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 $x_{\{1\}} < x_{\min}$ 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 $x_{\{K+1\}} = x_{\{K\}} + \Delta_K$, $x_{\{K+1\}} \geq x_{\max}$.
- 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 otwarty zaś z prawej domknięty, 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** (w tym przypadku: *gęstością okresu drgań wahadła*), a którą definiujemy jako $f_i := p_i/\Delta_i$, czyli stosunek częstości p_i do szerokości Δ_i przedziału histogramowania. W rezultacie otrzymujemy kolejne wiersze tabeli poniżej.

przedział	$(x_{\{1\}}, x_{\{2\}}]$	$(x_{\{2\}}, x_{\{3\}}]$	...	$(x_{\{K\}}, x_{\{K+1\}}]$	suma
krotność	n_1	n_2	...	n_K	N
częstość p_i	$\frac{n_1}{N}$	$\frac{n_2}{N}$	...	$\frac{n_K}{N}$	1
gęstość f_i [s^{-1}]	$\frac{n_1}{N\Delta_1}$	$\frac{n_2}{N\Delta_2}$	...	$\frac{n_K}{N\Delta_K}$	–

Zauważ, że wielkości f_i mają wymiar – w tym przypadku jest to odwrotność jednostki czasu, w której wyrażamy wyniki pomiaru okresu. Spełniają one także oczywisty związek

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

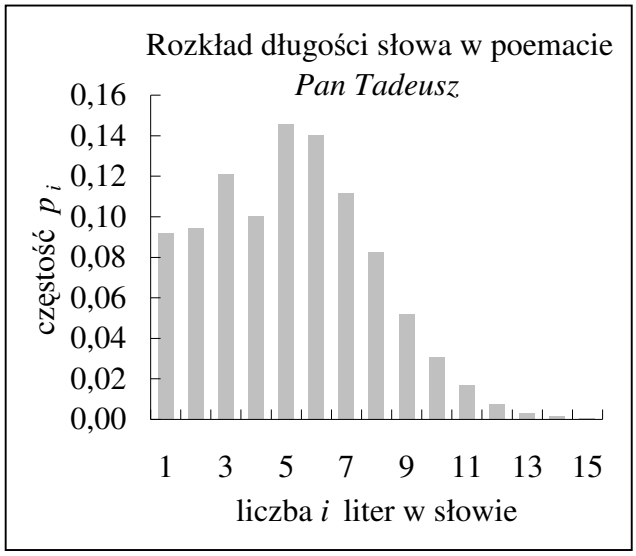
czyli pola powierzchni słupków histogramu sumują się do jedność – co jest definicją frazy: **histogram jest unormowany do jedności**. Najczęściej szerokości Δ_i przedziałów histogramowania wybieramy takie same dla każdego z przedziałów, co uprasza nieco obliczenia. Są jednak sytuacje (przykład poznamy w jednym z następnych ćwiczeń), kiedy to zmuszeni jesteśmy wybrać je różnymi (a w skrajnym przypadku sięgającymi nieskończoności).

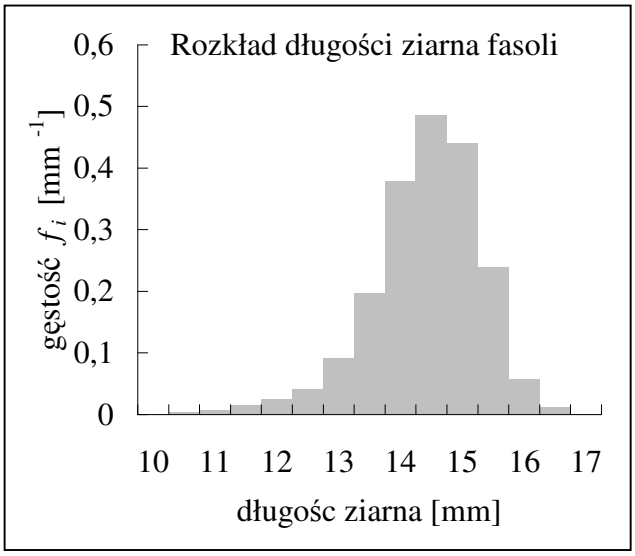
Histogram rysujemy, kreśląc słupki, o wysokości proporcjonalnej do wartości gęstości, na kolejnych przedziałach zaznaczonych na osi odciętych, czyli wielkości histogramowanej.

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 prezentując wykres, który mieści się w trzeciej ćwiartce, znaczniki będą wskazywać na zewnątrz treści 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 takimi wielkościami, jak liczba rozpadających się jąder atomowych w zadanym przedziale czasowym, długość słowa czyli liczba liter w słowie, liczba oczek na kostce do gier planszowych, liczba galaktyk w wybranym 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 opisują się liczbą całkowitą bądź zerem. Te pierwsze nazywamy wielkościami **ciągłymi**, zaś o tych drugich mówimy, że przedstawiają sobą wielkości **dyskretne**. 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 rozdzielone. Ilustrują to dwa rysunki poniżej. Lewy przedstawia częstość wielkości dyskretnej – liczby liter w słowach *Pana Tadeusza* (tekst poematu: Polska Biblioteka Internetowa, http://www.pbi.edu.pl/) zaś prawy ukazuje gęstość wielkości ciągłej – długości 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* oznacz tu liczbę ziaren fasoli na przedział długości, a nie gęstość masy ziarna fasoli).





O histogramach często mówimy, że przedstawiają sobą **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.