Metody matematyczne w transporcie

Zmienna losowa
Miary statystyczne i ich interpretacja

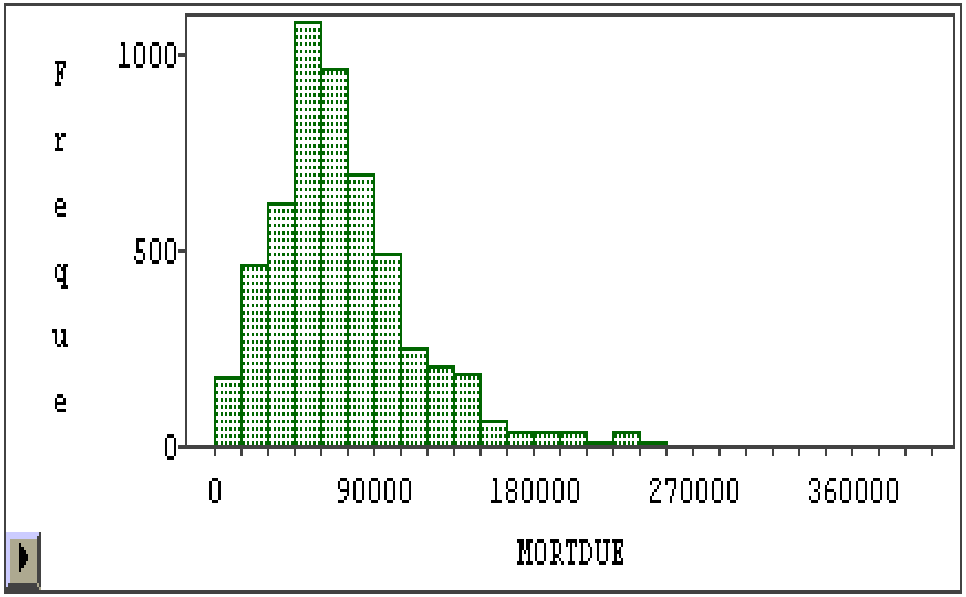
Marzena Nowakowska
Katedra Technologii Informatycznych
Wydział Zarządzania i Modelowania Komputerowego

Zmienna losowa

Zmienna losowa jest funkcją, która zdarzeniu elementarnemu przyporządkowuje pewną wartość. Wartość ta jest wielkością ilościową lub jakościową znaną po przeprowadzeniu badań lub doświadczenia, nie dającą się przewidzieć wcześniej.

Podstawowe pojęcia związane ze zmienną losową

- zmienna losowa dyskretna: ilościowa i jakościowa
- zmienna losowa ciągła
- rozkład zmiennej losowej
- realizacja zmiennej losowej
- szereg rozdzielczy
- histogram



Klasyfikacja zmiennych losowych

Zmienna losowa typu skokowego (dyskretna)

może przyjmować tylko skończoną lub co najwyżej przeliczalną liczbę wartości. Mogą to być wartości naturalne, całkowite lub jakościowe.

Zmienna losowa typu ciągłego (ciągła)

może przybierać wartości określone na zbiorze nieprzeliczalnym (najczęściej dowolna liczba rzeczywista z określonego przedziału na osi liczbowej).

Rozkład i dystrybuanta zmiennej losowej skokowej

Zmienna losowa typu skokowego

Rozkładem skokowej zmiennej losowej X nazywa się prawdopodobieństwo tego, że zmienna X przyjmie wartość x_i:

$P(X = x_i) = p_i$, $i = 1, 2, ...$

Funkcja $F(x_i) = P(X <= x_i)$ nazywa się dystrybuantą zmiennej losowej X, $F(x_i) = p_1 + p_2 + ... + p_i$.

Najpopularniejsze rozkłady zmiennej losowej skokowej:
zero-jedynkowy, dwumianowy, hipergeometryczny, Poissona

Dystrybuanta i gęstość prawdopodobieństwa zmiennej losowej ciągłej

Zmienna losowa typu ciągłego

Funkcja F(x) = P(X <= x) nazywa się dystrybuantą zmiennej losowej X, gdzie F(x) jest prawdopodobieństwem, więc $0 \leq F(x) \leq 1$.

Dla zmiennej losowej ciągłej zachodzi: P(X = stała) = 0.

