# Ćwiczenia ze statystyki

Jarosław Żygierewicz

10 listopada 2009

# Spis treści

# Rozdział 1

# Programowanie w Matlabie

## 1.1 Wstęp

Na naszych ćwiczeniach będziemy korzystać z Matlaba, gdyż można w nim łatwo i szybko implementować i testować algorytmy, w szczególności operujące na macierzach i wektorach.

## 1.2 Gdzie jest help?

Matlab ma rozbudowany system pomocy;

1. z command line

   ```
   help
   help funkcja_o_której_chcemy_sie_dowiedzieć
   ```

2. w menu Help

⇒ Proszę poszukać na różne sposoby informacji o funkcji `plot` Polecenie

```
help matlab/general
```

produkuje następujący wydruk (bardziej użyteczne komendy zostały **wytłuszczone**)

```
General purpose commands.
MATLAB Version 6.5 (R13) 20-Jun-2002

General information
  helpbrowser - Bring up the help browser.
  doc         - Complete on-line help, displayed in the help browser.
```

**help - M-file help, displayed at the command line.**

```
  helpwin     - M-file help, displayed in the help browser.
  lookfor     - Search all M-files for keyword.
  syntax      - Help on MATLAB command syntax.
  support     - Open MathWorks Technical Support Web Page.
  demo        - Run demonstrations.
```

**ver - MATLAB, SIMULINK, and toolbox version information.**

```
  version     - MATLAB version information.
  whatsnew    - Access Release Notes.

Managing the workspace.
```

**who** - List current variables.

**whos** - List current variables, long form.

```
  workspace   - Display Workspace Browser, a GUI for managing the workspace.
  pack        - Consolidate workspace memory.
```

**clear** - Clear variables and functions from memory.

**load** - Load workspace variables from disk

**save** - Save workspace variables to disk

**quit** - Quit MATLAB session.

```
Managing commands and functions.
  what        - List MATLAB-specific files in directory.
```

**type - List M-file.**

```
  edit        - Edit M-file.
  open        - Open files by extension.
  which       - Locate functions and files.
  pcode       - Create pre-parsed pseudo-code file (P-file).
  inmem       - List functions in memory.
  mex         - Compile MEX-function.

Managing the search path
  path        - Get/set search path.
  addpath     - Add directory to search path.
  rmpath      - Remove directory from search path.
  pathtool    - Modify search path.
  rehash      - Refresh function and file system caches.
  import      - Import Java packages into the current scope.

Controlling the command window.
  echo        - Echo commands in M-files.
  more        - Control paged output in command window.
  diary       - Save text of MATLAB session.
  format      - Set output format.
  beep        - Produce beep sound.

Operating system commands
```

**cd - Change current working directory.**

```
  copyfile    - Copy a file or directory.
  movefile    - Move a file or directory.
  delete      - Delete file.
  pwd         - Show (print) current working directory.
  dir         - List directory.
  fileattrib  - Get or set attributes of files and directories.
  isdir       - True if argument is a directory.
  mkdir       - Make directory.
  rmdir       - Remove directory.
  getenv      - Get environment variable.
  !           - Execute operating system command (see PUNCT).
  dos         - Execute DOS command and return result.
```

```
  unix         - Execute UNIX command and return result.
  system       - Execute system command and return result.
  perl         - Execute Perl command and return result.
  web          - Open Web browser on site or files.
  computer     - Computer type.
  isunix       - True for the UNIX version of MATLAB.
  ispc         - True for the PC (Windows) version of MATLAB.

Debugging M-files.
  debug        - List debugging commands.
  dbstop       - Set breakpoint.
  dbclear      - Remove breakpoint.
  dbcont       - Continue execution.
  dbdown       - Change local workspace context.
  dbstack      - Display function call stack.
  dbstatus     - List all breakpoints.
  dbstep       - Execute one or more lines.
  dbtype       - List M-file with line numbers.
  dbup         - Change local workspace context.
  dbquit       - Quit debug mode.
  dbmex        - Debug MEX-files (UNIX only).

Profiling M-files.
  profile      - Profile function execution time.
  profreport   - Generate profile report.

Tools to locate dependent functions of an M-file.
  depfun       - Locate dependent functions of an m-file.
  depdir       - Locate dependent directories of an m-file.
  inmem        - List functions in memory.
```

## 1.3 Zmienne - wektory, macierze

W matlabie wygodnie myśleć jest o zmiennych jako o macierzach: skalar to macierz $1 \times 1$ wektor to macierz $1 \times N$ lub $N \times 1$. Precyzja zmiennych jest domyślna (double) jeżeli nie każemy inaczej. Napiszcie w command line:

```
a=1;
```

`whos`

Spis podstawowych operacji na macierzach otrzymamy wpisując

`help matlab/elmat`

Najczęściej przeze mnie używane to:

**zeros** - produkuje macierz wypełnioną zerami.

**ones** - produkuje macierz wypełnioną jedynkami.

**eye** - macierz jednostkowa.

**repmat** - tworzy macierz złożoną z kopii podanej macierzy .

**rand** - macierz wypełniona liczbami z rozkładu płaskiego (0,1).

**randn** -macierz wypełniona liczbami z rozkładu normalnego o średniej 0 i wariancji 1.

```
    linspace    - Linearly spaced vector.
    logspace    - Logarithmically spaced vector.
    meshgrid    - X and Y arrays for 3-D plots.
```

```
  Basic array information.
```

**size** - Size of array.

**length** - Length of vector.

```
    ndims       - Number of dimensions.

    disp        - Display matrix or text.
    isempty     - True for empty array.

  Matrix manipulation.
    cat         - Concatenate arrays.
    reshape     - Change size.
```

```
    diag        - Diagonal matrices and diagonals of matrix.

    fliplr      - Flip matrix in left/right direction.
    flipud      - Flip matrix in up/down direction.
    flipdim     - Flip matrix along specified dimension.
    rot90       - Rotate matrix 90 degrees.
```

**:** - operator zasięgu (służy do robienie wektorów z równo odległymi elementami lub indeksowania fragmentów macierzy)

**find** - znajduje indeksy niezerowych elementów

**end** - indeks ostatniego elementu.

```
  Special variables and constants.
    ans         - Most recent answer.
    eps         - Floating point relative accuracy.
    pi          - 3.1415926535897....
    i, j        - Imaginary unit.
```

Macierze możemy wpisywać "z palca":

```
>> A=[1 2 3 4; 5 6 7 8; 8 9 1 2];
>> A

A =

     1     2     3     4
     5     6     7     8
     8     9     1     2

>> disp(A)
     1     2     3     4
     5     6     7     8
     8     9     1     2
```

Wczytywać je z plików lub uzyskiwać w wyniku działania funkcji.
Większość operacji działa na macierzach w sposób intuicyjny:
Transpozycja:

```
>> A'

ans =

     1     5     8
     2     6     9
     3     7     1
     4     8     2
```

Sumowanie

```
>> sum(A)

ans =

    14    17    11    14


>> sum(A')

ans =

    10    26    20
```

Możemy też jawnie podać wymiar po którym dana operacja ma być wykonana

```
>> sum(A,2)

ans =

    10
    26
    20
```

Działają też zwykłe operatory $+-$

```
>> B=[1 2;3 4]

B =

     1     2
     3     4

>> B+B

ans =

     2     4
     6     8

>> B-B

ans =

     0     0
     0     0
```

Operatory `* / ^` działają na "całych" macierzach

```
>> B/B

ans =

     1     0
     0     1

>> B*B

ans =

     7    10
    15    22
```

```
>> B./B

ans =

     1     1
     1     1

>> B.*B

ans =

     1     4
     9    16
```

Do elementu macierzy dostajemy się tak:

```
>> B(1,2)

ans =

     2
```

Teraz zmieniamy jego wartość:

```
>> B(1,2)=4;
>> B

B =

     1     4
     3     4

>>
```

Zwróćmy uwagę, że przy modyfikowaniu elementów macierzy jej rozmiar dostosowuje się automatycznie i może się zmienić!:

```
>> B(1,3)=4;
>> B
```

```
B =

     1     4     4
     3     4     0
```

Kontrola zakresu jest tylko przy pobieraniu elementów macierzy:

```
>> B(3,1)
??? Index exceeds matrix dimensions.
```

Operator dwukropek, :, jest jednym z najbardziej użytecznych operatorów w MATLABie. Występuje w kilku różnych formach wyrażenie 1:10 produkuje wektor wierszowy o elementach od 1 do 10.

```
>> 1:10

ans =

     1     2     3     4     5     6     7     8     9    10
```

Jeśli podamy inkrement to uzyskamy wektor o pożądanej różnicy między elementami np:

```
>> 10:-2.5:0

ans =

   10.0000    7.5000    5.0000    2.5000         0
```

Operator ten zastosowany w indeksie macierzy daje nam łatwy dostęp do jej fragmentów `A(1:k,j)` daje nam pierwszych k elementów j-tej kolumny macierzy A. Sam ":" oznacza wszystkie elementy danego wiersza lub kolumny np. sum(A(:,end)) oblicza sumę elementów ostatniej kolumny A

Do sklejania macierzy służy operator []. Np

```
B=ones(2,2);
C=[B B+1; B+2 B+3]
```

Możemy usuwać kolumny lub wiersze macierzy:

```
>> C=[1 2 3 4;5 6 7 8; 9 10 11 12; 13 14 15 16]
C(2,:)=[]
```

Do analizy danych statystycznych MATLAB używa danych zorientowanych kolumnowo. Każda kolumna w zestawie danych reprezentuje zmienną a wiersz obserwację (pomiar) tej zmiennej. Element (i,j) jest więc i-tą obserwacją j-tej zmiennej. Jako przykład rozważmy dane z trzema zmiennymi: Rytm serca, waga, ilość godzin ćwiczeń na tydzień. Dla pięciu obserwacji macierz z danymi wygląda np. tak:

```
D =
      72          134          3.2
      81          201          3.5
      69          156          7.1
      82          148          2.4
      75          170          1.2
```

Pierwszy wiersz zawiera rytm serca, wagę, ilość godzin ćwiczeń na tydzień dla pacjenta 1, drugi wiersz to samo dla pacjenta 2 itd. Możemy do tak przygotowanych danych zastosować jedną z licznych funkcji do analizy danych np. policzmy średnią i odchylenie standardowe poszczególnych zmiennych

```
mu = mean(D)
sigma = std(D)
```

## 1.4 Funkcje i skrypty

Kawałek kodu matlabowego zapisany w pliku tekstowym (z rozszerzeniem .m) to skrypt. Można go wykonać wpisując nazwę pliku (bez rozszerzenia). Skrypt ma dostęp do wszystkich zmiennych znajdujących się w workspace, zmienne wytworzone w skrypcie są widoczne w workspace.

Większość poleceń Matlaba to funkcje, niektóre są wbudowane i działają bardzo szybko, ale znaczna część jest napisana w plikach tekstowych, które są interpretowane w czasie wykonywania (działają wolniej). Ma to jednak tą zaletę, że możemy do takiej funkcji zajrzeć i dużo się nauczyć, albo ją zmodyfikować! :-)

W matlabie można też tworzyć własne funkcje — zbudowane z już istniejących. Plik zawierający funkcję musi nazywać się tak jak ta funkcja z roszerzeniem ".m" Pierwsza linia definiuje składnię wywołania funkcji np:

```
function [mean,stdev] = stat(x)
%STAT Interesting statistics.
n = length(x);
mean = sum(x) / n;
stdev = sqrt(sum((x - mean).^2)/n);
```

Powyższy kod definiuje funkcję stat (powinna być zapisana w pliku stat.m). Funkcja ta przyjmuje jako argument wektor `x` i zwraca dwie wartości `mean,stdev` zmienne używane wewnątrz funkcji są lokalne tzn. nie są widoczne w workspace.

