

# **Eksploracja danych**

## **Sprawozdanie nr 1 z ćwiczeń laboratoryjnych**

Tomasz Szkaradek  
Mateusz Niepokój  
Mariusz Biegański

Wydział Fizyki i Informatyki Stosowanej

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

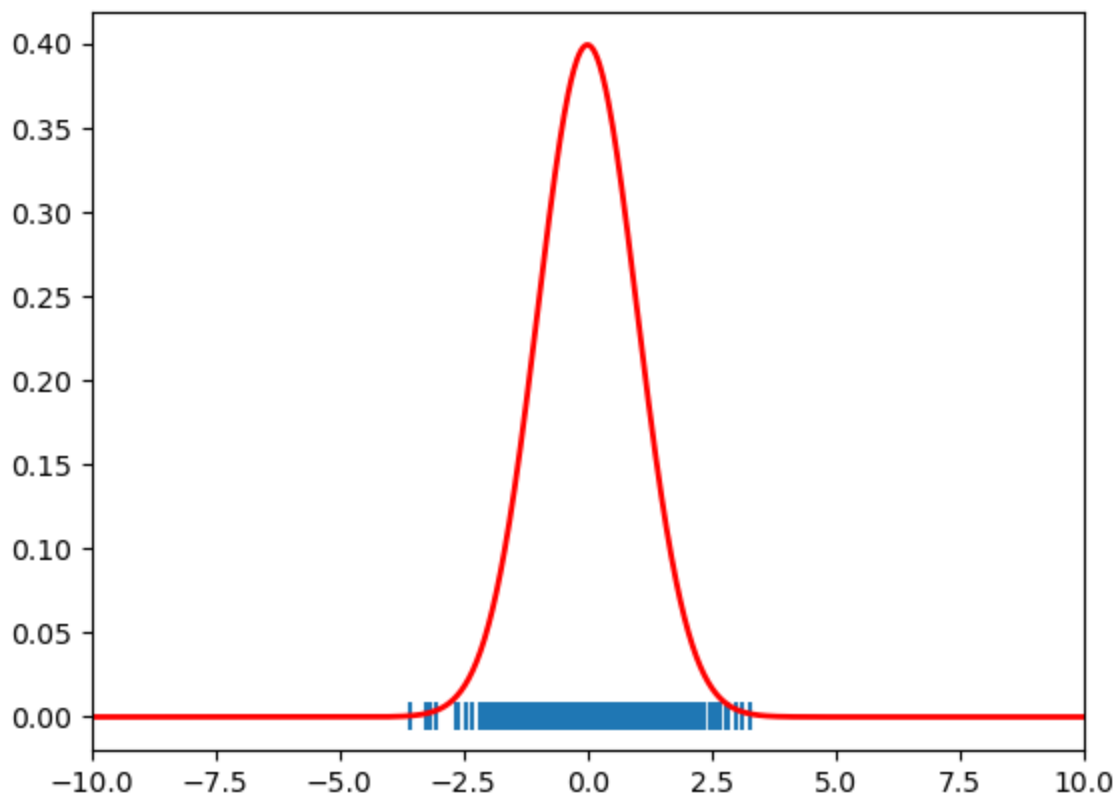
13 listopad 2023

### **Wstęp**

Podczas pierwszych zajęć laboratoryjnych skupiliśmy się na eksploracji oraz modyfikacji rozkładu normalnego.