Jeżeli dystrybuanta F(x) ma pochodną w punkcie x, to pochodna ta nazywa się gęstością prawdopodobieństwa f(x) zmiennej losowej X w punkcie x. Zachodzą zależności:

$$f(x) = F'(x) \qquad F(x) = \int_{-\infty}^{x} f(t)dt$$

Najpopularniejsze rozkłady zmiennej losowej ciągłej: prostokątny, trójkątny, **normalny**, lognormalny, gamma, beta

Histogram - definicje

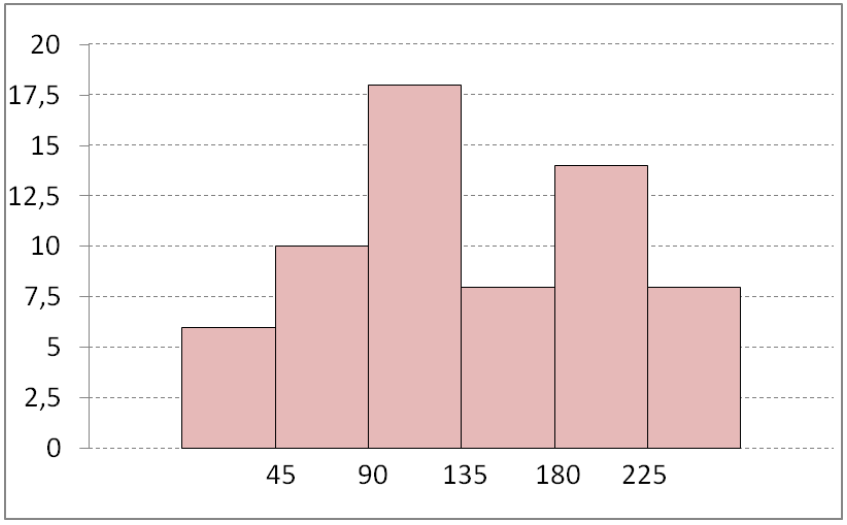
Wartość, jaką zmienna losowa przyjmie w wyniku badań lub doświadczenia nosi nazwę **realizacji zmiennej losowej**. Zbiór realizacji zmiennych losowych stanowi **zbiorowość statystyczną**, która może podlegać badaniu. W przypadku dużej liczebności zbiorowości statystycznej szereg statystyczny zestawia się w szereg rozdzielczy.

Szereg rozdzielczy powstaje poprzez podzielenie obszaru zmienności cechy (zmiennej losowej) na rozłączne przedziały (tzw. przedziały klasowe), zazwyczaj o tej samej długości, a następnie wyznaczenie liczby wystąpień wartości elementów analizowanego szeregu statystycznego w kolejnych przedziałach.

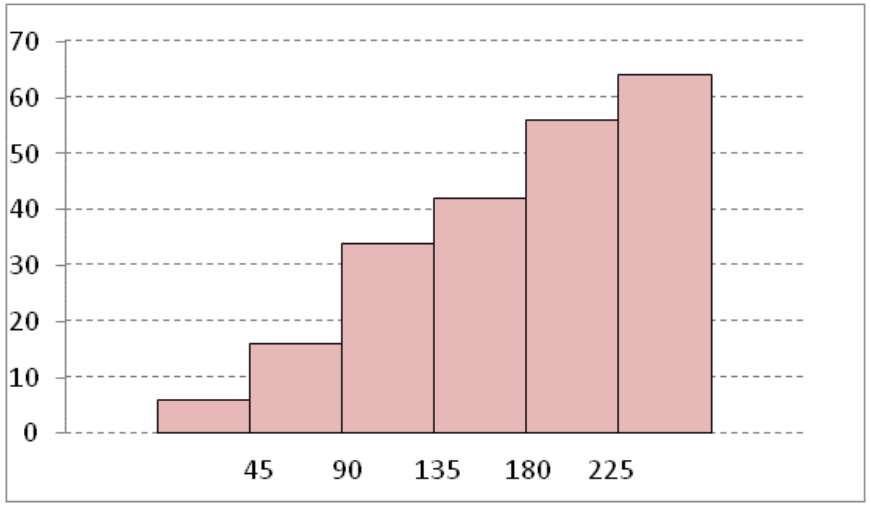
Histogram jest graficzną ilustracją szeregu rozdzielczego w postaci słupków liczebności (lub częstości względnej) względem wyróżnionych przedziałów.

Histogram skumulowany jest ilustracją szeregu rozdzielczego o liczebnościach (częstościach) skumulowanych.