Przykład wywołania tej funkcji:

```
x=1:10;
>> [m,s]=stat(x)
m =
    5.5000
s =
    2.8723
```

W jednym pliku możemy mieć zdefiniowanych więcej funkcji, z tym, że są one widoczne tylko dla funkcji zawartych w tym samym pliku np. powyższą funkcję stat można zaimplementować tak:

```
function [mean,stdev] = stat(x)
%STAT Interesting statistics.
n = length(x);
mean = avg(x,n);
stdev = sqrt(sum((x-avg(x,n)).^2)/n);

%------------------------
function mean = avg(x,n)
%MEAN subfunction
mean = sum(x)/n;
```

Powrót z funkcji następuje po osiągnięciu końca ciała funkcji. Wcześniejszy powrót warunkowy można uzyskać dzięki poleceniu `return`

## 1.5 Przydatne funkcje

**plot**

```
%wytwarzamy wektor t o elementach od 1 do 1024 co 1
>> t=1:256;
%rysujemy wektor t - domyślnie jest on łączony odcinkami prostej
>> plot(t)
% tu zobaczymy jakie mamy naprawdę elementy wektora
>> plot(t,'g.')
% dzielimy wszystkie elementy wektora przez 128
% (możemy sobie interpretować t jako czas 2s próbkowany co 1/128 sek).
>> t=t/128;
% robimy sinusa z okresem 1 (s)
>> x=sin(2*pi*t);
% rysujemy wektor x
>> plot(x)
% rysujemy wektor x względem wektora t
>> plot(t,x)
% rysujemy wektor x względem wektora
% t linią ciągłą na niebiesko i na tym
% tle rysujemy co piąty element x i t
% czerwonymi kółkami
>> plot(t,x,'b-',t(1:5:end),x(1:5:end),'ro')
```

**rand, hist**

```
 x=rand(300,1);
>> plot(x)
>> hist(x)
>> hist(x,20)
```

**randn**

```
 x=randn(3000,1);
 hist(x,20)
```

**ceil** `ceil` przydaje się szczególnie w połączeniu z `rand` do wytwarzania losowych prób z wektora:

```
>> x=1:128;
>> plot(x)
```

```
>> title('oto nasz wektor z liniowo rozmieszczonymi elementami od 1 do 128')
% mieszamy indeksy wektora x
>> y=x(ceil(length(x)*rand(size(x))));
% po kolei:
% size(x) - zwraca nam rozmiar x
% rand(size(x)) - robimy wektor o takim samym rozmiarze jak x złożony
% z liczb losowych z przedziału (0,1)
% length(x)*rand(size(x))) - z przedziału (0,1) robimy przedział (0, długość(x)
% na koniec zaokrąglamy do góry dzięki temu uzyskujemy liczby naturalne
% [1,długość(x)] - czyli prawidłowe indeksy matlaba
>> plot(y)
```

**boxplot** przydatny do zgrubnego obejrzenia rozkładu

```
boxplot(X,NOTCH,SYM,VERT,WHIS)
```

```
produces a box and whisker plot for
each column of X. The box has lines at the lower quartile (25 percentyl),
median, and upper quartile (75 percentyl) values.
The whiskers are lines extending from
each end of the box to show the extent of the rest of the data.
Outliers (wartości odstające) are data with values
beyond the ends of the whiskers.

NOTCH = 1 produces a notched-box plot. Notches represent a robust
estimate of the uncertainty about the medians for box to box comparison.
```

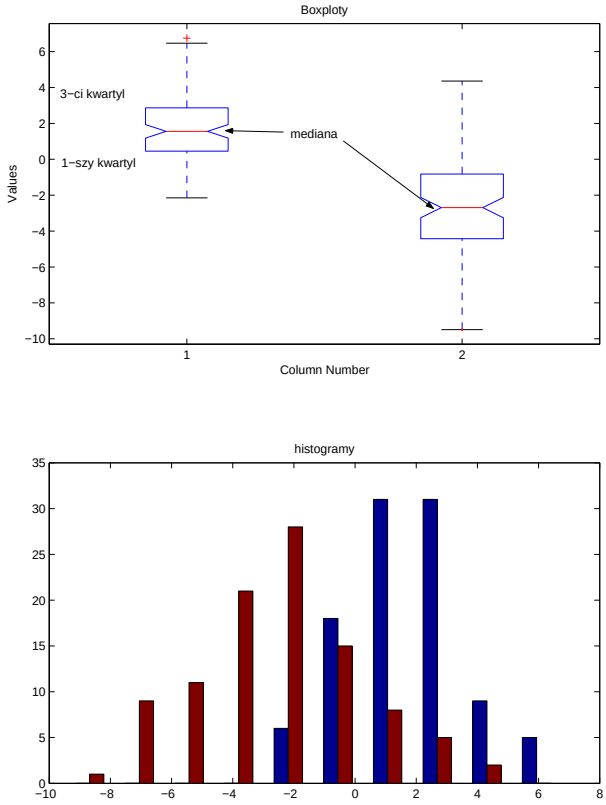
n1 = med + 1.57*(q3-q1)/sqrt(length(x));

n2 = med - 1.57*(q3-q1)/sqrt(length(x));

```
NOTCH = 0 (default) produces a rectangular box plot.
SYM sets the symbol for the outlier values if any (default='+').
VERT = 0 makes the boxes horizontal (default: VERT = 1, for vertical).
WHIS defines the maximum length of the whiskers as a function of the
IQR (inter quartile range odległość między 25 a 75 percentylem)(default = 1.5).
The whisker extends to the most extreme data
value within WHIS*IQR of the box.  If WHIS = 0 then BOXPLOT displays
all data values outside the box using the plotting symbol, SYM.

BOXPLOT(X,G,NOTCH,...) produces a box and whisker plot for the vector
```

```
X grouped by G.  G is a grouping variable defined as a vector, string
matrix, or cell array of strings.  G can also be a cell array of
several grouping variables (such as {G1 G2 G3}) to group the values
in X by each unique combination of grouping variable values.
```




Rysunek 1.1: Porównanie histogramu i boxplotów

```
x1=2*(randn(100,1)+1);
x2=3*(randn(100,1)-1);
z=[x1 x2];
subplot(211)
boxplot(z,1)
subplot(212)
hist(z)
```

**sort**

```
y=sort(x2)
plot(y)
```

**find**

```
x=randn(1000,1);
y=x(find(x>2));
length(y)
hist(x,20)
```

## 1.6 Instrukcje sterujące

Instrukcje sterujące matlaba:

**if** together with **else** and **elseif**, executes a group of statements based on some logical condition

```
if I == J
  A(I,J) = 2;
elseif abs(I-J) == 1
  A(I,J) = -1;
else
  A(I,J) = 0;
end
```

**switch** together with case and otherwise, executes different groups of statements depending on the value of some logical condition.

```
method = 'bilinear';

 switch method
   case {'linear','bilinear'}
     disp('Method is linear')
   case 'cubic'
     disp('Method is cubic')
   case 'nearest'
```

```
    disp('Method is nearest')
  otherwise
    disp('Unknown method.')
end

Method is linear
```

**while** executes a group of statements an indefinite number of times, based on some logical condition

```
i=10
while i>1
i=i-1;
disp(i)
end
```

**for** executes a group of statements a fixed number of times.

```
for k=1:10
disp(k*(1:10))
end
```

**continue** passes control to the next iteration of a for or while loop, skipping any remaining statements in the body of the loop

**break** terminates execution of a for or while loop

**try...catch** changes flow control if an error is detected during execution

**return** causes execution to return to the invoking function.

# Rozdział 2

# Generatory liczb pseudolosowych

## 2.1 Wstęp

Będziemy rozważać generatory typu $x_{n+1} = f(x_n, x_{n-1}, ..., x_{n-k})(modM)$. Zakładamy, że argumentami funkcji $f$ są liczby całkowite ze zbioru $0, 1, ..., M-1$. Dla ustalenia uwagi mogą to być generatory liniowe typu:

$$x_{n+1} = (ax_n + c)(modM)$$

Generatory takie mają niestety okres, po którym sekwencja liczb powtarza się.

## 2.2 Generatory

Poniżej zamieszczony jest kod pięciu przykładowych generatorów liczb pseudolosowych:

### 2.2.1 Generator 1

```
function y=gen1(x,N)
a=16807;
m=2147483647;
q=127773;
r=2836;
```

```
y=zeros(N,1);
for i=1:N
   hi=floor(x/q);
   lo=mod(x,q);
   test=a*lo-r*hi;
   if test>0
      x=test;
   else
      x=test+m;
   end
y(i)=x/m;
end
```

### 2.2.2 Generator 2

```
function y=gen2(x,N)
m=8191;
a=101;
c=1731;
y=zeros(N,1);
for i=1:N
   x=mod(a.*x+c,m);
   y(i)=x/m;
end
```

### 2.2.3 Generator 3

```
function y=gen3(x,N)
a=517;
m=32767;
c=6923;
y=zeros(N,1);
for i=1:N
   x=mod(a.*x+c,m);
   y(i)=x/m;
end
```

### 2.2.4 Generator 4

```
function y=gen4(x,N)
c=65536;
y=zeros(N,1);
for i=1:N
   x=x*25;
   x=mod(x,c);
   x=x*125;
   x=mod(x,c);
   y(i)=x/c;
end
```

### 2.2.5 Generator 5

```
function y=gen5(x,N)
a=16807;
rm=2147483647;
q=127773;
r=2836;
y=zeros(N,1);
for i=1:N
   A=floor(x/q);
   test=a*x - A*(a*q+r);
   if test>0
      x=test;
   else
      x=test+rm;
   end
   y(i)=x/rm;
end
```

## 2.3 Testy jakości

Dla wyżej wymienionych generatorów oraz dla generatora wbudowanego w Matlab, proszę wykonać następujące testy:

1. narysować histogram rozkładu gęstości prawdopodobieństwa

2. test zgodności momentów – czy momenty obliczone dla wygenerowanych liczb są takie jak przewiduje teoria

3. test korelacji na rysunku: sporządzić wykres gdzie na jednej osi są wartości $x_n$ na drugiej zaś wartości $x_{n+1}$.

Baterie testów jakości generatorów np:

- `http://csrc.nist.gov/groups/ST/toolkit/rng/index.html`
- `http://www.phy.duke.edu/~rgb/General/rand_rate.php`

# Rozdział 3

# Rozkłady prawdopodobieństwa

## 3.1 Kilka użytecznych rozkładów prawdopodobieństwa

Korzystając z omówionych generatorów liczb pseudolosowych o rozkładzie płaskim można wygenerować w zasadzie dowolny zadany rozkład gęstości prawdopodobieństwa.

### 3.1.1 Rozkład dwumianowy

Zmienna losowa, która zlicza liczbę sukcesów $k$ w $n$ próbach, gdzie $p$ jest prawdopodobieństwem sukcesu w pojedynczej próbie, podlega rozkładowi dwumianowemu:

$$P_n(k) = \left( \begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

**Problem** Jak z rozkładu płaskiego wytworzyć zmienne losowe o rozkładzie dwumianowym?

**Zadanie** Z oszacowań agencji wynika, że średnio 2 z 3 reklam spotyka się z pozytywnym odzewem. Akcja marketingowa obejmuje 12 reklam. Niech $X$ oznacza liczbę reklam skutecznych. Czy $X$ podlega rozkładowi dwumianowemu? Jakie jest prawdopodobieństwo, że 10 reklam będzie skutecznych? Prawdopodobieństwo obliczyć ze wzoru oraz korzystając z symulacji.

[Odp: p=0.127]

### 3.1.2 Rozkład Poissona

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Rozkładowi Poissona podlegają zmienne losowe zliczające w jednostce czasu ilość zdarzeń o niskim prawdopodobieństwie zajścia. Np. ilość rozpadów promieniotwórczych na jednostkę czasu.

**Przykład** Lekarz pełniący dyżur w szpitalu jest wzywany do pacjentów średnio 3 ray w ciągu nocy. Załóżmy, że liczba wezwań na noc podlega rozkładowi Poissona. Jakie jest prawdopodobieństwo, że noc upłynie lekarzowi spokojnie? U nas $\mu = 3$, $x = 0$ więc

$$P(0) = \frac{3^0 e^{-3}}{0!} = e^{-3} = 0.0498$$

**Problem** Jak ze zmiennych podlegających rozkładowi płaskiemu uzyskać zmienne podlegające rozkładowi Poissona?

### 3.1.3 Rozkład Gaussa — rozkład normalny

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

gdzie: $\mu$ – średnia, $\sigma$ – odchylenie standardowe ($\sigma^2$ –wariancja).

Najbardziej efektywny sposób wytwarzania zmiennych losowych o rozkładzie normalnym polega na zastosowaniu Centralnego Twierdzenia Granicznego.

$$X = \sum_{n=1}^{N} Y_n - \frac{N}{2}$$

gdzie $Y$ zmienna losowa z rozkładu płaskiego $(0, 1)$

**Zadanie** Proszę zrobić histogramy `histfit` 10 000 liczb $X$ uzyskanych dla $N = 1, 2, ..., 12$

**Problem** Jakie parametry charakteryzują rozkład do którego zbiegają sumy?

Rozkład o średniej 0 i wariancji 1 (notacja $N(0,1)$) jest nazywany rozkładem standardowym i często jest oznaczany literą $Z$. Dokonując odpowiedniej transformacji można z rozkładu $Z$ uzyskać dowolny inny rozkład normalny.

**Zadanie** Proszę uzyskać i narysować rozkład $N(2,9)$.

**Przykład** Producent silników twierdzi, że jego silniki mają średnią moc 220KM a odchylenie standardowe wynosi 15 KM. Potencjalny klient testuje 100 silników. Jakie jest prawdopodobieństwo, że średnia z próby będzie mniejsza niż 217 KM?

Przypomnijmy, że z CTG dla dużych liczebności próby $n$ $\bar{x} \sim N(\mu, \sigma^2/n)$. Zatem szukamy

$$P(\bar{x} < 217) = P\left(Z < \frac{217-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z < \frac{217-220}{\frac{15}{\sqrt{100}}}\right) = P(Z < -2) = 0.028$$

```
% ze wzoru:
p=normcdf(-2)

% symulacja

mu=220;
sig=15;
N_prob=100;

m_kryt=217;

N_rep=1e5;
srednia=zeros(1,N_rep);
for i=1:N_rep
    seria=sig*randn(1,100)+mu;
    srednia(i)=mean(seria);
end
[n,x]=hist(srednia,30);
bar(x,n)
line([m_kryt m_kryt],[0 max(n)],'Color',[1 0 0])
```

```
p1=sum(srednia<=m_kryt)/N_rep
```

Gdy nie znamy odchylenia standardowego populacji $\sigma$ używamy w jego miejsce estymatora wariancji $S^2$ danego wzorem:

$$S^2 = \frac{\sum(\bar{x} - x_i)^2}{n-1}$$

Wtedy rozkład $Y = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ nie podlega rozkładowi normalnemu. Jeśli populacja $X$ podlega rozkładowi normalnemu to $Y$ podlega **rozkładowi $t$ z $n-1$ stopniami swobody**.