### **Zadanie 1**

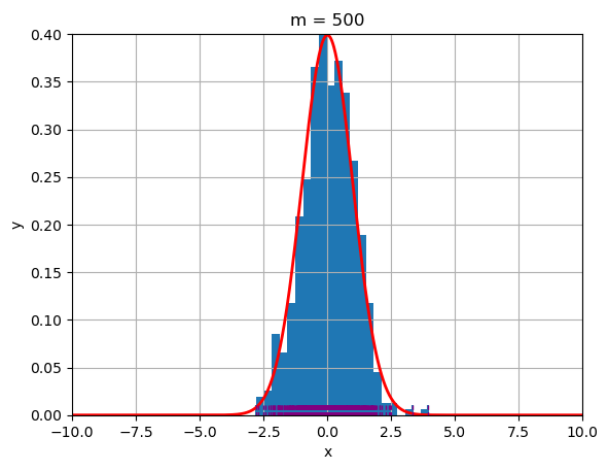
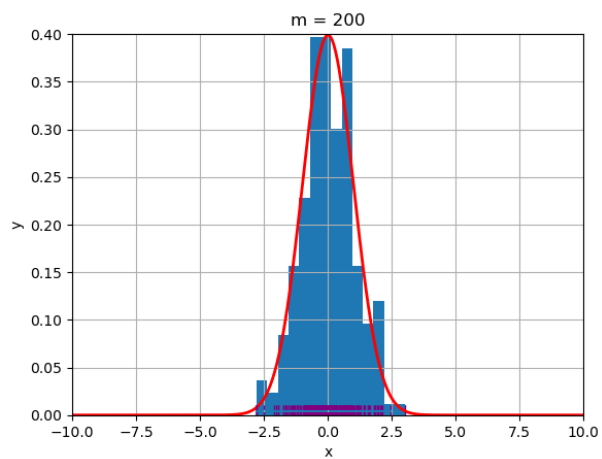
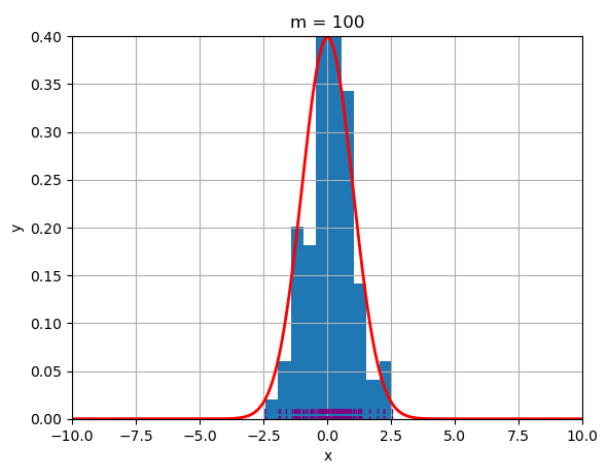
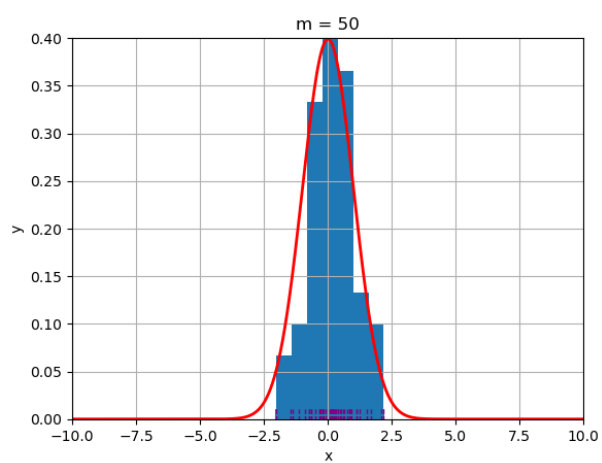
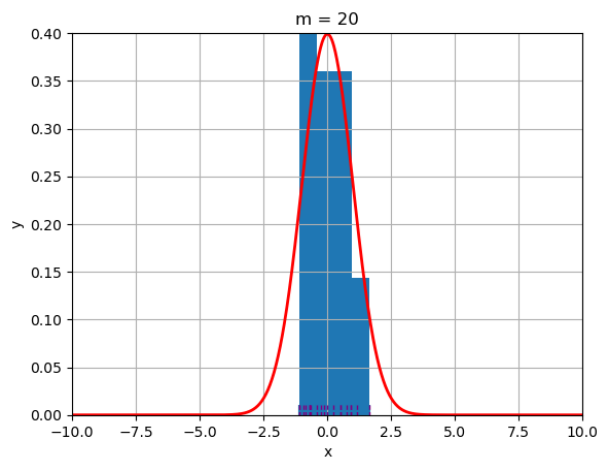
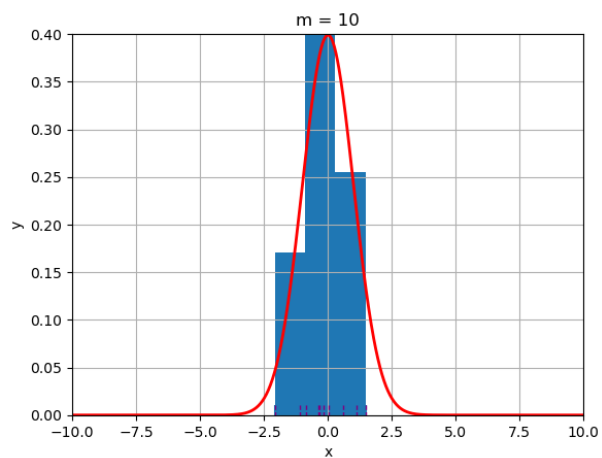
Na początku ćwiczeń stworzyliśmy wykres rozkładu normalnego Gaussa, korzystając z języka Python i biblioteki numerycznej numpy. Do wizualizacji użyliśmy również biblioteki matplotlib. Ustaliliśmy wartość ziarna wykorzystywanego w generatorze liczb pseudolosowych na `seed=1`.

