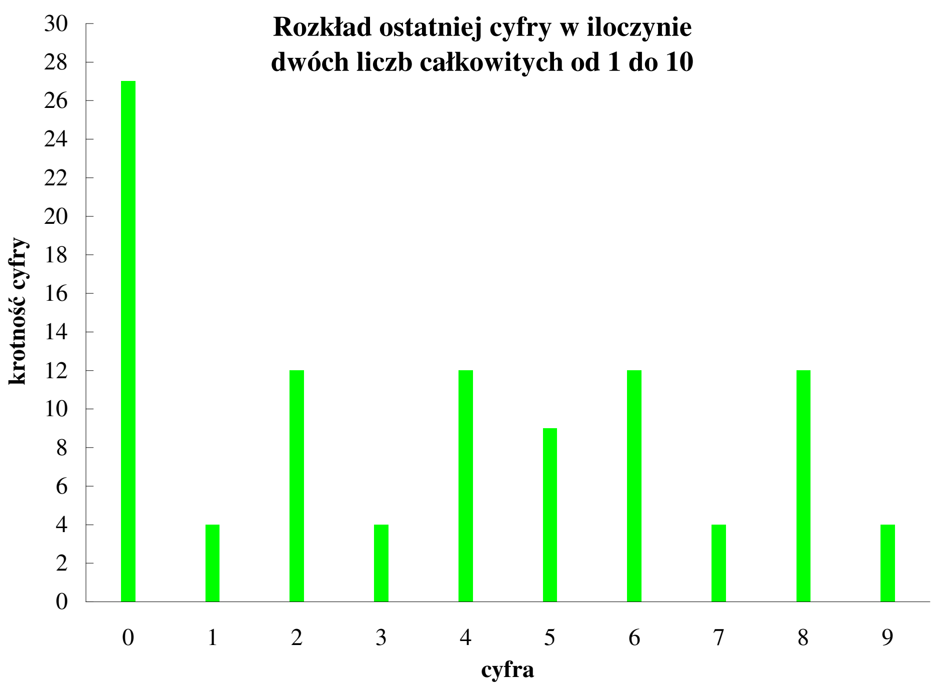

Figure 6: Rozkład ostatnej cyfry w iloczynach liczb całkowitych od 1 do 10

3 Bezpośredni pomiar pojedynczej wielkości fizycznej

Istotą pomiaru jest porównanie interesującej nas wielkości z przyjętym wzorcem $\longrightarrow$ jednostką mierzonej wielkości. Zgodnie z polskim prawem w pomiarach żywamy jednostek Układu SI. Jego konstrukcja jest wielkim osiągnięciem cywilizacji: jednakowy, precyzyjnie zdefiniowany układ jednostek został przyjęty na całym świecie – poza Stanami Zjednoczonymi, Liberią i Mjanmą (Birmą). W naukach przyrodniczych także w Stanach Zjednoczonych stosowany jest Układ SI. Od chwili przyjęcia we Francji definicji metra i sekundy, w roku 1791, oraz kilograma, w roku 1795, definicje jednostek układu były wielokrotnie zmieniane w celu zwiększenia precyzji wzorców, a z tym także dokładności wykonywanych pomiarów. Wraz z odkrywaniem nowych zjawisk wprowadzano także dalsze jednostki.

Od 2018 roku jednostki Układu SI: amper, kelwin, sekunda, metr, kilogram, kandela i mol oraz ich pochodne definiowane są poprzez przyjęcie jako dokładnych wartości podstawowych stałych fizycznych:

- prędkości światła w próżni $c = 299792458$ m/s
- stałej Plancka $h = 6,62607015 cdot 10^{-34}$ J s
- ładunku elementarnego $e = 1,602176634 \cdot 10^{-19}$ C
- stałej Boltzmanna $k = 1,380649 \cdot 10^{-23}$ J/K
- stałej Avogadry $N_A = 6,02214076 \cdot 10^{23}$ 1/mol
- częstotliwości nadsubtelnego przejścia w atomie cezu 133 $\Delta\nu_{Cs} = 9192631770$ Hz
- skuteczności świetlnej monochromatycznego promieniowania o częstotliwości $540 \cdot 10^{12}$ Hz, $K_{cd} = 683$ lm/W

Definicje jednostek Układu SI można znaleźć na stronie Głównego Urzędu Miar:

https://www.gum.gov.pl/

lub na stronie Międzynarodowego Biura Miar i Wag (BIPM) – oficjalny tekst po angielsku:

https://www.bipm.org/en/measurement-units/

Wykonywanie pomiarów

Wykonując pomiary nie korzystamy bezpośrednio z wzorców odpowiednich jednostek, ale używamy odpowienio wykalibrowanych przyrządów. Podczas pomiaru staramy się, tak dalece jak jest to możliwe, wyeliminować wszelkie czynniki, które mogłyby zakłócić jego przebieg. Dokładność uzyskanego wyniku zależy od:

- przyjętej metody pomiaru
- sposobu i dokładności eliminowania czynników zakłócających jego przebieg
- jakości przyrządów pomiarowych $\longrightarrow$ dokładności i powtarzalności odtwarzania przez nie wzorca mierzonej jednostki

Podczas następnych wykładów zajmiemy się analizą wpływu wymienionych wyżej czynników na dokładność pomiaru i budowaniem modeli matematycznych procesu pomiaru. Będziemy zakładali, że **dokładna wartość** mierzonej wielkości istnieje i jest dobrze zdefiniowaną, mianowaną liczbą rzeczywistą, a pomiar służy jak najdokładniejszemu poznaniu tej wartości. Należy jednak pamiętać, że w wielu sytuacjach mamy do czynienia z wielkościami, dla których pojęcie dokładnej wartości ma ograniczony sens. Tak jest np. w przypadku masy człowieka – z każdym oddechem nasza masa nieznacznie się zmienia (wydychamy dwutlenek węgla powstający w procesach przemiany materii). Nie będziemy dalej wracali do tego zagadnienia.

Założymy, że:

- wielkość mierzona ma jednoznacznie określoną, ale nie znaną nam wartość μ;
- jako wynik pomiaru otrzymujemy wartość x;
- różnicę $\varepsilon = x - \mu$ nazywamy **błędem pomiaru**;
  ponieważ nie znamy dokładnie wynosi mu, to nie jest nam także znana dokładna warotść błędu ε;

Na podstawie serii pomiarów będziemy starali się zdobyć wiedzę o rozkładzie wartości ε, w szczególności będziemy oceniali zakres w jakim mieści się ε:

$$-u \leq \varepsilon \leq u.$$

Wartość u nazywamy **niepewnością pomiaru**.