KONSTRUKCJA HISTOGRAMU

Przy konstruowaniu histogramu przydatne jest wyznaczenie wartości najmniejszej x_{min} i największej, x_{max}, w próbce. Wartości te pozwalają ocenić rozpiętość histogramu. Niech symbol N oznacza liczebność próbki. Sam histogram budujemy
w następujący sposób:

- Ustalamy dolną krawędź $x_1 \le x_{min}$ histogramu oraz szerokość Δ_i przedziałów, czyli cały zakres wartości wielkości histogramowanej dzielimy na przedziały:
  od x_1 do $x_2 = x_1 + \Delta_1$, od x_2 do $x_3 = x_2 + \Delta_2$, od x_3 do $x_4 = x_3 + \Delta_3$ itd., aż do ostatniego, K-tego przedziału od wartości $x_K = x_{K-1} + \Delta_{K-1}$ do wartości
  $x_{K+1} = x_K + \Delta_K$, $x_{K+1} > x_{max}$.
- Następnie ustalamy, do którego przedziału należy każda z kolejnych wartości z próbki, otrzymując liczby n_i danych w każdym z przedziałów, zwane liczebnościami bądź krotnościami. W trakcie ustalania, do którego przedziału histogramowania należy włączyć daną wartość, możemy natknąć się na sytuację, w której wartość ta wypada na granicy przedziałów, a więc może zostać zaklasyfikowana zarówno do tego, w którym rozważana wartość stanowi górną granicę lub też do następnego przedziału obejmującego większe wartości zmiennej histogramowanej. Najczęściej przyjmujemy konwencję,
  w której przedział histogramowania jest z lewej strony domknięty zaś z prawej otwarty, jak to sugeruje opis w poprzednim punkcie.
- W następnym kroku dla każdego przedziału histogramu konstruujemy częstość $p_i := n_i/N$ oraz wielkość f_i, którą zwiemy gęstością wielkości histogramowanej, którą definiujemy jako $f_i := p_i/\Delta_i$, czyli stosunek częstości p_i do szerokości Δ_i przedziału histogramowania. W rezultacie otrzymujemy Tabelę 1.

Tabela1. Konstrukcja krotności, czestości i gęstości histogramowania

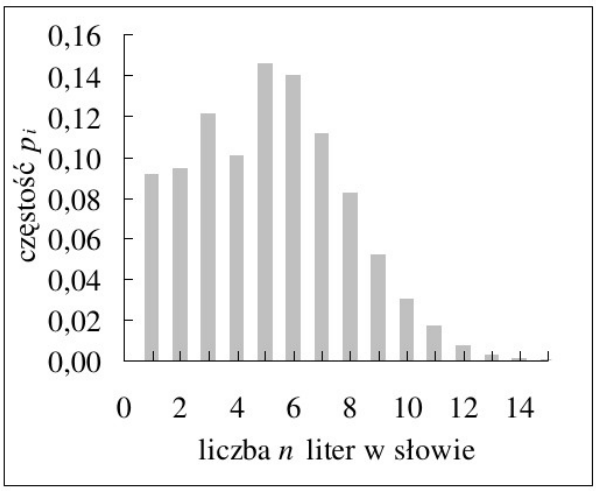
Przedział	$[x_1 , x_2 [$	$[x_2 , x_3 [$	….	$[x_K , x_{K+1} [$	suma
Krotność	n_1	n_2	….	n_K	$\sum_{i=1}^{i=K} n_i = N$
Częstość p_i	$\frac{n_1}{N}$	$\frac{n_2}{N}$	….	$\frac{n_K}{N}$	$\sum_{i=1}^{i=K} p_i = 1$
Gęstość f_i [s^{-1}]	$\frac{n_1}{N\Delta_1}$	$\frac{n_2}{N\Delta_2}$	….	$\frac{n_K}{N\Delta_K}$	$\sum_{i=1}^{i=K} f_i \Delta_i = 1$

Zauważ, że gęstości f_i mają wymiar – w przypadku histogramu zmierzonych wartości okresu wahadła wymiarem jest odwrotność jednostki czasu, w której wyrażamy wyniki pomiaru okresu. Spełniają one także oczywisty związek:

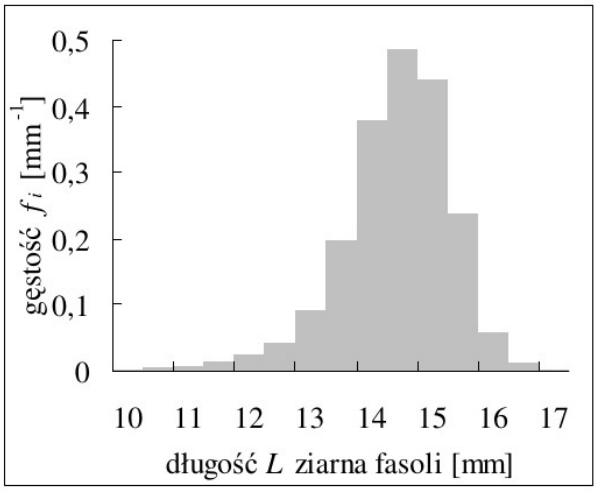
$$\sum_{i=1}^{i=K} f_i \Delta_i = 1,$$

czyli pola powierzchni słupków sumują się do jedności – co jest definicja określenia: **histogram jest unormowany do jedności**. Najczęściej szerokości Δ_i przedziałów histogramowania wybieramy takie same dla każdego z przedziałów, co upraszcza nieco obliczenia. Są jednak sytuacje, kiedy zmuszeni jesteśmy wybrać różne szerokości, a w skrajnym przypadku sięgające nieskończoności. Histogram rysujemy, kreśląc słupki, o wysokości proporcjonalnej do wartości krotności, częstości, czy gęstości w kolejnych przedziałach, czyli do wielkości histogramowanej

Zwróć uwagę na niektóre elementy graficzne takiego rysunku. Histogram winien mieć tytuł, osie należy opisać zarówno słownie, jak i symbolem, jak również podać, w nawiasach kwadratowych lub okrągłych, jednostkę wielkości występującej na osiach. Normy wymagają, aby znaczniki na osiach zwrócone były ku dodatnim kierunkom osi, co powoduje że w przypadku, gdy kreślone wielkości wypadają w pierwszej ćwiartce, znaczniki te „wchodzą" do rysunku, zaś gdy wykres mieści się w trzeciej ćwiartce, znaczniki będą wskazywać na zewnątrz rysunku.
W odniesieniu do wszystkich elementów graficznych prezentowanych w opracowaniach naukowych obowiązuje jeszcze jedna zasada: powinny być one „ascetyczne" w swym obliczu – wszelkie gradienty, tła, linie siatek, trzeci wymiar i tym podobne „dodatki" powinny się pojawiać jedynie wtedy, gdy wynikają z istoty prezentowanej wielkości lub też intencją autora jest zwrócenie uwagi czytelnika na dany aspekt.
Z histogramami związana jest konwencja, której należy bezwzględnie przestrzegać. Otóż, istnieją dwa typy wielkości, które histogramujemy. Rozważmy takie wielkości jak czas, masa, długość, temperatura, ciśnienie,... i skonfrontujmy je z takimi wielkościami, jak liczba rozpadających się jąder atomowych, liczba liter w słowie, liczba oczek na kostce do gier planszowych, liczba galaktyk w kącie bryłowym,... . Te pierwsze maja tę własność, że *a priori* mogą przyjmować dowolną wartość, także wyrażoną liczbą niewymierną (nie możemy wykluczyć, że kulka ma masę np. e^{π} g), podczas gdy te drugie opisywane są liczbami całkowitymi.
Te pierwsze nazywamy **wielkościami ciągłymi**, zaś o tych drugich mówimy, że są **wielkościami dyskretnymi**. To rozróżnienie znajduje swe odbicie na histogramie – słupki histogramu wielkości ciągłej zawsze rysujemy połączone ze sobą (nawet, jeśli histogram przedstawia częstości, a nawet krotności), a słupki histogramu wielkości dyskretnej rysujemy jako rozdzielone. Ilustrują to dwa rysunki. Rys. 1 przedstawia częstość wielkości dyskretnej – liczby i liter w słowach poematu *Pan Tadeusz* (tekst poematu: Polska Biblioteka Internetowa, http://www.pbi.edu.pl/) zaś Rys. 2 ukazuje gęstość wielkości ciągłej – długości *L* liczby 9439 ziaren fasoli (S.J. Pretorius, *Biometrika*, **22**, (1930), 110; dane za: M.G. Kendall i A. Stuart, The Advanced Theory of Statistics, *Charles Gryffin & Co. Ltd.*, London, 1958) – zwróć uwagę, że słowo gęstość oznacza tu ułamek liczby ziaren fasoli trafiających do wybranego przedziału, odniesiony do długości przedziału, a nie gęstość masy ziarna fasoli.



Rys. 1. Rozkład długości *n* słowa w poemacie *Pan Tadeusz*



Rys. 2. Rozkład długości *L* ziarna fasoli

O histogramach mówimy, że przedstawiają rozkład wielkości histogramowanej, np. rozkład wartości zmierzonego okresu drgań. Termin ten również stosujemy, gdy ilustrujemy częstości, a nawet wtedy, gdy na histogramie ukazujemy jedynie liczby danych (krotności n_i) w klasach.