**Zadanie** Teraz kilka najprostszych zastosowań rozkładu normalnego:

1. Znajdźmy prawdopodobieństwo, że $Z < -2.47$. Proszę zrobić to na dwa sposoby: raz z użyciem wygenerowanego rozkładu normalnego dla $N = 12$, drugi raz z użyciem funkcji `normcdf` [odp: p=0.0068]

2. Znaleźć prawdopodobieństwo $P(|Z| < 2)$ [Odp: p=0.9545]

3. Koncentracja zanieczyszczeń w półprzewodniku używanym do produkcji procesorów podlega rozkładowi normalnemu o średniej 127 cząsteczek na milion i odchyleniu standardowemu 22. Półprzewodnik może zostać użyty jedynie gdy koncentracja zanieczyszczen spada poniżej 150 cząstek na milion. Jaka proporcja półprzewodników nadaje się do użycia? Prawdopodobiństwo obliczyć korzystając z dystrybuanty rozkładu normalnego oraz z symulacji. [Opd: p=0.852]

# Rozdział 4

# Przedziały ufności

**Przedział ufności** (CI) odzwierciedla zarówno wielkość badanej grupy jak i zmienność analizowanej cechy wewnątrz tej grupy. Średnia będąca wynikiem przeprowadzonych badań nie jest równa rzeczywistej średniej populacyjnej. Rozbieżność między uzyskanym wynikiem a rzeczywistą średnią populacji zależy od wielkości badanej grupy oraz zmienności badanej cechy w jej obrębie. Jeśli badana grupa jest niewielka i ma dużą zmienność analizowanej cechy wówczas rozbieżność między średnią uzyskaną a rzeczywistą może być znaczna. Natomiast, jeśli badana grupa jest dużą z niewielką zmiennością danych wówczas uzyskana średnia będzie prawdopodobnie bardzo bliska średniej populacyjnej. CI jest określany z różnym procentem "zaufania", np. 90 czy też 95%. Najczęściej używa się 95% przedziału ufności, który przy założeniu, że grupa badana była zgromadzona w sposób losowy wskazuje z 95% pewnością, że w zakresie przedziału ufności znajduje się rzeczywista średnia populacyjna. Przedział ufności jest wskaźnikiem precyzji wykonanych pomiarów.

Pewnych problemów koncepcyjnych nastręcza konstrukcja przedziałów ufności. Dla danej znanej populacji możemy obliczyć średnią populacji oraz prawdopodobieństwo uzyskania konkretnej wartości średniej przy losowaniu próby o zadanej liczebności z tejże populacji, możemy więc określić prawdopodobieństwo $P$, że odległość średniej z próby i średniej z populacji jest $D$. Mając do dyspozycji tylko próbę możemy albo założyć, że pochodzi ona z jakiegoś znanego rozkładu i wyliczyć z niego wartości krytyczne albo zakładając, że jest ona reprezentatywna możemy metodą bootstrapu "wytworzyć" wiele innych prób z badanej populacji i oszacować jakie są granice, w które wpada żądana frakcja średnich (np.:90%, 95%).

**Bootstrap** jest związany z pobieraniem próby. Najkorzystniejszą sytu-

acją jest ta, w której dla oszacowania różnych parametrów statystycznych populacji mamy możliwość pobierania z tej populacji wielu prób. Jeśli jest to niemożliwe możemy posłużyć się pobieranim wielokrotnie prób z tej próby którą posiadamy. Postępowanie takie jest sensowne pod warunkiem, że próba, która służy nam do generowania innych możliwych pobrań próby jest *reprezentatywna*. **W bootstrapie losujemy ze zwracaniem** (dlaczego?).

## 4.1 Przedział ufności dla średniej

Przedział ufności $(1-\alpha)*100$ % dla średniej $\mu$, gdy znamy odch. std. $\sigma$ i próba pochodzi z rozkładu normalnego lub jest dostatecznie duża:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

### 4.1.1 Przykład

Wyciągamy losową próbę ($n = 25$) z populacji o rozkładzie normalnym. Dostajemy średnią z próby $\bar{x} = 122$. Załóżmy, że znamy standardowe odchylenie populacji $\sigma = 20$. Oblicz przedział ufności 95 % dla średniej populacji $\mu$. Co zrobić aby zmniejszyć obliczony przedział 10-krotnie?

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 122 \pm 1.96 * \frac{20}{\sqrt{25}} = 122 \pm 7.84 = [114.16\ 129.84]$$

Możemy być pewni na 95%, że nieznana średnia populacji $\mu$ znajduje się pomiędzy 114.16 a 129.84. Jeśli chcemy zmniejszyć przedział ufności 10-krotnie, musimy pobrać 100 razy większą próbę tj. n = 2500.

### 4.1.2 Zadanie

Importer win musi zbadać średnią zawartość alkoholu w nowej partii win francuskich. Z doświadczenia z poprzednimi gatunkami wina, przyjmuje on, że standardowe odchylenie w populacji wynosi 1.2 %. Importer wybrał losową próbę 60 butelek nowego wina i otrzymał średnią z próby 9.3 %. Znaleźć przedział ufności 90 % dla średniej zawartości alkoholu w nowej partii win. Odp. [9.045 9.55]

Przedział ufności $(1-\alpha)*100\%$ dla średniej $\mu$, gdy nie znamy odch. std. $\sigma$ i próba pochodzi z rozkładu normalnego:

$$\left[\bar{x} - t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

gdzie $t_{\alpha/2}$ jest wartością, która odcina obszar $\alpha/2$ z rozkładu $t$ z $n-1$ stopniami swobody.

### 4.1.3 Przykład

Lekarz chce zbadać średni czas trwania kuracji tj. od podania leku do ustąpienia objawów w pewnej chorobie. Losowa próba 15 pacjentów dała średni czas $\bar{x} = 10.37$ dnia i odchylenie standardowe $s = 3.5$ dnia. Zakładając normalny rozkład w populacji czasów trwania kuracji znaleźć 95 % przedział ufności dla średniego czasu trwania kuracji.

Znajdujemy wartość z rozkładu $t$ o $n-1$ $(= 14)$ stopniach swobody, która odcina obszar $\alpha/2 = 0.025$. $t_{0.025} = 2.145$. Dostajemy więc $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 10.37 \pm 2.145 * \frac{3.5}{\sqrt{15}} = [8.4312.31]$ Lekarz może być pewny na 95 %, że od podania leku do ustąpienia objawów upłynie czas pomiędzy 8.43 a 12.31 dnia

### 4.1.4 Zadanie

Producent opon rowerowych chce oszacować średni dystans jaki można przejechać na oponie pewnego rodzaju zanim opona się zużyje. Pobrano losową próbę 32 opon, opona jest używana aż do przetarcia i odległość przejechana na każdej oponie jest rejestrowana. Dane, w tysiącach kilometrów, są następujące:
32, 33, 28, 37, 29, 30, 25, 27, 39, 40, 26, 26, 27, 30, 25, 30, 31, 29, 24, 36, 25, 37, 37, 20, 22, 35, 23, 28, 30, 36, 40, 41.
Znaleźć 99 % przedział ufności dla średniego przebiegu opon tego rodzaju. Odp. [27.76 33.36]

## 4.2 Przedział ufności dla wariancji

Przedział ufności $(1-\alpha)*100\%$ dla wariancji populacji $\sigma^2$, gdy rozkład populacji jest normalny:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]$$

gdzie $\chi^2_{\alpha/2}$ jest wartością, która odcina na prawo obszar $\alpha/2$ z rozkładu chi - kwadrat z $n-1$ stopniami swobody. $\chi^2_{1-\alpha/2}$ jest wartością, która odcina na lewo obszar $\alpha/2$ z rozkładu chi - kwadrat z $n-1$ stopniami swobody (lub równoważnie: odcina na prawo obszar $1-\alpha/2$).

Skąd taki wzór? Wystarczy zauważyć, że:

$$\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$$

podlega rozkładowi $\chi^2$ o $N-1$ stopniach swobody, zaś estymator wariancji to

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$$

### 4.2.1 Zadanie

Automat do kawy nalewa kawę do kubków. Jeśli średnia porcja kawy w kubku odbiega od normy, maszynę można wyregulować. Jeśli jednak wariancja porcji kawy jest zbyt duża, maszyna jest 'out of control" i wymaga reperacji. Od czasu do czasu przeprowadzana jest kontrola wariancji porcji kawy. Odbywa się to poprzez wybór losowej próby napełnionych kubków i policzenie wariancji próby. Losowa próba 30 kubków dała wariancję próby $s^2 = 18.54$. Obliczyć 95 % przedział ufności dla wariancji populacji $\sigma^2$. Zadanie rozwiązać za pomocą powyższego wzoru oraz za pomocą symulacji.

Odp. [11.765 33.604]

## 4.3 Rozmiar próby

Gdy wyciągamy próbę, często ważne jest jaki jest minimalny rozmiar próby, który zapewni nam żądaną precyzję wyniku.

Musimy odpowiedzieć sobie na trzy pytania:

1. Jak nasze oszacowanie nieznanego parametru ma być bliskie prawdziwej wartości? Odpowiedź oznaczmy $D$ (dystans).

2. Jaki chcemy mieć poziom ufności, że nasze oszacowanie i prawdziwa wartość parametru są od siebie oddalone o nie więcej niż D?

3. Jakie jest oszacowanie wariancji w populacji?

Jeśli nie znamy odpowiedzi na pkt. 3 przeprowadzamy tzw. pilot study i szacujemy $\sigma$ na podstawie odchylenia std. próby. Minimalny rozmiar próby potrzebny do oszacowania średniej populacji $\mu$ wynosi:

$$n = \frac{z^2_{\alpha/2}\sigma^2}{D^2}$$

### 4.3.1 Zadanie

Biuro podróży chce oszacować średnią ilość pieniędzy wydaną na wakacje przez osoby korzystające z jego usług. Ludzie przeprowadzający analizę chcieliby móc oszacować średni koszt wakacji z dokładnością do 200 zł na poziomie ufności 95 %. Z poprzednich doświadczeń tego biura podróży wynika, że odchylenie standardowe w populacji wynosi $\sigma = 400$ zł. Jaka będzie minimalna wielkość próby?

Odp. $n = 15.366$ wiec wielkość próby wynosi 16 (zaokrąglamy w górę)

## 4.4 Przykład z bootstrapem

Rozważmy sądę przedwyborczą, mamy dwóch kandydatów na prezydenta. Ankietowano 1500 osób. 840 osób deklarowało poparcie dla kandydata A zaś 660 dla kandydata B. Na ile pewny może być kandydat A swojego zwycięstwa?

1. Jak dokładnie brzmi pytanie? Jaki jest 95% przedział ufności dla poparcia kandydata A w całej populacji? Czy też innymi słowami: W jakim przedziale na 95% znajduje się proporcja glosujących popierających kandydata A.

2. Nasze najlepsze mniemanie o własnościach "świata" z którego pochodzą dane otrzymujemy ze zwykłej proporcji. Wynika z niej, że kandydat A ma poparcie 56% zaś kandydat B poparcie 44% wyborców.

3. Przypiszmy do kandydata A 1 zaś do B 0

4. Pobranie ankiety modelujemy przez pobranie losowo 1500 próbek z modelu naszego "świata" czyli wektora złożonego z 56 zer i 44 jedynek. Wynikiem jednej ankiety jest proporcja popierających kandydata A (lub B)

5. Zbieramy rozkład proporcji – musimy w tym celu "przeprowadzić" wielokrotnie ankietę. Narysujmy histogram.

6. Chcemy znaleźć 95% przedział ufności musimy znaleźć percentyl 2.5 oraz 97.5. Liczby te stanowią poszukiwany przedział ufności.

```
x=ones(1,100);
x(55:100)=0;
Nboot=100000;
A=zeros(1,Nboot);
for i=1:Nboot
    %ankieta=x(ceil(100*rand(1,1500)));
    ankieta=randsample(x,1500,true);
    A(i)=sum(ankieta)/1500;
end
lo=prctile(A,2.5);
hi=prctile(A,97.5);
disp(sprintf('przedzial ufnosci: %.3f - %.3f\n',lo,hi))
[n,x]=hist(A,30);
bar(x,n)
line([lo lo] , [0 max(n)] ,'Color',[1 0 0] )
line([hi hi],[0 max(n)],'Color',[1 0 0] )
```

## 4.5 Zadania

### 4.5.1 Przyrost masy w nowej diecie

Producent karmy dla zwierząt chciał przetestować nowy rodzaj karmy. Próbki podawał 12 zwierzakom przez 4 tygodnie. Po tym czasie zanotował następujące przyrosty masy:
15.43 16.92 14.43 12.94 15.92 17.42 18.91 16.92 14.93 14.49 15.92 15.43 kg
średni przyrost wynosi 15.80 kg. producent widzi jednak, że w próbie jest

dość znaczny rozrzut pomiędzy poszczególnymi zwierzętami $12.94 - 18.91$ i nie jest pewien czy można reklamować nowy pordukt podając średni przyrost 15.8 kg. Podejrzewa, że inna grupa zwierząt może mieć zupełnie inną średnią.

