Elementy statystyki

STA - Wykład 1

dr hab. Waldemar Wołyński
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza

Programy do statystycznej analizy danych

Komercyjne:

a) **Statistica**
URL http://www.statsoft.com
URL http://www.statsoft.pl

b) **SAS**
URL http://www.sas.com

c) **SPSS**
URL http://www.spss.com
URL http://www.spss.pl

Niekomercyjne:

a) **R**
URL http://www.r-project.org

Literatura

1. T. Górecki, *Podstawy statystyki z przykładami w R*, BTC 2011.
2. Ł. Komsta, *Wprowadzenie do środowiska R* (http://www.r-project.org).
3. P. Biecek, *Przewodnik po pakiecie R*, GIS 2014.
4. M. Gągolewski, *Programowanie w języku R*, PWN 2014.
5. W.N. Venables, D. M. Smith and the R Development Core Team, *An Introduction to R* (http://www.r-project.org).
6. J. Verzani, *simpleR - Using R for Introductory Statistics* (http://www.r-project.org).

Program **R**

Program **R** jest zaawansowanym pakietem statystycznym i językiem programowania istniejącym na platformy Windows, Unix oraz MacOS. Objęty jest licencją **GNU GPL**.

Pierwsza wersja **R** (początek lat 90) została napisana przez Roberta Gentlemana i Ross Ihake pracujących na Wydziale Statystyki Uniwerstetu w Auckland. Obecnie rozwojem **R** kieruje fundacja "The R Foundation for Statistical Computing".

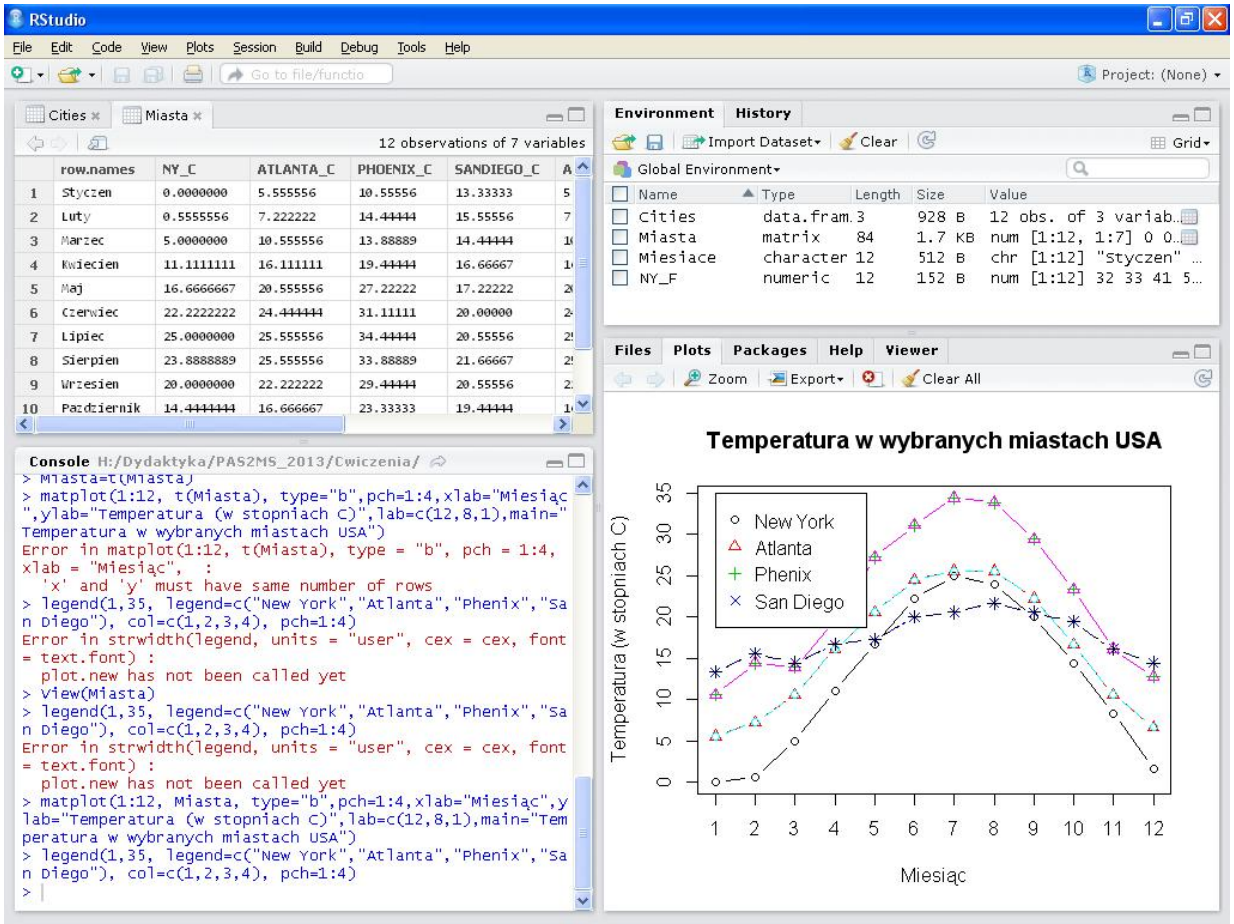
Język **R** był wzorowany na języku **S** opracowanym w AT&T Bell Laboratories i stosowanym w programie S-PLUS.