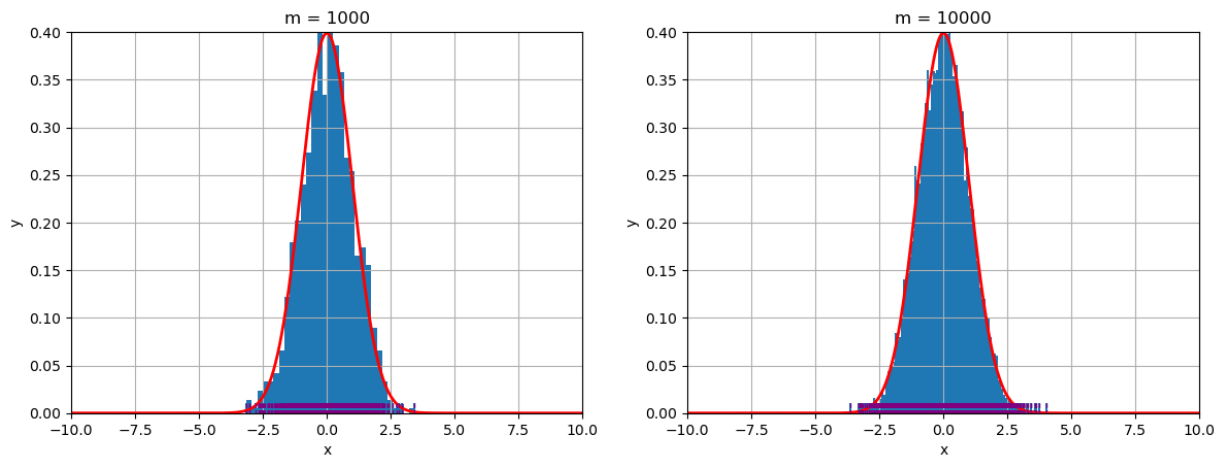


*Rys. 1: Wykres punktowy, wraz z funkcją ilustrującą rozkład punktów dla rozkładu normalnego Gaussa(0, 1) dla 1000 elementów.*

## **Zadanie 2**

W kolejnym kroku na podstawie wygenerowanych próbek o różnej liczności  $m=[10, 20, 50, 100, 200, 500, 1000, 10000]$  z rozkładu normalnego  $N(0, 1)$  wygenerowano kolejne histogramy:



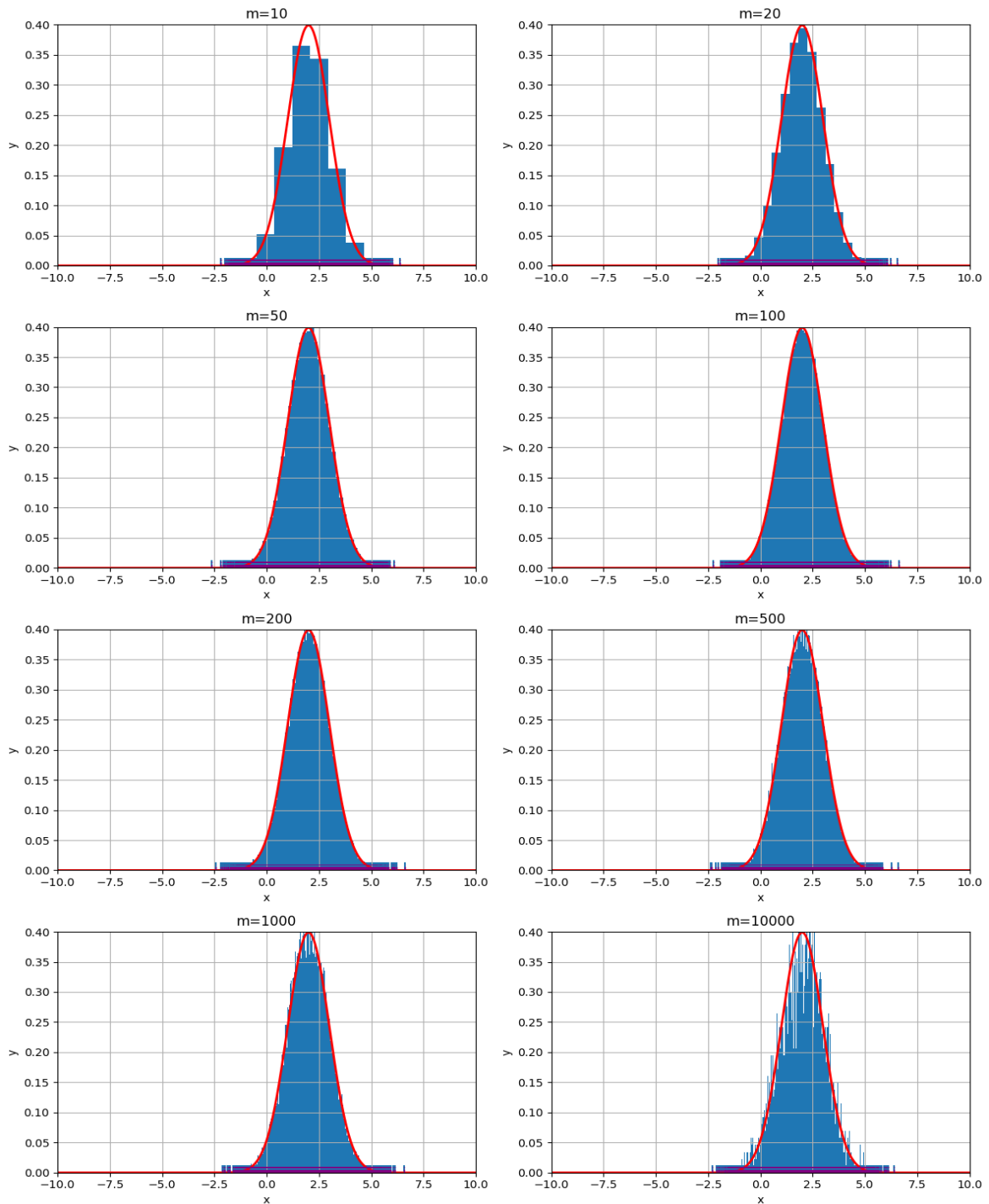


*Rys. 2: Losowe próbki z rozkładu normalnego o różnej liczności.*

Wraz z coraz większą ilością próbek na histogramie dostrzeżono coraz lepsze dopasowanie wyników do rozkładu Gaussa. Zgodnie z oczekiwaniami najliczniejsze wartości znajdują się w punkcie 0 oraz wartości nie przekraczają progu 0,5.

### Zadanie 3

W tym zadaniu zamiast rozkładu  $N(0,1)$  użyto rozkładu  $N(2,1)$  i narysowano wykres.