- Używając powyższych danych znajdźmy szansę, że w innej losowej próbie 12 zwierząt uzyskamy średni przyrost masy poniżej 15 kg.
- Wynik zilustrować przy pomocy histogramu.
- Jaki byłby wynik przy założeniu, że masy zwierząt pochodzą z rozkładu normalnego?

[odp: $p \approx 0.03$, $p_{normalne} = 0.053$]

### 4.5.2 Średnica drzew

Ogrodnik eksperymentuje z nowym gatunkiem drzew. Posadził 20 sztuk i po dwóch latach zmierzył następujące średnice pni (w cm):
8.5 7.6 9.3 5.5 11.4 6.9 6.5 12.9 8.7 4.8 4.2 8.1 6.5 5.8 6.7 2.4 11.1 7.1 8.8 7.2

- Proszę obejrzeć boxplot.
- Proszę znaleźć średnią średnicę i 90% przedział ufności dla średniej.
- Proszę znaleźć medianę i 90% przedział ufności dla mediany.
- Wynik zilustrować przy pomocy histogramu.

```
Odp:
srednia i jej przedzial ufnosci[6.57 7.50 8.39]
mediana i jej przedzial ufnosci[6.50 7.15 8.50]
```

### 4.5.3 Zawartość aluminium w Tebańskich naczyniach.

Zawartość procentowa aluminium w 18 antycznych naczyniach z Teb była następująca:
11.4 13.4 13.5 13.8 13.9 14.4 14.5 15 15.1 15.8 16 16.3 16.5 16.9 17 17.2 17.5 19.0
Jaka jest mediana procentowej zawartości aluminium i jaki jest 95% przedział ufności.

```
Odp: mediana i jej 95 proc przedzial ufnosci[14.15 15.45 16.65]
```

### 4.5.4 Przedział ufności dla różnicy dwóch średnich

Mamy 7 myszy, którym podano środek, który miał poprawić ich przeżywalność po operacji oraz 9 myszy kontrolnych, którym owego środka nie podano. Myszy traktowane specjalnie przeżyły
94 38 23 197 99 16 141 dni
a myszy traktowane standardowo:
52 10 40 104 51 27 146 30 46 dni
Srednia różnica wynosi 30.63 dni dłużej dla myszy traktowanych po nowemu. Pytanie, na które chcielibyśmy znać odpowiedź to: Czy nowy środek faktycznie poprawia przeżywalność.

Skonstruujmy przedział ufności 95% dla średniej różnicy w przeżywalności.

Uwaga: przy tym problemie każdą z grup traktujemy jako reprezentantów bardzo dużych populacji.

```
Odp:
oryginalna roznica srednich: 30.63
przedzial ufnosci z reprobkowania [-20.5 84.7]
przedzial ufnosci z zalozeniem normalnosci [-31.4 92.7]
```

### 4.5.5 Przedział ufności dla proporcji

W badaniach nad cholesterolem u ludzi stwierdzono, że w grupie 135 badanych z wysokim poziomem cholesterolu 10 osób przeszło zawał serca. Pytanie: Na ile pewni możemy być, że jeśli weźmiemy dużo większą grupę pod uwagę to proporcja zawałowców będzie podobna, czyli konkretnie jaki jest 95% przedział ufności dla proporcji 10/135 ? Obejrzeć histogram. Jakie wnioski?

```
Odp: proporcja i jej 95 proc przedzial ufnosci[0.03 0.07 0.12]
```

### 4.5.6 Bezrobotni

W próbce 200 osób 7 procent jest bezrobotnych. Określić 95% przedział ufności dla prawdziwej średniej w populacji.

```
Odp: średnia i jej 95 proc przedzial ufnosci [0.010 0.035 0.060]
```

### 4.5.7 Żywotność baterii

W próbce 20 testowanych baterii stwierdzono średni czas życia 28.85 miesiąca. Określić 95% przedział ufności dla średniej. Wartości dla badanej próbki były następujące:
30 32 31 28 31 29 29 24 30 31 28 28 32 31 24 23 31 27 27 31 miesięcy
Obejrzeć rozkład przy pomocy histfit i zbadać jaki wpływ na przedział ufności ma przyjęcie założenia o normalności rozkładu czasów życia.

`Odp: srednia i jej 95 proc przedzial ufnosci[27.65 28.85 29.95]`

### 4.5.8 Pomiary

Mamy 10 pomiarów pewnej wielkości:
0.02 0.026 0.023 0.017 0.022 0.019 0.018 0.018 0.017 0.022
Proszę znaleźć średnią i 95% przedział ufności.

`Odp: średnia i jej 95 proc przedział ufności [0.0185 0.0202 0.0220]`

Czy pomiarów jest wystarczająco dużo aby sensownie wyznaczyć średnią i przedział ufności?

`Wsk: obliczyć średnie dla 1e6 powtórzeń i obejrzeć histogramy dla 10, 20, 30, 1`

# Rozdział 5

# Testowanie hipotez dotyczących jednej lub dwóch populacji

## 5.1 Wstęp

### 5.1.1 Hipoteza zerowa i alternatywna

Podstawą sukcesu w statystycznym testowaniu hipotez jest prawidłowe ich sformułowanie. Hipotezy muszą być rozłączne. Tak kombinujemy, żeby jako hipoteza zerowa wyszło to co chcemy odrzucić, gdyż błąd takiej decyzji możemy bezpośrednio kontrolować. Logika testowania jest następująca: tworzymy funkcję od zmiennych losowych, dla której przy spełnineniu przez owe zmienne hipotezy zerowej potrafimy podać prawdopodobieństwa z jakimi przyjmuje ona różne wartości. Ta funkcja nazywana jest statystyką. Następnie obliczamy wartość tej funkcji dla badanej próby. Jeśli prawdopodobieństwo osiągnięcia otrzymanej bądź jeszcze bardziej ekstremalnej wartości statystyki jest niskie to wątpimy, że nasze dane są zgodne z hipotezą zerową i jesteśmy skłonni przyjąć hipotezę alternatywną.

### 5.1.2 Różne podejścia do tego samego problemu

Zasadniczym problemem jest wybór staytstyki. Mamy dwie zasadnicze możliwosci:

- znamy rozkład prawdopodobieństwa, z którego pochodzą nasze dane, lub umiemy je przetranformować do znanego rozkładu, wtedy bierzemy

"z półki" klasyczny test parametryczny np. test-t, $\chi^2$, F itp

- nie znamy rozkładu prawdopodobieństwa naszych danych albo nie chcemy nic o nim zakładać. W tym wypadku znowu mamy dwie możliwości:

  - korzystamy z klasycznego testu nieparametrycznego np.:

    **Wilcoxon rank sum test** — `ranksum` — testuje hipotezę zerową, że dwie próby $X$ i $Y$, które ze sobą porównujemy pochodzą z populacji o takiej samej medianie. Próby $X$ i $Y$ **nie są** sparowane.

    **Wilcoxon signed rank test** —`signrank` — testuje hipotezę zerową, że dwie próby $X$ i $Y$, które ze sobą porównujemy pochodzą z populacji o takiej samej medianie. Próby $X$ i $Y$ **są** sparowane.

    **Test znaków** — `signtest` — testuje hipotezę zerową, że dwie próby $X$ i $Y$, które ze sobą porównujemy pochodzą z populacji o takiej samej medianie. Próby $X$ i $Y$ **są** sparowane.

    Jeśli wystarczy czasu to proponuję przećwiczenie tych testów na zadaniach, które robiliśmy przy resamplingu i przy klasycznych testach parametrycznych. Proszę się przy tym zastanowić, które testy mają większą, a które mniejszą moc.

  - wytwarzamy rozkład statystyki na podstawie naszych danych przez repróbkowanie. W podejściu repróbkowania tworzymy statystyczny model badanego procesu i następnie badamy w drodze symulacji prawdopodobieństwa generowania przez ten model interesujących nas sytuacji. Największą uwagę musimy tu poświęcić na prawidłowe sformułowanie modelu, a następnie precyzyjne określenie prawdopodobieństwo jakiego zdarzenia nas naprawdę interesuje.

### 5.1.3 Poziom $p$

Poziom $p$ jest to wartość prawdopodobińśtwa, że wobec posiadanych danych hipoteza zerowa jest prawdziwa. Najczęściej porównujemy go z zadanym, wcześniej ustalonym poziomem istotności $\alpha$, przy którym możemy odrzucić hipotezę zerową dysponując otrzymaną wartością statystyki testowej.

## 5.2 Formułowanie hipotez

### 5.2.1 Przykład: Napromieniowywanie muszek owocowych

Załóżmy, że wymyśliliśmy metodę napromieniowywania muszek owocowych powodującą taką mutację, że potomstwo ich nie będzie miało jednakowej szansy na bycie samcem lub samiczką. W pierwszych 20 zbadanych przypadkach uzyskujemy 14 samców i 6 samiczek.

**Pytanie naukowe:** Czy wyniki eksperymentu potwierdzają, że nasza metoda zaburza proporcję płci?

Najpierw musimy przetłumaczyć pytanie naukowe na pytanie statystyczne.

**Pytanie statystyczne:** Jakie jest prawdopodobieństwo uzyskania zaobserwowanej próbki jeśli rzeczywista proporcja płci jest 1:1?

Z tego pytania wynikają dwie możliwe hipotezy:

**Hipoteza zerowa:** Nowa metoda nie zaburza proporcji płci 1:1. Zaobserwowana próbka pochodzi z populacji, w której proporcja płci *jest* 1:1

**Hipoteza przeciwna:** Zaobserwowana próbka pochodzi z populacji, w której proporcja płci *nie jest* 1:1.

**Prawdopodobieństwo, które musimy oszacować:** Jakie jest prawdopodobieństwo uzyskania 14 lub więcej jedynek w serii 20 prób, jeśli prawdopodobieństwo jedynki jest 1/2?

1. Oznaczmy 1 — samiec 0 — samiczka.
2. Zróbmy wektor 20 elementowy zawierający 10 zer i 10 jedynek.
3. Wylosujmy ze zwracaniem nowy wektor 20 elementowy. (Jest to nasz model uzyskiwania 20 elementowej próbki z populacji o proporcji 1:1.) Zapamiętajmy ilość jedynek.
4. Powtórzmy poprzedni krok 1000 razy
5. Zróbmy histogram ilości jedynek.

6. Policzmy ile razy zdarzyło sie 14 lub więcej jedynek (to odpowiada 14 lub więcej samców) i dodajmy do tego ilość przypadków gdy mieliśmy 6 lub mniej jedynek (to odpowiada 14 lub więcej samiczek). Wynik podzielmy przez ilość losowań (1000).

Powyższa procedura opisuje test dwustronny. Testu dwustronnego musimy użyć jeśli nie mamy *istotnych* powodów,żeby wierzyć,że nowa metoda działa *jedynie na zwiększenie* szansy pojawienia się samca.

## 5.3 Testowanie hipotez na temat średniej

Firma rozwożąca przesyłki po mieście deklaruje średni czas dostarczenia 28 minut. Przeprowadźmy test tej hipotezy.

$$H_0 : \mu = 28$$
$$H_1 : \mu \neq 28$$

Wybieramy losową próbę 100 przesyłek, mierzymy czas dostarczenia, liczymy średnią z próby $\bar{x} = 31.5$ minut i odchylenie standardowe próby $s = 5$ minut. Konstruujemy przedziały ufności 95 % dla średniej $\mu$. Dla dużej próby możemy użyć rozkładu normalnego

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}} = 31.5 \pm 1.96 * \frac{5}{\sqrt{100}} = 31.5 \pm 0.98 = [30.52 \quad 32.48]$$

Możemy więc być na 95% pewni, że nieznana średnia leży pomiędzy 30.52 a 32.48 a więc na 95% nie leży poza tym przedziałem. Skoro $H_0$ podaje $\mu = 28$ (poza przedziałem), możemy odrzucić tą hipotezę. Jeśli $H_0$ jest prawdziwe, istnieje prawdopodobieństwo 0.05, że skonstruowany przedział nie będzie zawierał $\mu$. Istnieje zatem prawdopodobieństwo 0.05 popełnienia błądu I-go rodzaju. Mówimy, że przeprowadziliśmy test na poziomie istotności 0.05.

Można to zagadnienie rozwiązać w inny sposób. Obliczmy

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = (31.5 - 28)/(5/10) = 7 > z_{\alpha/2} = 1.96$$

więc odrzucamy $H_0$ na poziomie $\alpha = 0.05$.

## 5.4 Testowanie hipotez na temat wariancji

Do testowania hipotez na temat wariancji używamy statystyki chi-kwadrat o $n-1$ stopniach swobody:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

gdzie $\sigma_0^2$ jest wartością wariancji podaną w $H_0$.