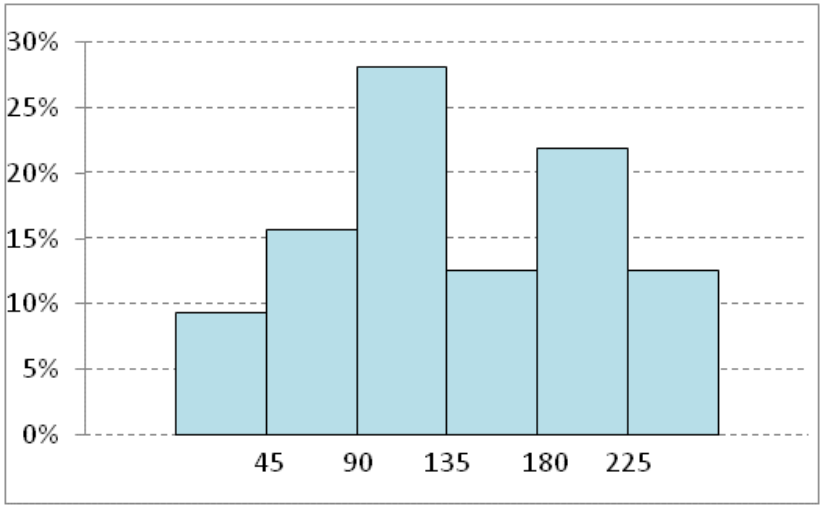
Histogram zmiennej losowej



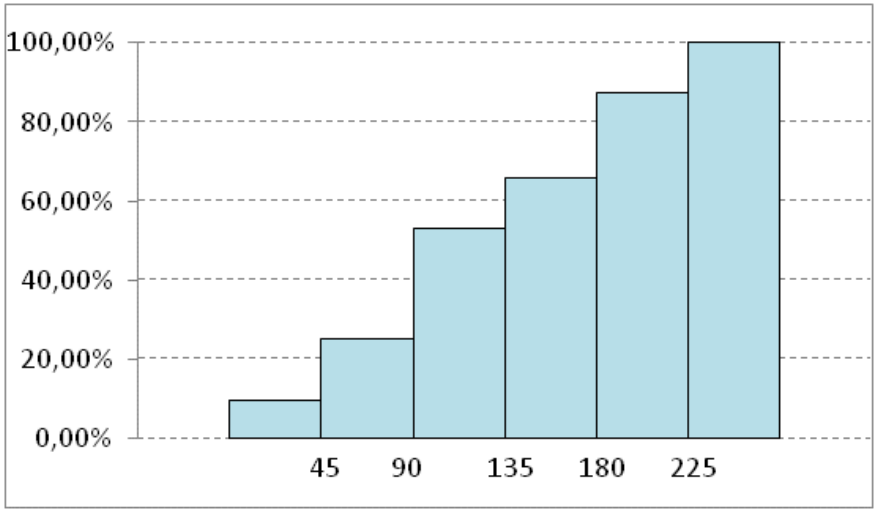
Histogram częstości (frequency)



Histogram częstości skumulowanej



Histogram gęstości (density)



Histogram gęstości skumulowanej

Miary statystyczne

W zastosowaniach praktycznych rozpatruje się funkcję rozkładu prawdopodobieństwa czy dystrybuantę zmiennej losowej, oraz dodatkowo parametry opisujące zasadnicze cechy rozkładu.

Za pomocą miar położenia (pozycyjne), rozproszenia i kształtu rozkładu (pozostałe) można w sposób syntetyczny opisać zbiór danych.

Miary pozycyjne

Minimum – **Min**

Maksimum – **Max**

Modalna Mo (dominanta) – **Mode**

Mediana Me (wartość środkowa) – **Median**

Średnia Śr (wartość oczekiwana) – **Mean**

Kwantyl rzędu p $\in$ [0, 1]

Kwartyle rzędu:

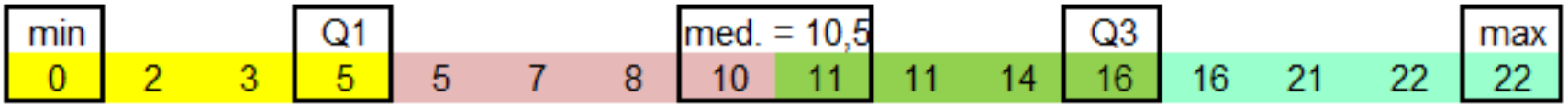
0 = minimum (**Q0)**, 2 = mediana (**Q2)**, 4 = maksimum (**Q4)**

Kwartyl 1 K1 – **Q1**

Kwartyl 3 K3 – **Q3**

Znaczenie niektórych z ww. miar jest uzależnione od uporządkowania niemalejąco zbioru danych.