Język **R** jest językiem interpretowanym, a nie kompilowanym (kolejne komendy interpretowane są linia po linii lub wykonywane jako skrypt).

Największą siłą **R** jest około 20 000 bibliotek funkcji napisanych przez setki osób z całego świata, przeznaczonych do najróżniejszych zastosowań. Każda biblioteka dostarczana jest z pełną dokumentacją.

Program **RStudio**

Istnieje wiele programów (nakładek) ułatwiających prace z programem **R** np. Rcmdr, RKWard, RStudio.



Strona domowa: www.rstudio.com

Programowanie w **R**

Język programowania **R** opiera się na zmiennych i funkcjach. Zmienne nie muszą być deklarowane.

Uwagi:

- Klasycznym operatorem przypisania jest $<-$, można również wykorzystywać znak $=$.
- Jeżeli chcemy, aby wynik przypisania został wyświetlony na ekranie, należy przypisanie zamknąć w nawiasy $()$.
- Jeśli chcemy, aby kilka wyrażeń było zapisanych w jednej linii, to musimy oddzielić je średnikiem.
- Komentarz poprzedzamy znakiem hash $\#$, wszystko do końca linii jest już komentarzem.
- **R** odróżnia wielkie i małe litery.
- W celu określenia kolejności działań używamy nawiasów okrągłych.
- Do grupowania wyrażeń używamy nawiasów klamrowych.

Wektory

vector(...)

Podstawowe funkcje:

- **mode** - zwraca typ elementów wektora;
- **length** - zwraca długość wektora.

Elementy dodatkowe:

- **wek[3]** - odwołanie do trzeciego elementu wektora 'wek';
- **c()** - tworzenie wektora poprzez złączanie, np. c(1,3,6);
- **:** - generuje liczby z podanego przedziału, np. 1:4;
- **seq** - generuje liczby z podanego przedziału, przy czym można podać krok (by) lub długość (length), np. seq(0,3,by=0.5);
- **rep** - generuje ciąg składający się z powtórzeń innego ciągu, np. rep(1:3,2).

Ramki danych

data.frame(...)

Podstawowe funkcje:

- **nrow** - zwraca liczbę wierszy;
- **ncol** - zwraca liczbę kolumn;
- **rownames** - zmiana nazwy wiersza;
- **colnames** - zmiana nazwy kolumny.

Elementy dodatkowe:

- **ark[1,3]** - odwołanie do elementu w pierwszym wierszu i trzeciej kolumnie;
- **ark[,2]** - odwołanie do elementów drugiej kolumny;
- **ark$wiek** - odwołanie do elementów kolumny o nazwie ’wiek’ (zmiennej: ’wiek’);
- **ark[-3,]** - usunięcie trzeciego wiersza;
- **attach(ark)** - dołączenie do przestrzeni nazw wszystkich nazw kolumn ramki danych ’ark’.

Listy

list(...)