### 5.4.1 Przykład

Do produkcji baterii używane są metalowe płytki o średniej średnicy 5mm. Jeśli wariancja średnicy płytki jest nie większa niż $1mm^2$, produkcja jest kontynuowana. Jeśli wariancja przekracza $1mm^2$ proces produkcji trzeba przerwać. Kontroler przeprowadza test: $H_0 : \sigma^2 \leqslant 1$ i $H_1 : \sigma^2 > 1$. Wybiera losową próbę 31 płytek i znajduje wariancję próby $s^2 = 1.62$. Czy daje to podstawy do przerwania produkcji ?

$\chi^2 = 48.6$. Znajdujemy poziom $p$ dla tej wartości $\chi^2$ z 30 stopniami swobody.

Odp. Odrzucamy $H_0$

## 5.5 Błąd drugiego rodzaju. Moc testu.

Błąd II-go rodzaju popełniamy wtedy gdy nie odrzucamy $H_0$ a prawdziwe jest $H_1$.

### 5.5.1 Przykład

Załóżmy następujący test:

$$H_0 : \mu = 60$$

$$H_1 : \mu = 65$$

Niech rozmiar próby wynosi $n = 100$ a odchylenie standardowe w populacji $\sigma = 20$. Powinniśmy tu zastosować test jednostronny (mamy tylko dwie możliwości: $\mu = 60$ lub 65). Znajdźmy punkt krytyczny $C$ dla poziomu istotności $\alpha = 0.05$:

$$C = \mu_0 + 1.645\frac{\sigma}{\sqrt{n}} = 60 + 1.645(20/10) = 63.29$$

Błąd pierwszego rodzaju popełnimy gdy $\bar{x} > C$ i prawdziwe będzie $H_0$. Prawdopodobieństwo błędu pierwszego rodzaju ustaliliśmy z góry na poziomie $\alpha = 0.05$.

$$\alpha = P(\bar{x} > C | \mu = \mu_0)$$

Błąd drugiego rodzaju popełnimy gdy $\bar{x} < C$ a prawdziwe będzie $H_1$. Prawdopodobieństwo popełnienia tego błędu wynosi:

$$\beta = P(\bar{x} < C | \mu = \mu_1) = P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < \frac{C - \mu_0}{\sigma/\sqrt{n}}\right) =$$

$$= P\left(Z < \frac{63.29 - 65}{20/10}\right) = P(Z < -0.855) = 0.1963$$

A moc testu czyli prawdopodobieństwo odrzucenia hipotezy zerowej podczas gdy nie jest ona prawdziwa wynosi $1 - \beta = 0.8037$.

## 5.6 Porównanie dwóch populacji

Dla przypomnienia: Jeśli mamy dwie próbki danych $x_1$ o liczebności $n_1$ i estymowanej wariancji $s_1^2$ i $x_2$ o liczebności $n_2$ i estymowanej wariancji $s_2^2$ pochodzących z rozkładu normalnego o takiej samej wariancji $\sigma$ to:

1. wspólna wariancja może być estymowana jako:

$$\sigma^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

2. wariancja różnicy średnich może być estymowana jako:

$$\sigma_\Delta^2 = s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2 = \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2 = \frac{n_1 + n_2}{n_1 n_2}s^2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_\Delta}$$

pochodzi z rokładu $t$ o $n_1 + n_2 - 2$ stopniach swobody.

### 5.6.1 Przykład

Producent odtwarzaczy CD chce sprawdzić czy małe obniżenie ceny produktu wpłynie korzystnie na sprzedaż. Losowa próba 15 tygodni sprzedaży przed obniżką dała średni dochód 6598 zł i standardowe odchylenie 844 zł. Losowa próba 12 tygodni sprzedaży w trakcie promocji dała średnią 6870 i odchylenie standardowe 669 zł. Czy dane te wykazują poprawę sprzedaży w trakcie promocji ?

Wskazówka: założyć jednakową wariancję w populacji sprzedaży przed i po obniżce. Zastosować test $t$.

Odp. Nie możemy odrzucić $H_0$. Nie mamy podstaw by uznać że mała obniżka cen poprawiła sprzedaż.

### 5.6.2 Do testowania równości wariancji w dwóch populacjach stosuje się test $F$:

$$F_{(n1-1,n2-1)} = \frac{s_1^2}{s_2^2}$$

W przykładzie powyżej założyliśmy równość wariancji. Korzystając z testu $F$ sprawdzić czy założenie było uzasadnione.

### 5.6.3 Karma dla świń

Badamy dwie nowe karmy dla świń; nazwijmy je A i B. Mamy dwie grupy po 12 zwierząt. Uzyskane przyrosty masy są następujące:
A: 31 34 29 26 32 35 38 34 31 29 32 31
B: 26 24 28 29 30 29 31 29 32 26 28 32
Czy któraś z karm daje istotnie większe przyrosty masy?

```
clf
echo on
% Badamy dwie nowe karmy dla świń; nazwijmy je A i B.
% Mamy dwie grupy po 12  zwierząt.
% Uzyskane przyrosty masy są następujące:
A=[ 31 34 29 26 32 35 38 34 31 29 32 31];
B=[26 24 28 29 30 29 31 29 32 26 28 32];
% Czy któraś z  karm daje istotnie większe przyrosty masy?
```

```
% badamy normalnosc rozkladow A i B
subplot(221)
normplot(A)
title('A')
subplot(222)
normplot(B)
title('B')
subplot(223)
boxplot([A' B'],'notch','on')

% Poniewaz nie jest wazne , ktora karma jest
% lepsza budujemy test dwustronny.
% Zakladajac ze dane pochodza z rozkladu normalnego:
[h p_t]=ttest2(A,B)

% nie zakładajac nic o rozkladzie mozemy przetestowac ta hipoteze przy
% pomocy testu Wilcoxona
[p_w h]=ranksum(A,B)

% i na koniec to samo sprawdzimy za pomocš testu reprobkowanego
% zgodnie z hipoteza zerowa probka A i  B pochodza z tej samej populacji.
% Nasza najlepsza wiedza o owej populacji to polaczone probki A i B:
POP=[A B];
N=length(POP);
NA=length(A);
NB=length(B);
% Zasymulujemy N_rep razy wyciagniecie z POP prob o rozmiarach NA i NB i
% zobaczymy jak czesto zdarza sie roznica srednich taka jak w oryginalnym
% pomiarze lub jeszcze wieksza.
N_rep=1e4;
% oryginalna roznica srednich i median:
mi_0=abs(mean(A)-mean(B));
T_0=abs(mean(A)-mean(B))/std(POP);
me_0=abs(median(A)-median(B));
mi=zeros(1,N_rep);
T=zeros(1,N_rep);
me=zeros(1,N_rep);
```

```
echo off
for i=1:N_rep
    AA=randsample(POP,NA,true);
    BB=randsample(POP,NB,true);
    mi(i)=abs(mean(AA)-mean(BB)); % abs bo test dwustronny
    T(i)=abs(mean(AA)-mean(BB))/std([AA BB]);
    me(i)=abs(median(AA)-median(BB));
end
echo on
p_mi=length(find(mi>=mi_0))/N_rep
p_T=length(find(T>=T_0))/N_rep
p_me=length(find(me>=me_0))/N_rep

echo off

subplot(224)
text(0.2, 0.5,sprintf(' p_t=%.3f\n p_T=%.3f\n p_{mi}=%.3f\n p_w=%.3f\n p_{me}=%
axis off
```

## 5.7 Założenie normalności rozkładu

We wszystkich wspomnianych powyżej klasycznych testach statystycznych $t$, $z$, $F$, $\chi^2$ istotnym założeniem jest to, że dane wejściowe w próbie mają rezkład normalny. W powyższych zadaniach po prostu to zakładaliśmy, ale w praktyce, kiedy dostajemy próbę do analizy, musimy sami sprawdzić, czy możemy uznać ją za pochodzącą z rozkładu normalnego. Do weryfikacji takiej hipotezy służą narzędzia graficzne:

- histogram z naniesionym fitem rozkładu normalnego `histfit`
- wykres wartości w próbie, wzg. prawdopodobieństwa uzyskania takiej wartości w rozkładzie normalnym `normplot`

oraz testy nieprametryczne:

- test Lillieforsa ( `lillietest`) — test oparty na badaniu maksymalnej różnicy pomiędzy dystrybuantą empiryczną (z próby) a dystrybuantą rozkładu normalnego o takiej samej średniej i wariancji jak oszacowana z próby

- test Kolmogorova-Smirnova (`kstest` ) — test ten także jest oparty na badaniu maksymalnej różnicy pomiędzy dystrybuantą empiryczną (z próby) a dystrybuantą rozkładu normalnego. Istotna różnica wzg. poprzedniego testu polega na tym, że tu musimy *apriori* skądinąd znać parametry odpowiedniego rozkładu normalnego.

### 5.7.1 Przykład

Proszę wygenerować 100 liczb z rozkładu normalnego. Liczby te obejrzyjmy na wykresach `histfit` oraz `normplot` i zbadamy ich normalność testem Lillieforsa.

```
x = normrnd(0,1,100,1);
figure(1)
subplot(221)
normplot(x);
subplot(222)
histfit(x,15)
[H,P,LSTAT,CV]=lillietest(x);
title(sprintf('h: %d,p: %.2f,lstat: %.2f, cv: %.2f', H,P,LSTAT,CV))
y=x(1:10)
subplot(223)
normplot(y);
subplot(224)
histfit(y,15)
[H,P,LSTAT,CV]=lillietest(y);
title(sprintf('h: %d,p: %.2f,lstat: %.2f, cv: %.2f', H,P,LSTAT,CV))
```

A teraz popsujmy normalność danych:

```
xx=x.^3;
figure(2)
subplot(221)
normplot(xx);
subplot(222)
histfit(xx,15)
[H,P,LSTAT,CV]=lillietest(xx);
title(sprintf('h: %d,p: %.2f,lstat: %.2f, cv: %.2f', H,P,LSTAT,CV))
yy=xx(1:10);
```

```
subplot(223)
normplot(yy);
subplot(224)
histfit(yy,15)
[H,P,LSTAT,CV]=lillietest(yy);
title(sprintf('h: %d,p: %.2f,lstat: %.2f, cv: %.2f', H,P,LSTAT,CV))
```

Proszę zapuścić skrypt kilka razy i zwrócić uwagę na to, jak trudno jest ocenić normalność danych przy małych próbach.

### 5.7.2 Przykład

Często normalność danych można poprawić przez zastosowanie odpowiedniej transformacji.

```
load carsmall
[h p l c] = lillietest(Weight);
[h p l c]
histfit(Weight)
```

Widać, że dane nie są normalne, są mocno skośne w prawą stronę. Przy tym typie odstępstwa od normalności często pomaga logarytmowanie:

```
w=log(Weight);
[h p l c] = lillietest(w);
[h p l c]
histfit(w)
```

Przy mniejszym stopniu skośności można spróbować $\sqrt{\phantom{x}}$. Ogólnie: zanim zaczniemy analizować dane dobrze jest je pooglądać na różnych wykresach i chwilę *pomyśleć*.

## 5.8 Przykłady różne

### 5.8.1 Linie lotnicze

Linie lotnicze, projektując nowy samolot chcą sprawdzić czy średnia waga bagażu ręcznego zabieranego przez pasażerów nie zmieniła się od czasu poprzednich badań i wynosi wciąż 12 kg. Analiza ma być przeprowadzona na

poziomie istotności $\alpha = 0.05$. Analityk pobrał próbę bagażu ręcznego 144 pasażerów i obliczył wartość średnią z próby $\bar{x} = 14.6$ kg i odchylenie standardowe z próby $s = 7.8$. Przeprowadź test hipotezy, że $\mu = 12$.

```
mu_0=12;
a=0.05;
N=144;
x=14.6;
s=7.8;

% test dotyczy sredniej wiec jej std:
std_mu=s/sqrt(N);

% odchylenie std obliczylismy z proby => stosujemy test t
t=(mu_0-x)/std_mu
% test jest dwustronny mamy wiec dwie wartosci krytyczna t :
disp([ tinv(a,N-1) tinv(1-a,N-1)])
%'Odp: Wyliczone t lezy poza obszarem akceptacji hipotezy zerowej,
% zatej odrzucamy hipoteze zerowa i akceptujemy alternatywna.'
```

Odp. Hipotezę można odrzucić na zadanym poziomie istotności.

### 5.8.2 Agencja nieruchomości

Agencja nieruchomości w Japonii podała, że ceny gruntu w centrum Tokio wzrosły o 49% w ciągu ostatniego roku. Inwestor chcąc przetestować te dane, znajduje próbę 18 nieruchomości w centrum Tokio, dla których zna cenę obecna i sprzed roku. Dla każdej nieruchomości oblicza procentowy wzrost wartości a następnie znajduje średnią i odchylenie standardowe z próby. Statystyki próby wynoszą $\bar{x} = 38\%$ i $s = 14\%$. Przeprowadź test na poziomie istotności $\alpha = 0.01$.