Wszystkie miary pozycyjne mają miano cechy



Miary rozproszenia

Rozstęp R – **Range**

miano cechy

Różnica między wartością największą i najmniejszą

Rozstęp międzykwartylowy IQR – **InterQuartile Range**

miano cechy

Różnica między kwartylem 3 i kwartylem 1

Odchylenie standardowe OS – **Std Deviation**

miano cechy

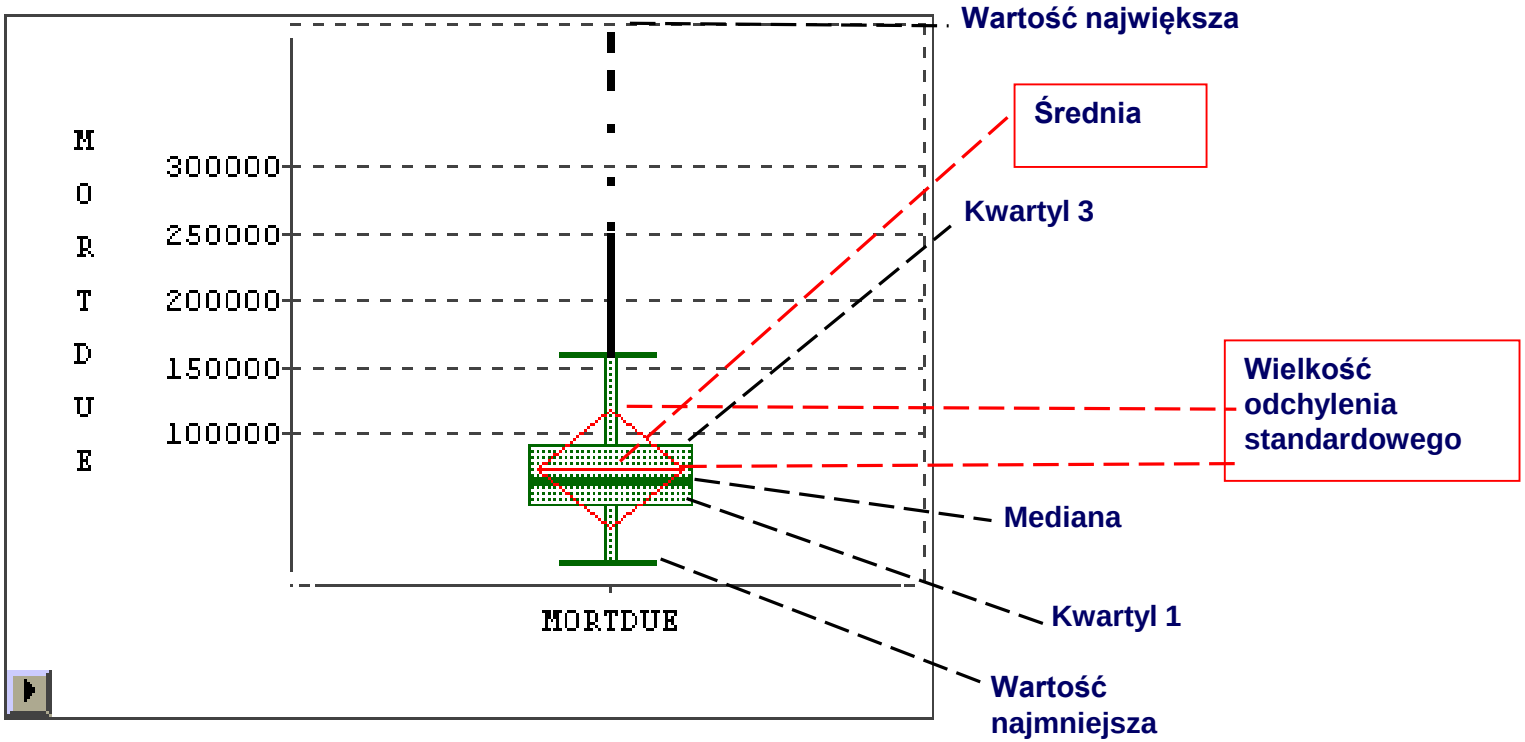
Pierwiastek kwadratowy z wariancji

Współczynnik zmienności WZ – **Coeff Variation**

brak miana

Iloraz odchylenia standardowego i średniej; wyrażany w %

Wykres pudełkowy z wąsami



Q3-Q1 = IQR (InterQuartile Range)
Długość wąsa pudełka ≤ 1.5 * IQR

Miary pozostałe – skośność

Współczynnik asymetrii WA (skośność) – **Skewness**

asymetria prawostronna dla WA > 0:

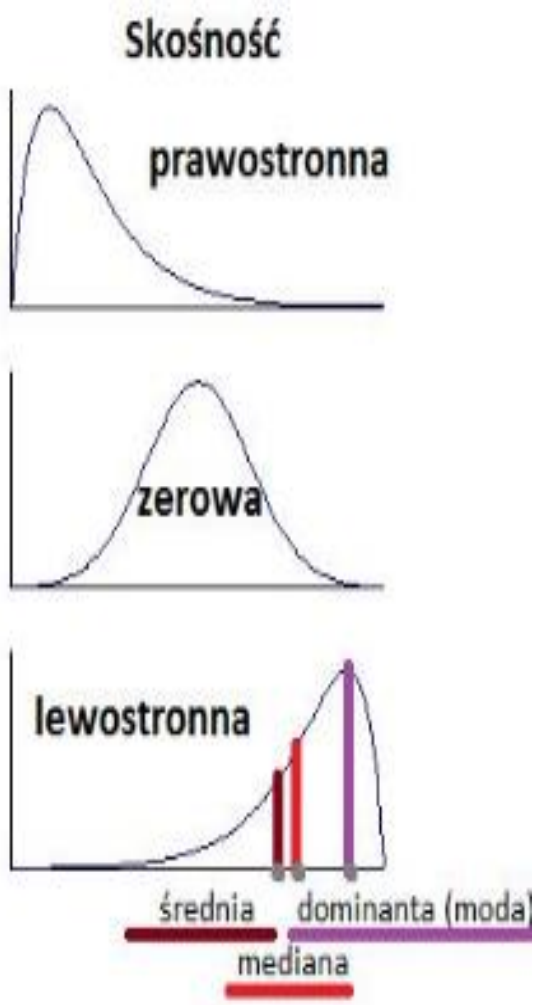
Mo < Me < Śr

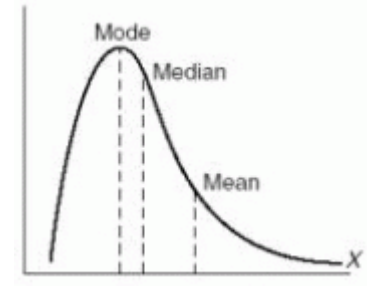
rozkład cechy symetryczny dla WA = 0

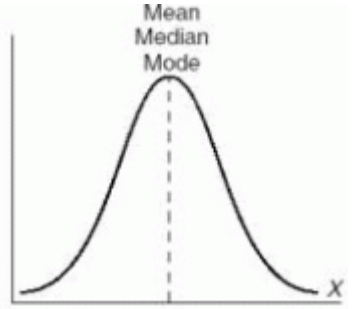
Mo = Me = Śr

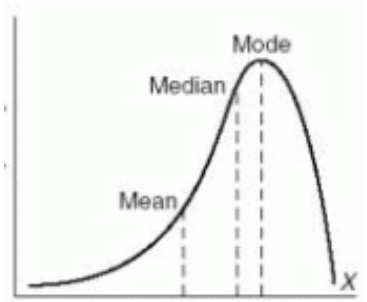
asymetria lewostronna dla WA < 0

Mo > Me > Śr





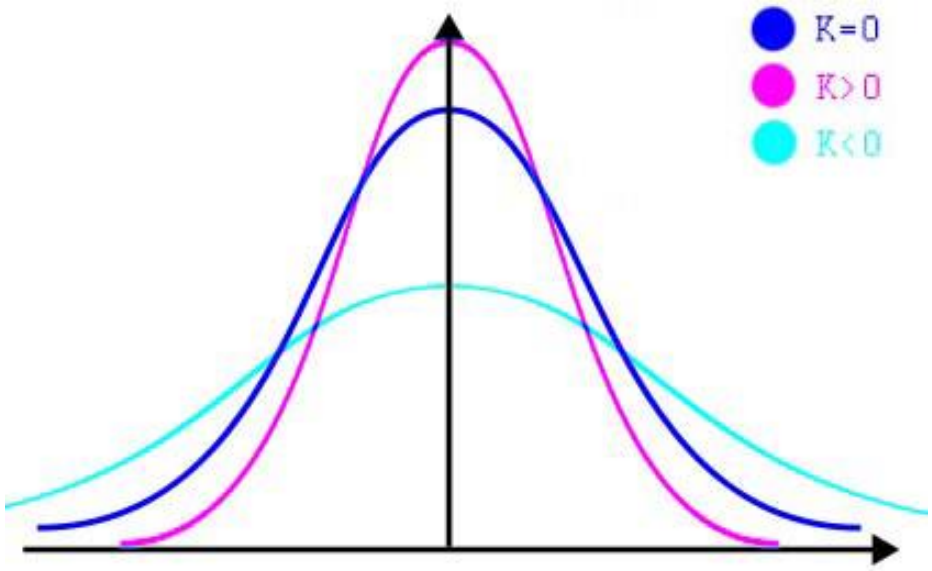




Miary pozostałe – kurtoza

Współczynnik spłaszczenia WS (kurtoza, eksces) – **Kurtosis**

- rozkład zbliżony do normalnego WS = 0 (K = 0)
- rozkład wysmukły dla WS > 0 (K > 0)
- rozkład płaski dla WS <0 (K < 0)



Środowisko systemu SAS

Editor **–** obszar do definiowania programów z wykorzystaniem języka programowania SAS'a zwanego 4GL, języka makr oraz zestawu dostępnych procedur systemowych

Log **–** „dziennik pokładowy", do którego są wyprowadzane komunikaty systemu o wykonanych operacjach oraz ostrzeżenia o błędach, jeśli takie się pojawią

Output **–** okno wyników programu SAS'owego

Explorer **–** centralne okno dostępu do zasobów systemu: katalogów, tabel (czyli plików z danymi zapisanych w formacie SAS'a), bibliotek

Results **–** okno do zarządzania wynikami przetwarzania danych

Kontekstowe: menu główne i pasek narzędzi

Biblioteki programu SAS

Biblioteka (*Library*) – referencja identyfikująca folder na dysku. Foldery te zawierają zasoby tematyczne, najczęściej pliki SAS'owe.
Organizacja dostępu – poprzez operację skojarzenia w oknie *New Library*. Otwarcie okna realizuje się z paska narzędzi za pomocą przycisku *Add New Library* lub z aktywnego okna *Eksplorer* za pomocą menu *File*|*New/Library*