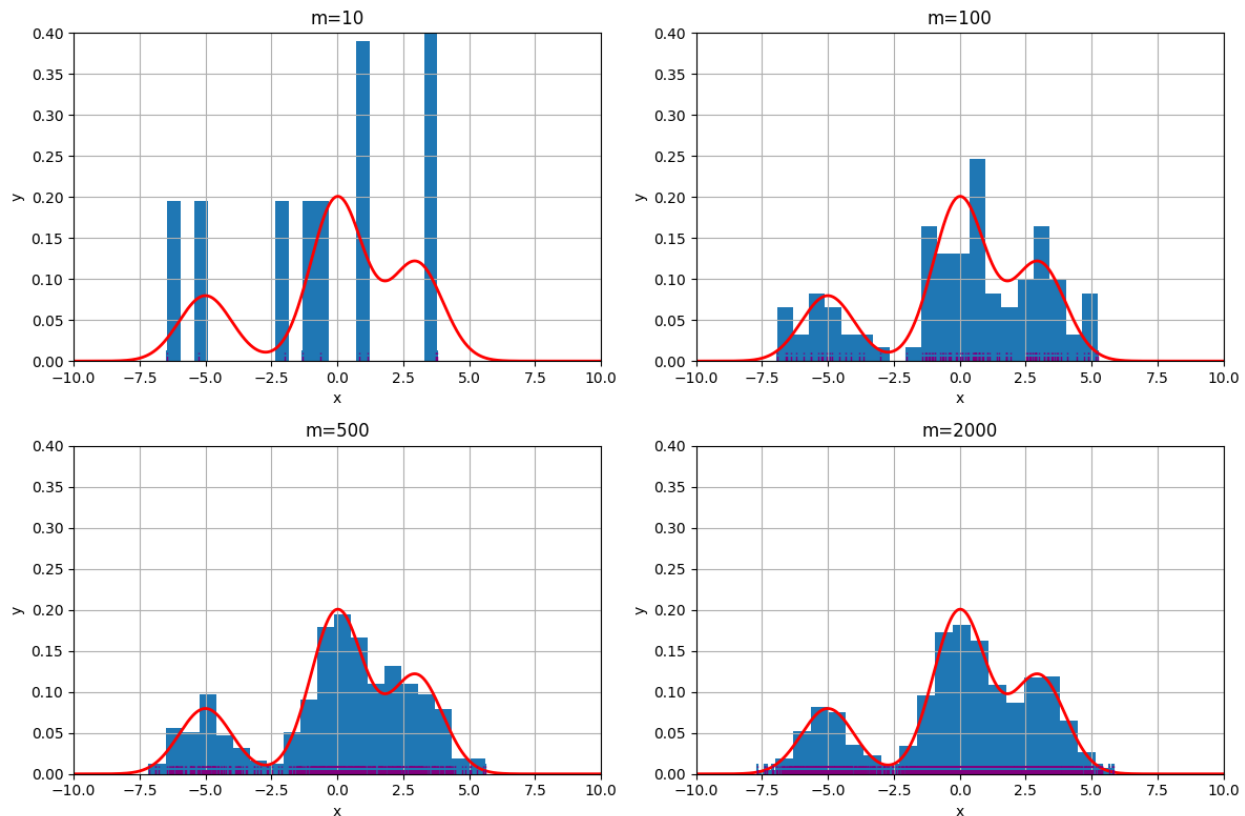


Rys. 3: Wykres wraz z histogramem rozkładu gaussa  $N(2,1)$  dla 50 elementów.

Skutkiem takiego działania jest przesunięcie wierzchołka wykresu o dwa punkty w stronę dodatniej osi X.

#### Zadanie 4

Zadanie polegało na wygładzanie wykresu, którego dane stanowią sumę trzech zbiorów liczbowych z rozkładu normalnego zgodnie z: 50% (1, 0) oraz 30% (3, 1) oraz 20% z (-5, 1).