Wskazówka: Statystyka próby jest mała a odchylenie standardowe populacji nie znane więc należy skorzystać z rozkładu $t$ o $n - 1 = 17$ stopni swobody.

```
mu_o=49;
x=38;
s=14;
N=18;
```

```
a=0.01;

t=(x-mu_o)/(s/sqrt(N));

% Odp: zaobserwowanie równie malej lub mniejszej wartosci srednieg wzrostu
% cen przy przwadziwej hipotezie zerowej i podanym rozmiarze próbki wynosi

p=tcdf(t,N-1)
% jest to znacznie mniej niz zalozony poziom istotnosci więc odrzucamy
% hipotezę zerowš
```

Odp. Odrzucamy $H_0 : \mu_0 = 49$, na poziomie istotności 0.01.

### 5.8.3 Czy zabiegi bio-inżynieryjne zwiększają częstość narodzin krów?

Załóżmy, że krowy są bardziej wartościowe od byków. Bio-inżynier twierdzi, że przy pomocy pewnych zabiegów jest w stanie spowodować zwiększenie szansy na urodzenie się krowy powyżej 50%. W jego eksperymencie na 10 urodzonych zwierząt 9 było krowami, a tylko 1 bykiem. Czy powinnniśmy wierzyć temu bio-inżynierowi? Jakia jest szansa na uzyskanie takiego, bądź bardziej ekstremalnego wyniku przy założeniu, że procedura stosowana przez naszego inżyniera nia ma żadnych efektów? W tym problemie dla odmiany założymy, że w normalnych warunkach 100 spośród 206 cieląt to krowy. Zadanie rozwiązać metodą parametryczną i przez repróbkowanie.

```
% sposób pierwszy:
% zmienna urodzenie byka/krowy podlega rozkladowi dwumianowemu
p=100/206;
N=10;
p_bino=1-binocdf(8,N,p) %prawdopodobieństwo wylosowania 9 lub 10 krów w 10 prob
% W jeggo mwetodzie chyba cos jest

% sposob drugi: repróbkowanie
% model swiata z ktorego pochodza byki(0)/krowy(1):
w=[ones(1,100) zeros(1,106)];

N_rep=1e5;
```

```
wynik=zeros(1,N_rep);

for i=1:N_rep
    wynik(i)=sum(randsample(w,10,true));
end
p_rep=length(find(wynik>=9))/N_rep
```

### 5.8.4 Porównanie lekarstwa na raka i placebo

Badamy skuteczność leku na raka. Mamy grupę 12 chorych: 6 osobom podajemy lek — poprawa wystąpiła u 5 osób, pozostałym sześciu osobom podajemy placebo — poprawa wystąpiła u 2 osób. Czy te wyniki upoważniają do stwierdzenia, że lek istotnie zwiększa szansę poprawy?

Uwaga: W tym zadaniu porównujemy dwie grupy ze sobą.

- Jaka jest hipoteza zerowa?
- Rozkład jakiej wielkości musimy zbadać?

```
% Zakladamy hipotezy
% H0: lek nie daje poprawy
% H1: lek daje poprawe

% zgodnie z H0 obie proby pochodza ze swiata:
w=[ones(1,7) zeros(1,5)]; % jedynki = wystapila poprawa

% reprobkujemy
N_rep=1e5;

n_l=5; % ilosc popraw w grupie leku
n_p=2; % ilosc popraw w grupie placebo
% jako statystyke testowa przyjmiemy roznice w poprawach miedzy grupami
% w tym problemie istotne jest zwiekszenie ilosci popraw wiec stosujemy test
% jednostronny

st_0= n_l - n_p;
st_rep=zeros(1,N_rep);
for i=1:N_rep
    n_l_rep   = sum(randsample(w,6,true));
```

```
    n_p_rep   = sum(randsample(w,6,true));
    st_rep(i) = n_l_rep - n_p_rep; % wartosc statystyki w i-tym reprobkowaniu
end

% jaka proporcja statystyk reprobkowanych daje wynik taki jak oryginalnie
% lub bardziej ekstremalny?
p=length(find(st_rep>=st_0))/N_rep
```

### 5.8.5 Lek przeciwdepresyjny

Poniższa tabela prezentuje wyniki 9 pacjentów wykonujących pewien test diagnostyczny przed podaniem leku i po podaniu leku.

```
przed po
1.83 0.878
0.50 0.647
1.62 0.598
2.48 2.05
1.68 1.06
1.88 1.29
1.55 1.06
3.06 3.14
1.3 1.29
```

Skonstruować test, który pozwoli stwierdzić czy lek jest skuteczny. Porównać różne wersje testu bootstrapową (losowanie z powtórzeniami), permutacyjną, test parametryczny i nieparametryczny. Jakie założenia przyjmujemy przy każdej z wersji testu?

Uwaga: w tym zadaniu mamy dwie grupy "przed" i "po" ale oprócz tego istnieje ścisły porządek w parach, bez sensu jest porównywanie "przed" od jednego pacjenta z "po" drugiego pacjenta. Musimy stosować testy, które biorą ten porządek pod uwagę (testy pairwise).

```
%przed po
A=[1.83 0.878
0.50 0.647
1.62 0.598
2.48 2.05
```

```
1.68 1.06
1.88 1.29
1.55 1.06
3.06 3.14
1.3 1.29];
%Skonstruować test, który pozwoli stwierdzić czy lek jest skuteczny.
%Porównać dwie wersje testu bootstrapową (losowanie z powtórzeniami) i permutac

% Jako miare tego czy lek jest skuteczny przyjmiemy różnicę (po-przed)
% Bedziemy wierzyli ze lek dziala jesli ta roznica jest istotnie rozna od
% zera

r=A(:,2)-A(:,1);
disp('srednia roznica:')
mr=mean(r);
disp(mr)
% H0: wszystko jedno ktory pomiar jest przed a ktory po (r_srednie równe
% zero)
% H1: r_srednie różne zero


% Musimy wytwarzac symulowane r zgodnie z hipoteza zerowa, ze nie ma
% znaczenia która wartosc jest przed, a która po. trzeba tylko uwarzac żeby
% nie pomieszac pacjentow

% wersja bootstrpowa: bierzemy pacjentow z powtorzeniami - zakładamy
% reprezentatywna grupa bardzo duzej populacji

N_rep=1e4;
r_boot=zeros(1,N_rep);
N=length(r);
for i=1:N_rep
    ix=randsample(N,N,true); % wybieramy pacjentow z powtorzeniami
    B=A(ix,:);
    for j=1:N   % mieszamy losowo przypisujemy przypadki do grupy przed i po
        s=rand(1,1);
        if s>0.5
            przed(j)=B(j,1);
```

```
            po(j)=B(j,2);
        else
            przed(j)=B(j,2);
            po(j)=B(j,1);
        end
    end
    rr=po-przed;
    r_boot(i)=mean(rr);
end
subplot(221)
hist(r_boot,30)
disp('dla repróbkowanego testu dwustronnego: ')
p_h0=length(find(abs(r_boot)>abs(mr)))/N_rep;
disp(sprintf('p_H0: %.3f',p_h0))
title(sprintf('test repróbkowany p_{H0}: %.3f',p_h0))


% wersja permutacyjna: korzystamy za każdym razem ze wszystkich dostepnych
% pacjentow

N_rep=1e4;
r_boot=zeros(1,N_rep);
N=length(r);
for i=1:N_rep
    for j=1:N   % mieszamy losowo przypisujemy przypadki do grupy przed i po
        s=rand(1,1);
        if s>0.5
            przed(j)=A(j,1);
            po(j)=A(j,2);
        else
            przed(j)=A(j,2);
            po(j)=A(j,1);
        end
    end
    rr=po-przed;
    r_boot(i)=mean(rr);
end
subplot(222)
```

```
hist(r_boot,30)
disp('dla permutacyjnego testu dwustronnego: ')
p_h0=length(find(abs(r_boot)>abs(mr)))/N_rep;
disp(sprintf('p_H0: %.3f',p_h0))
title(sprintf('test permutacyjny p_{H0}: %.3f',p_h0))

% jesli mielibysmy jakies przeslanki, zeby zalozyc normalnosc naszych
% danych wejsciowych to mozna zastosowac test t
disp('dla parowanego testu t dwustronnego: ')
subplot(223)
normplot(r)
[h, p]=ttest(A(:,2),A(:,1))
title(sprintf('parowany test t p_{H0}: %.3f',p))

subplot(224)
axis
axis off
disp('Testy nieparametryczne testuja czy mediana roznicy jest 0')
p_st = signtest(A(:,2),A(:,1)); % zal: dane pochodza z dowolnej ciaglej dystryb
s1=sprintf(' test znkow p_{H0}: %.3f',p_st)
p_sr = signrank(A(:,2),A(:,1)); % zal: dane pochodza z dowolnej ciaglej symetry
s2=sprintf(' test rang p_{H0}: %.3f',p_sr)
str(1)={'Testy nieparametryczne testuja'};
str(2)={'czy mediana roznicy jest 0'};
str(3)={' '};
str(4)={s1};
str(5)={s2};
text(0.1, 0.7,str)
```

## 5.9 Zadania

### 5.9.1 Zanieczyszczenie środowiska

Agencja ochrony środowiska ustaliła limit na koncentrację zanieczyszczeń emitowanych przez fabryki. Załóżmy, że dopuszczalny poziom zanieczyszczeń wynosi: 55 cząstek na milion (cz/m) w promieniu dwóch kilometrów od fabryki. Kontrola przeprowadza 100 pomiarów o różnej porze dnia i roku w

promieniu dwóch km. od pewnej fabryki. Średnia z próby wyniosła 60 cz/m a odchylenie standardowe $s = 20$ cz/m. Czy dane te są wystarczające by uznać, że fabryka łamie prawo ?

Fabryka łamie prawo jeśli emituje zanieczyszczenia na poziomie wyższym niż dopuszczalny więc należy przeprowadzić test jednostronny (w tym przypadku prawostronny).

Odp. Możemy odrzucić hipotezę $H_0$ (głoszącą, że fabryka nie łamie prawa) na poziomie $\alpha = 0{,}01$.

Czy moglibyśmy odrzucić $H_0$ na tym samym poziomie stosując test dwustronny? Jest ważne aby w zależności od problemu wybrać odpowiedni test: jedno- lub dwustronny.

### 5.9.2 Wzrost mocy turbiny

Turbina hydroelektryczna generuje moc średnią 25.2 kW. Po unowocześnieniu maszyny chcemy przetestować czy średnia moc generowana się zmieniła (na + lub -). Przeprowadzono 115 pomiarów, które dały średnią 26.1 kW i odch. std. 3.2 kW. Przeprowadzić test statystyczny, znaleźć poziom $p$ i zinterpretować wynik.

Odp.: Poziom $p = 0.0026$. Oznacza to, że hipoteza $H_0$ jest bardzo mało prawdopodobna i możemy ją odrzucić.

### 5.9.3 Sonda

Władze miasta chciałyby wiedzieć, czy przyznać koncesję operatorowi sieci kablowej. W tym celu zleciły nam przeprowadzenie sondy wśród mieszkańców. Zapytaliśmy o zdanie 50 przypadkowo wybranych osób. 30 osób powiedziało "tak" a 20 "nie". Na ile pewnie otrzymane wyniki wskazują, że mieszkańcy chcą tej kablówki?

Celem naszych badań jest uniknięcie błędu polegającego na tym, że powiemy iż większość miezkańców chce kablówki podczas gdy tak na prawdę to nie chce.

Wskazówka: Granicznym przypadkiem popełnienia tego błędu jest proporcja 1:1 zwolenników i przeciwników kablówki. Jeśli przeciwników kablówki byłoby jeszcze więcej to uzyskanie naszych wyników byłoby jeszcze mniej prawdopodobne.

### 5.9.4 Wybory prezydenckie

W ankiecie uzyskaliśmy 840 głosów popierających kandydaturę A i 660 kandydaturę B. Jaka jest szansa, że tak naprawdę kandydat B ma poparcie 50% lub większe? Jakie jest prawdopodobieństwo pojawienia sie zaobserwowanej próbki lub próbki wskazującej na jeszcze większe poparcie dla kandydata A, jeśli w rzeczywistości poparcie kandydata A byłoby 50% lub mniej.

### 5.9.5 Czy stosunek do marihuany się zmienił?

Rozważmy dwie ankiety przeprowadzone w USA, pytano 1500 respondentów o stosunek do legalizacji marihuany. Pierwszą ankietę przeprowadzono w 1980, wówczas za legalizacją opowiadało się 52% a drugą w 1985 i za legalizacją było 46%. Czy wyniki tych dwóch ankiet są istotnie różne?

Z jaką proporcją powinniśmy porównywać te wyniki? Jaka jest hipoteza zerowa?

### 5.9.6 Zawały serca i cholesterol

Badano grupę 605 osób. 135 osób z tej gurpy miało wysoki poziom cholesterolu a 470 niski. W grupie z wysokim poziomem cholesterolu odnotowano 10 przypadków zawału serca a w grupie z niskim poziomem 21, w czasie 16 lat obserwacji. Nasze pytanie brzmi: Czy możemy uznać, że wysoki poziom cholesterolu zwiększa ryzyko zawału serca? Innymi słowy: czy możemy założyć, że obie grupy pochodzą z tej samej "populacji"?