Biblioteki wbudowane: SASHELP, SASUSER, WORK

Zbiory danych programu SAS

Struktura taka jak struktura tabeli w bazie danych - stąd nazwa tabela
W języku analiz statystycznych każdy wiersz to obserwacja, a każda kolumna to zmienna lub cecha.

Dwa typy danych:
- tekstowy (litery, cyfry i znaki specjalne)
- liczbowy (cyfry, litera E w przypadku notacji naukowej, kropka jako separator części ułamkowej, znak minus przed liczbą).

Wartości liczbowe są przechowywane w postaci liczby zmiennoprzecinkowej.

Dana typu daty jest reprezentowana jako liczba dni od 1960-01-01.

Dana typu czasu jest reprezentowane jako liczba sekund jaka upłynęła od północy do podanego czasu.

Pliki są przechowywane w bibliotekach - dostęp referencyjny:
biblioteka.zbior, np. SASUSER.NKWDL

Tworzenie plików w środowisku SAS

W trybie bezpośrednim w oknie zbliżonym do arkusza kalkulacyjnego

Poprzez import z pliku innego formatu:

- Microsoft Excel
- Microsoft Access
- pliki formatu *dbf*
- zewnętrzne pliki z separatorami (deliminatorami); nie mają domyślnego rozszerzenia
- zewnętrzne pliki, w których separatorami wartości w kolejnych kolumnach są przecinki (*.*csv* – *Comma-Separated Values*) lub tabulatory (*.*tsv* – *Tab-Separated Values*)
- zewnętrzne pliki, w których separatorami wartości w kolejnych kolumnach są znaki tabulacji (*.txt)

Możliwość eksportu do plików innych formatów.