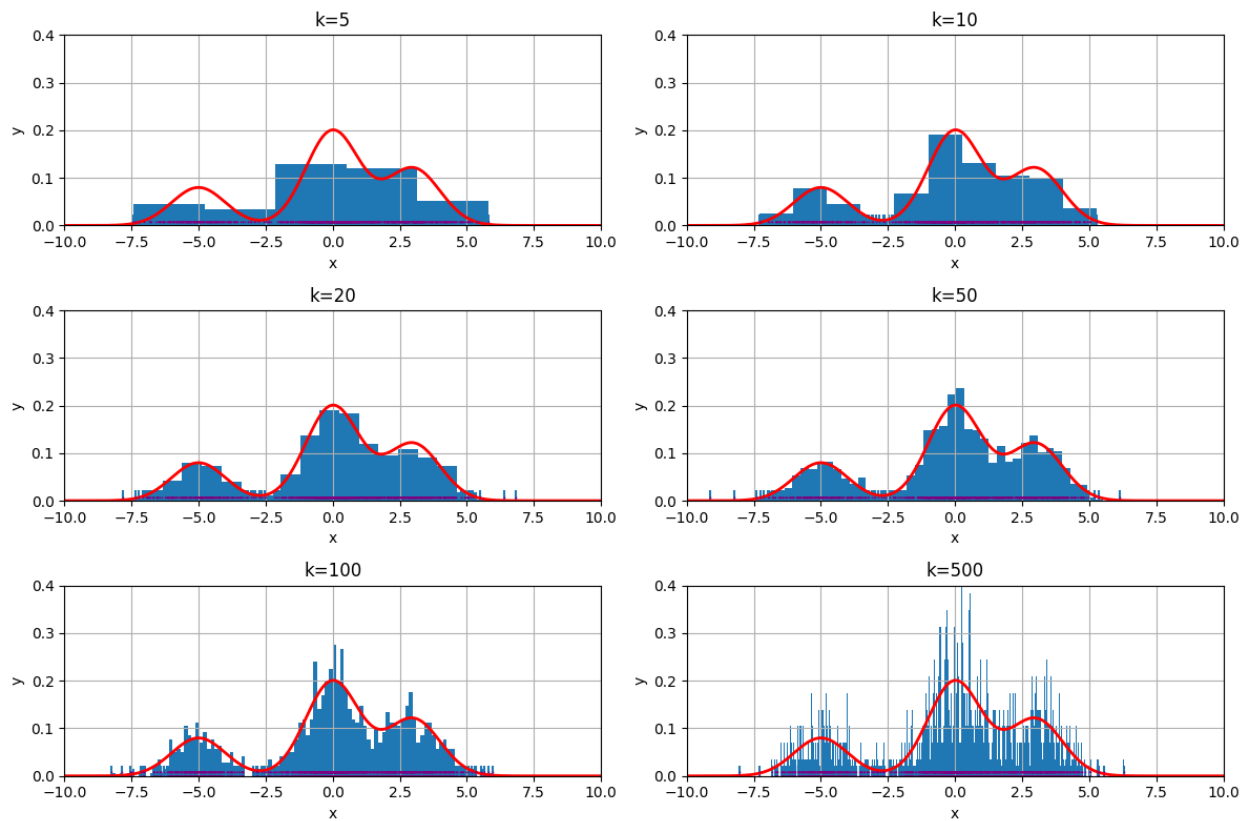
*Rys. 4: Wykresy będące sumą trzech zbiorów rozkładów normalnych.*

Z wykresu można odczytać, że w miarę wzrostu liczby elementów, kolejne przedziały histogramu coraz bardziej zbliżają się do wyestymowanej funkcji rozkładu. Histogram zdaje się dobrze odzwierciedlać analizowany zbiór danych, zwłaszcza dla 500 elementów.

#### Zadanie 5

Ostatnim zadaniem było porównanie różnych wartości  $k$  (ilość słupków histogramu) dla danych z zadania 4 i liczbie elementów równej 1000.

,



*Rys. 5: Wykresy będące sumą trzech zbiorów rozkładów normalnych.*

Na podstawie Rys. 5 doszło do wniosku, że  $k=[20, 50]$  najlepiej oddaje kształt wzorcowemu rozkładowi zaznaczonemu na wykresie linią czerwoną. Dla zbyt małych wartości histogram jest niedopełniony i pozostawia puste pola pod krzywą, z kolei dla zbyt dużych wartości pojawiają się przebiecia, które zakłócają wygenerowane dane.