### 5.9.7 Czy gęstości planet się różnią?

Rozważmy pięć planet znanych w antycznym świecie. Chcemy zbadać, czy planety wewnętrzne Merkury (0.68) i Wenus (0.94) mają istotnie większe gęstości niż planety zewnętrzne Mars(0.71) Jowisz (0.24) i Saturn(0.12)?

# Rozdział 6

# Porównywanie więcej niż dwóch grup

## 6.1 Problem wielokrotności testów

Rozważmy przykład. Wielokrotnie losujemy po dwie próby z tego samego rozkładu normalnego i badamy testem $t$ czy średnie są jednakowe.

```
N=10000;
h=zeros(1,N);
p=zeros(1,N);
for i=1:N
    x=randn(1,100);
    y=randn(1,100);
    [h(i),p(i)]=ttest2(x,y);
end
hist(p,30);
frakcja_odrzucona=sum(h)/N
```

Czy wynik jest zgodny z oczekiwaniami? Co z tego wynika?

## 6.2 ANOVA

Celem analizy wariancji jest zbadanie czy dane pochodzące z kilku grup mają tą samą średnią. ANOVA zakłada, że dane dają się opisać następującym

modelem liniowym:

$$y_{i,j} = \alpha_{.,j} + \epsilon_{i,j}$$

gdzie:

- $y_{i,j}$ — macierz obserwacji, każda kolumna $j$ odpowiada jednej grupie, wiersze odpowiadają przypadkom $i$
- $\alpha_{.,j}$ — macierz średnich, notacja $_{.,j}$ oznacza , że dla wszystkich przypadków w grupie $j$ mamy tą samą wartość średniej
- $\epsilon_{i,j}$ — macierz przypadkowych zaburzeń

W tym języku celem ANOVy jest zbadanie hipotezy zerowej, że wszystkie $\alpha_{.,j}$ są równe. Zakładamy, że $\epsilon_{i,j}$ są **niezależnymi** zmiennymi losowymi pochodzącymi z **tego samego** rozkładu **normalnego**

### 6.2.1 Przykład

Jako przykład wykorzystamy dane z matlabowego pliku `hogg`. Dane dotyczą ilości bakterii w różnych dostawach mleka. Załadujmy i obejrzyjmy dane:

```
load hogg
hogg
```

Wykonajmy analizę:

```
[p,tbl,stats] = anova1(hogg);
```

Pojawiło nam się okno z tabelką i z boxplotami. Boxploty przydają się aby graficznie zweryfikować czy średnie pomiędzy grupami są różne i ocenić czy wariancje w grupach są podobne. Tabelka przedstawia w standardowy sposób (większość pakietów statystycznych rysuje taką tabelkę) wyniki. Jak ją interpretować?

| Source | SS — składowe wariancji | df — ilość stopni swobody | MS — wariancja na stopień swobody | F — wartość statystyki | Prob >F — prawdopodobieństwo zaobserwowania wartości statystyki F równie lub bardziej ekstremalnej niż otrzymana w sąsiedniej rubryczce |
|---|---|---|---|---|---|
| Columns | Wariancja pomiędzy grupami ($s^2_{pom}$) | ilość grup - 1 $k-1$ | $\frac{s^2_{pom}}{k-1}$ | $\frac{\frac{s^2_{pom}}{k-1}}{\frac{s^2_{wew}}{n-k}}$ | |
| Error | Wariancja niewytłumaczona wariancją międzygrupową ($s^2_{wew}$) | ilość obserwacji -ilość grup $n-k$ | $\frac{s^2_{wew}}{n-k}$ | | |
| Total | Całkowita wariancja | ilość obserwacji -1 $(n-1)$ | | | |

Dane z pliku hogg:

```
hogg =

    24    14    11     7    19
    15     7     9     7    24
    21    12     7     4    19
    27    17    13     7    15
    33    14    12    12    10
    23    16    18    18    20
```

Obliczenia, które trzeba wykonać:

```
m_kolumny=mean(hogg)

m_kolumny =

   23.8333   13.3333   11.6667    9.1667   17.8333

m_total=mean(mean(hogg))
```

```
m_total =

   15.1667
```

Suma wariancji pochodzącej od tego, że kolumny nie mają takiej samej średniej:

```
[N_wierszy N_kolumn]=size(hogg);
ss_miedzykolumnami=sum(N_wierszy*(m_total-m_kolumny).^2)

ss_miedzykolumnami =

  803.0000
```

Stopni swobody jest o jeden mniej niż kolumn. średnio na st. swobody przypada więc $803/4 = 200.75$

Całkowita wariancja wewnątrz-grupowa

```
ss_wewnatrzkolumn=sum((hogg(:,1)-m_kolumny(1)).^2)+...
sum((hogg(:,2)-m_kolumny(2)).^2)+...
sum((hogg(:,3)-m_kolumny(3)).^2)+...
sum((hogg(:,4)-m_kolumny(4)).^2)+...
sum((hogg(:,5)-m_kolumny(5)).^2)
```

Ilość st.swobody jest mniejsza od ilości przypadków o ilość kolumn. średnio na st. swobody przypada więc: $557.16/25 = 22.28$

Przy założeniu hipotezy $H_0$ liczba $200.75/22.28 = 9.01$ pochodzi z rozkładu $F(4, 25)$. Prawdopodobieństwo zobaczenia tatiej liczby w powyższym rozkładzie jest 0.0001 czyli dość małe. Przy poziomie istotności testu 5% odrzucamy więc $H_0$.

### 6.2.2 Które średnie są różne?

Jeśli wynik tego testu pozwala nam odrzucić hipotezę zerową o równości średnich to rodzi się pytanie które średnie są różne. Jest wiele testów które na tym etapie można wykonać aby odpowiedzieć na to pytanie. W matlabie jest zaimplementowany test oparty na badaniu przedziałów ufności dla średnich.

```
[c,m]=multcompare(stats)
```

`stats` jest macierzą zwracaną przez `anova1` Macierz `c` zawiera następujące informacje: numer pierwszej grupy, numer drugiej grupy, różnica między średnimi tych dwóch grup, następne trzy kolumny to dolny brzeg przedziału ufności (95%), różnica średnich , górny brzeg przedziału ufności na tę różnicę. Macierz `m` w kolejnych wierszach zawiera średnie i błędy średnich dla poszczególnych grup. Te informacje są także prezentowane w oknie graficznym otwieranym przez funkcję `multcompare`

### 6.2.3 Przykład

Testowano skuteczność trzech leków, każdy na grupie 10 myszy. Uzyskano następujące wartości wskaźnika skuteczności (masa śledziony poddana transformacji log — w celu poprawienia normalności danych):

1. 19 45 26 23 36 23 26 33 22 29
2. 40 28 26 15 24 26 36 27 28 19
3. 32 26 30 17 23 24 29 20 27 19

Przeprowadzić analizę wariancji. Czy są podstawy aby sądzić, że leki mają różną skuteczność?

```
G1=[19 45 26 23 36 23 26 33 22 29];
G2=[40 28 26 15 24 26 36 27 28 19];
G3=[32 26 30 17 23 24 29 20 27 19];
X=[G1' G2' G3'];
GR={'G1' 'G2' 'G3'};
anova1(X,GR)
```

Wynik: ANOVA Table

| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Columns | 62.6 | 2 | 31.3 | 0.67684 | 0.51664 |
| Error | 1248.6 | 27 | 46.2444 | | |
| Total | 1311.2 | 29 | | | |

Nie ma podstaw aby odrzucić hipotezę, że wszystkie leki mają podobną skuteczność.

## 6.3 Dwu czynnikowa analiza wariancji

Cel tej analizy (ang Two-way ANOVA) to także zbadanie, czy średnie pomiędzy grupami danych są jednakowe. W tym przypadku dane są jednak grupowane przez dwa czynniki. Odpowiada to następującemu modelowi liniowemu:

$$y_{i,j,k} = \mu + \alpha_{.,j} + \beta_{i,.} + \gamma_{i,j} + \epsilon_{i,j,k}$$

gdzie:

- $y_{i,j,k}$ — macierz obserwacji, $j$ odpowiada grupie charakteryzowanej przez czynnik 1 ,$k$ odpowiada grupie charakteryzowanej przez czynnik 2, $i$ odpowiada za przypadki
- $\mu$ — średnia po wszystkich przypadkach
- $\alpha_{.,j}$ — macierz odchyleń do średniej $\mu$ które można przypisać czynnikowi 1
- $\beta_{i,.}$ — macierz odchyleń do średniej $\mu$ które można przypisać czynnikowi 2
- $\gamma_{i,j}$ — macierz interakcji
- $\epsilon_{i,j,k}$ — macierz przypadkowych zaburzeń

### 6.3.1 Przykład

Mamy dwie fabryki i każda z nich robi te same trzy modele samochodów. Badamy czy różne modele samochodów zrobione w różnych fabrykach różnią się ilością mil przejechanych na jednym galonie benzyny:

```
load mileage
mileage

mileage =

   33.3000   34.5000   37.4000
   33.4000   34.8000   36.8000 Fabryka 1
   32.9000   33.8000   37.6000
-----------------------------------------------------------------
```

```
   32.6000   33.4000   36.6000
   32.5000   33.7000   37.0000 Fabryka 2
   33.0000   33.9000   36.7000

   model  1  model 2    model 3

cars = 3;

[p, tbl, stats]=anova2(mileage, cars);
p
    0.0000    0.0039    0.8411
```

- Kolumny (modele) mają `p = 0`, — odrzucamy hipotezę zerową, że modele nie różnią się. Możemy teraz wykonać `multcompare(stats)` i zobaczyć które średnie się różnią
- W rzędach ( fabryki) mamy`p=0.0039` — wygląda na to, że jedna z fabryk robi wyraźnie lepsze samochody
- Dla czynnika interakcji mamy `p=0.8411` — nie mamy powodu aby powiedzieć , że pewien model jest wykonywany w jednej z fabryk lepiej niż w drugiej a inny gorzej

W tej implementacji `anova2` wymaga zbalansowanych danych tzn. ilość obserwacji dla każdej kombinacji czynników jest taka sama.

Łatwo można sobie wyobrazić rozszerzenie tej metodologii na dowolną ilość przypadków (N-way ANOVA) i na niezbalansowane dane — jest to zaimplementowane w funkcji `anovan` .

## 6.4 Nieparametryczne odpowiedniki ANOVY

### 6.4.1 Test Kruskala-Wallisa

Jest to nieprametryczny odpowiednik ANOVy z jednym czynnikiem. Test oparty jest na badaniu rang zamiast samych danych. Zakładamy jedynie, że dane pochodzą z rozkładu ciągłego.

```
load hogg
p=kruskalwallis(hogg)
```

Wyjście wygląda analogicznie jak dla `anova1`. Statystyką testową jest w tym przypadku $\chi^2$

Kruskal-Wallis ANOVA Table

| Source | SS | df | MS | Chi-sq | Prob>Chi-sq |
|---|---|---|---|---|---|
| Columns | 1302.25 | 4 | 325.563 | 16.91 | 0.002 |
| Error | 930.75 | 25 | 37.23 | | |
| Total | 2233 | 29 | | | |

### 6.4.2 Test Friedmana

Jest to odpowiednik dwuczynnikowej analizy wariancji, z tym że nie pozwala na testowanie istotności czynnika interakcji. Można testować jedynie czynniki główne i to w dodatku każdy osobno. Zakładamy w nim, że dane pochodzą z tej samej populacji i że wszystkie *obserwacje są niezależne*. Dla danych z samochodami:

```
load mileage
mileage =

   33.3000   34.5000   37.4000
   33.4000   34.8000   36.8000
   32.9000   33.8000   37.6000
   32.6000   33.4000   36.6000
   32.5000   33.7000   37.0000
   33.0000   33.9000   36.7000

p=friedman(mileage,3)
p =

   7.4659e-04
```

i dostajemy tabelkę analogiczną do ANOVy. W ten sposób przetestowaliśmy istotność czynnika model (kolumny) na zużycie benzyny. Wynika stąd, że modele różnią się średnim zużyciem paliwa.

Żeby zrobić test dla fabryk trzeba przetransformować dane:

```
x=reshape(mileage, [3 2 3])
x=permute(x, [1 3 2])
```

```
x=reshape(x,[9 2])

x =

   33.3000   32.6000
   33.4000   32.5000
   32.9000   33.0000
   34.5000   33.4000
   34.8000   33.7000
   33.8000   33.9000
   37.4000   36.6000
   36.8000   37.0000
   37.6000   36.7000
p=friedman(x,3)

p =

    0.0082
```

Czyli średnie zużycie paliwa w samochodach (nie rozróżniając modeli) produkowanych przez te dwie fabryki jest różne.