Podstawowe funkcje:

- **length** - zwraca liczbę elementów listy.

Elementy dodatkowe:

- **lista[[3]]** - odwołanie do trzeciego elementu listy;
- **lista$dane** - odwołanie do elementu listy o nazwie 'dane'.

Uwaga: Większość funkcji w **R** zwraca wynik w postaci listy.

Otwieranie/import danych

- **load(...)** - otwieranie danych zapisanych w formacie programu **R** (dla plików z rozszerzeniem 'RData');
- **read.table(...)** - import danych z plików tekstowych;
- **read.csv2(...)** - import danych z plików csv (np. Excel).

Uwaga: Do zapisu/exportu danych stosujemy odpowiednio funkcje: **save**, **write.table** oraz **write.csv2**.

Wykresy

Podstawowe typy wykresów:

- **plot** - wykres punktowy;
- **barplot** - wykres słupkowy;
- **hist** - histogram;
- **pie** - wykres kołowy;
- **boxplot** - wykres "pudełko z wąsami".

Popularne parametry wykresów:

- **main** - tytuł wykresu;
- **xlab, ylab** - tytuły osi;
- **lty** - typ linii;
- **lwd** - grubość linii;
- **col** - kolory punktów, linii, itp.

Funkcje

nazwa $<-$ **function**(argumenty) ciało

Instrukcje warunkowe:

if(warunek) wyrażenie1 **else** wyrażenie2

ifelse(warunek,a,b)

switch(zmienna, wartość1=akcja1, wartość2=akcja2, ...)

Pętle:

for(licznik **in** start:koniec) wyrażenie

while(warunek) wyrażenie

repeat wyrażenie

Rozkład empiryczny

Statystyka opisowa

Niech $\boldsymbol{x} = (x_1, \ldots, x_n)'$ będzie próbką, tzn. $x_1, \ldots, x_n$ są obserwacjami zmiennej (cechy) X.
Zadaniem **statystyki opisowej** jest prezentacja rozkładu cechy X w próbce (rozkładu empirycznego), przy pomocy tabeli lub wykresu. Często wystarczające jest jedynie podanie kilku liczb charakteryzujących ten rozkład.

Metody opisu rozkładu empirycznego:

2. Tabelaryczny
   R: *table* – szereg rozdzielczy (liczebności),
   prop.table – szereg rozdzielczy (proporcje, częstości),
   cut – dla cechy ilościowej ciągłej podział na przedziały klasowe.
3. Graficzny
   R: *barplot* – wykres słupkowy (cecha jakościowa lub ilościowa dyskretna),
   pie – wykres kołowy (cecha jakościowa lub ilościowa dyskretna),
   hist – histogram (cecha ilościowa ciągła).

3. Statystyki opisowe

- klasyczne – bazujące na uśrednianiu obserwowanych wartości w próbce, np. moment zwykły rzędu r:

$$\mu_r = \frac{1}{n}\sum_{k=1}^{n} x_k^r$$

- pozycyjne – bazujące na posortowanych rosnąco wartościach w próbce, np. dolny kwartyl:

$$Q_1 = \frac{1}{2}(x_{(i)} + x_{(j)}),$$

gdzie

$$i = \lceil \frac{n+1}{4} \rceil, \;\; j = \lceil \frac{n}{4} \rceil$$

lub górny kwartyl:

$$Q_3 = \frac{1}{2}(x_{(i)} + x_{(j)}),$$

gdzie

$$i = \lceil \frac{3(n+1)}{4} \rceil, \;\; j = \lceil \frac{3n}{4} \rceil.$$

Statystyki opisowe

Charakterystyki tendencji centralnej rozkładu empirycznego:

- **średnia**, **R:** *mean*

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

- **mediana**, **R:** *median*

$$Me = \begin{cases} x_{(\frac{n+1}{2})}, & n - \text{nieparzyste}, \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & n - \text{parzyste}. \end{cases}$$

Statystyki opisowe

Charakterystyki rozrzutu rozkładu empirycznego:

- **odchylenie standardowe**, **R:** *sd*

$$s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

- **współczynnik zmienności**

$$v = \frac{s}{\bar{x}}100$$