## 6.5 Jeszcze inaczej: repróbkowanie

### 6.5.1 Przykład: Napromieniowywanie muszek owocowych, ciąg dalszy

Załóżmy, że testujemy 4 metody napromieniowywania muszek owocowych potencjalnie powodujące taką mutację, że potomstwo ich nie będzie miało jednakowej szansy na bycie samcem lub samiczką. Badamy serię 20 pierwszych osobników z każdej z metod. Uzyskujemy następujące wwyniki:

- W metodzie A: 14 samców i 6 samiczek.
- W metodzie B: 10 samców 10 samiczek
- W metodzie C: 12 samców 8 samiczek

- W metodzie D: 9 samców 11 samiczek

Metody B, C i D na pierwszy rzut oka nie rokują nadziei. Istnieje więc pokusa, żeby je odrzucić i zająć się analizą metody A, redukując problem do rozważanego w zeszłym tygodniu. Jak sie zaraz przekonamy jest to bardzo błędny sposób myślenia.

**Hipoteza zerowa:** Nowe metody nie zaburzają proporcji płci 1:1. Zaobserwowana próbka pochodzi z populacji, w której proporcja płci *jest* 1:1. (Uwaga: tym razem zaobserwowaną próbkę stanowią 4 zestawy po 20 osobników – przy założeniu hipotezy zerowej każdy jest ciągnięty z tej samej populacji z proporcją płci 1:1)

**Hipoteza przeciwna:** Zaobserwowana próbka pochodzi z populacji, w której proporcja płci *nie jest* 1:1.

1. Oznaczmy 1 — samiec 0 — samiczka.
2. Zróbmy wektor 20 elementowy zawierający 10 zer i 10 jedynek.
3. Wylosujmy ze zwracaniem **cztery** nowe wektory po 20 elementów. Jeśli w którymkolwiek z wektorów występuje 14 lub więcej jedynek albo 6 lub mniej jedynek to zapamiętujemy sukces.
4. Powtórzmy poprzedni krok 1000 razy
5. Zróbmy histogram ilości sukcesów
6. Policzmy ile razy zdarzyło się sukcesy. Wynik podzielmy przez ilość losowań (1000).

Jako istotny powód dla odrzucenia hipotezy zerowej przyjmujemy to, że przy jej założeniu otrzymany wynik jest *zaskakujący, mało prawdopodobny.* Zauważmy, że wykonywanie jednocześnie kilku testów istotnie zmienia to jaki wynik jest dla nas *bardzo zaskakujący.* Powyższa procedura opisuje test dwustronny. Testu dwustronnego musimy użyć jeśli nie mamy *istotnych* powodów,żeby wierzyć,że nowa metoda działa *jedynie na zwiększenie* szansy pojawienia się samca.

## 6.6 Zadania

### 6.6.1 Tymidyna a rak

Dwie grupy szczurów otrzymywały zastrzyki tymidyny zawierające ślady trytu. Dodatkowo jednej grupie podano substancję rakotwórczą. Badanie wchłaniania tymidyny przez skórę szczurów w funkcji czasu przeprowadza się zliczając rozpady jąder trytu na jednostkę powierzchni skóry. Po transformacji logarytmicznej ($x = 50 \log x' - 100$) uzyskano następujące liczby:

| Nr obserwacji | Zastrzyki | czas po wstrzyknięciu w godzinach | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 4 | 8 | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 48 |
| 1 | tymidyna | 34 | 54 | 44 | 51 | 62 | 61 | 59 | 66 | 52 | 52 |
| 2 | | 40 | 57 | 52 | 46 | 61 | 70 | 67 | 59 | 63 | 50 |
| 3 | | 38 | 40 | 53 | 51 | 54 | 64 | 58 | 67 | 60 | 44 |
| 4 | | 36 | 43 | 51 | 49 | 60 | 68 | 66 | 58 | 59 | 52 |
| 1 | tymidyna | 28 | 23 | 42 | 43 | 31 | 32 | 25 | 24 | 26 | 26 |
| 2 | + substancja | 32 | 23 | 41 | 48 | 45 | 38 | 27 | 26 | 31 | 27 |
| 3 | rakotwórcza | 34 | 29 | 34 | 36 | 41 | 32 | 27 | 32 | 25 | 27 |
| 4 | | 27 | 30 | 39 | 43 | 37 | 34 | 28 | 30 | 26 | 30 |

Jakie są wnioski z analizy wariancji ?

```
t1=[34 54 44 51 62 61 59 66 52 52];
t2=[40 57 52 46 61 70 67 59 63 50];
t3=[38 40 53 51 54 64 58 67 60 44];
t4=[36 43 51 49 60 68 66 58 59 52];
tr1=[28 23 42 43 31 32 25 24 26 26];
tr2=[32 23 41 48 45 38 27 26 31 27];
tr3=[34 29 34 36 41 32 27 32 25 27];
tr4=[27 30 39 43 37 34 28 30 26 30];

T=[t1 t2 t3 t4];
TR=[tr1 tr2 tr3 tr4];
X=[T TR];
czas=[4 8 12 16 20 24 28 32 36 48 ];
GR_czas=[czas czas czas czas czas czas czas czas];
for i=1:40
```

```
GR_subs{i}='t';
end
for i=41:80
GR_subs{i}='tr';
end

GR={GR_czas GR_subs }
anovan(X,GR,'model','interaction','sstype',3,'varnames',{'czas';'substancja'})
```

### 6.6.2 Czy metody resocjalizacyjne różnią się?

Chcemy zbadać czy cztery metody resocjalizacji w istotny sposób dają różne wyniki. Badamy każdą na grupie 20 młodocianych przestępców. Jako wynik negatywny traktujemy ilość osób, które po terapii ponownie popełniły przestępstwa. Otrzymane wyniki są następujące:

- A: 3
- B: 10
- C: 10
- D: 13

Pytanie: czy któraś z terapii jest lepsza?
Wskazówka: Hipoteza 0: Wszystkie wyniki pochodzą z tej samej populacji — nasza najlepsza wiedza o tej populacji to proporcja ponownych przestępców (3+10+10+13)/80. Możemy badać największą różnicę pomiędzy czterema próbami lub odstępstwo od średniej każdej z prób osobno.

Odp: p≈0.01

### 6.6.3 Efekty łączenia firm: porównywanie danych parowanych

Poniższe dane przedstawiają wyniki finansowe uzyskiwane przez 33 zestawy firm. Pierwszy typ to firmy po połaczeniu, drugi to firmy o rozmiarze zbliżonym do rozmiaru firm, które uległy połączeniu, trzeci to firmy o rozmiarze zbliżonym do rozmiaru firmy powstałej po połączeniu.

Nr: | połączone | nie łączone mniejsze | nie łączone większe

```
 1 -0.2000     0.0256          0
 2  -0.3483    -0.1250     0.0805
 3 0.0751     0.0632    -0.0231
 4 0.1261    -0.0420     0.1647
 5 -0.1017     0.0800     0.2778
 6  0.0378     0.1491     0.4302
 7 0.1162     0.1518     0.1429
 8 -0.0984     0.0377     0.0400
 9 0.0214     0.0766     0.0111
10 -0.0171     0.2843     0.1891
11  -0.3648     0.1391     0.0389
12  0.0881     0.0387     0.0948
13 -0.2632     0.0564     0.0451
14 -0.0494     0.0537     0.0083
15 0.0115     0.0481     0.0948
16 0.0097     0.1982     0.0609
17 0.0714     0.4208    -0.0248
18  0.0018    0.0743     0.0532
19 0.0048    -0.0071     0.0501
20 -0.0540    0.1715    0.1095
21 0.0227     0.0279    -0.0225
22  0.0598    0.0486     0.1671
23   -0.0599     0.0264     0.0207
24 -0.0886    -0.0593     0.0771
25 -0.0248    -0.0184     0.0596
26 0.0764     0.0126     0.0346
27 -0.0017    -0.0455     0.0536
28 -0.2198     0.3431     0.0428
29 0.3824     0.2210     0.1158
30 -0.0068     0.2549     0.2370
31 -0.1630     0.0112    0.1905
32 0.1918     0.1505     0.1520
33  0.0612     0.1704     0.0935
```

Chcemy zweryfikować, czy łączenie ma istotny wpływ na obniżenie wyników finansowych.

Rozwiązanie przez Monte-Carlo: Przekształcić dane do postaci rang. Policzyć sumę rang w każdym typie firmy. Jako statystykę testową użyć różnicy

między maksymalną i minimalną sumą rang. Odp: p≈0.004

Inna statystyka testowa: suma kwadratów rang Odp: p≈0.001

Który test nieparametryczny można tu zastosować?

### 6.6.4 Czy lekarstwo działa?

Badano czy lek A zapobiega zbyt niskiej wadze narodzeniowej.

```
grupa leczona grupa kontrolna
  6.9  6.4
7.6  6.7
  7.3  5.4
7.6  8.2
6.8  5.3
7.2  6.6
8.0  5.8
5.5  5.7
5.8  6.2
7.3  7.1
8.2  7.0
6.9  6.9
6.8  5.6
5.7  4.2
8.6 6.8
Średnia: 7.08 6.26
```

Czy przedstawione wyniki potwierdzają, że w grupie leczonej wagi narodzeniowe są wyższe niż w grupie kontrolnej? Skonstruować test permutacyjny (mieszamy losowo przynależność do grup) i bootstrapowy (losujemy z powtórzeniami ze wspólnego *universum*).

Odp: p≈0.01

### 6.6.5 Karma dla świń raz jeszcze

Badamy cztery nowe karmy dla świń; nazwijmy je A, B, C, D. Mamy dwie grupy po 12 zwierząt. Uzyskane przyrosty masy są następujące:
A: 31 34 29 26 32 35 38 34 31 29 32 31
B: 26 24 28 29 30 29 31 29 32 26 28 32
C: 30 30 32 31 29 27 25 30 31 32 34 33

D: 32 25 31 26 32 27 28 29 29 28 23 25

Czy któraś z karm daje istotnie większe przyrosty masy? Najpierw musimy się upewnić, czy nie można założyć, że wszystkie karmy dają takie same przyrosty. Dopiero po odrzuceniu tej możliwości sensowne staje się pytanie o to, która karma jest lepsza.

Odp: Dla zbiorczej H0 p=0.0055. Karma A jest lepsza niż B i D;

# Rozdział 7

# Modele liniowe

## 7.1 Efekty jednego czynnika w różnych grupach

Mamy sytuację, w której badamy efekt $y$ pewnego czynnika $x$ w różnych grupach. Zakładamy, że sensowny jest liniowy związek między $x$ a $y$, tzn. możemy ten związek wyrazić jedynym z modeli:

**ta sama średnia**

$$y = \alpha + \epsilon$$

**różne średnia**

$$y = \alpha + \alpha_i + \epsilon$$

**to samo nachylenie (korelacja)**

$$y = \alpha + \beta x + \epsilon$$

**różne średnie ale to samo nachylenie (proste równoległe)**

$$y = (\alpha + \alpha_i) + \beta x + \epsilon$$

**różne proste**

$$y = (\alpha + \alpha_i) + (\beta + \beta_i)x + \epsilon$$

Do dopasowanie i analizy takich modeli służy analiza kowariancji.

### 7.1.1 Przykład

Zbadamy jak masa samochodu wpływa na zużycie paliwa i czy ta zależność jest taka sama dla różnych roczników.

Poniższa komenda powoduje, że `aoctool` dopasowuje odrębną linię do wektorów kolumnowych `Weight` i `MPG` dla każdej z grup zdefiniowanych przez `Model_Year`. W poniższym wywołaniu zmienną zależną jest `MPG`, a niezależną `Weight`.

```
load carsmall
[h,atab,ctab,stats] = aoctool(Weight,MPG,Model_Year);
```

Pytania:

1. Jak odczytać dopasowane modele?
2. Jakie są odchylenia standardowe dla dopasowanych parametrów?
3. Dopasowane nachylenia są dość podobne, czy są istotnie różne?
4. Jakie są przedziały ufności dla dopasowanych prostych (dla śrenich wartości obserwacji)?
5. Jakie są przedziały ufności dla obserwacji ?

### 7.1.2 Przykład: Rozmiary żołędzi

W USA rośnie 50 gatunków dębów. W badaniach wzięto pod uwagę 28 gatunków rosnących w regionie atlantyckim i 11 gatunków z regionu kalifornijskiego. Ineresującą kwestią jest to, czy średnie rozmiary żołędzi (objętości) są związane z regionem, z którego pochodzą? Czy zasięg występowania danego gatunku jest związany z rozmiarami żołędzi?

Proszę obejrzeć histogramy danych, jeśli trzeba zastosować transformację danych w celu poprawienia normalności, a następnie zbadanie hipotez:

1. w obu regionach średnie rozmiary żołędzi są takie same
2. zasięgi występowania drzew mają taki sam związek z rozmiarami żołędzi w obu regionach

Nazwy zmiennych:

1. Gatunek: łacińska nazwa gatunku
2. Region: atlantycki lub kalifornijski
3. Zasięg: Powierzchnia na której występuje gatunek w $km^2 \times 100$
4. Rozmiar: rozmiar żołędzia $cm^3$
5. Wysokość: wysokość drzewa $m$

**Dane:**

```
Gatunek Region Zasięg Rozmiar Wysokość
Quercus alba L. Atlantic 24196 1.4 27
Quercus bicolor Willd.   Atlantic 7900 3.4 21
Quercus macrocarpa Michx. Atlantic 23038 9.1 25
Quercus prinoides Willd. Atlantic 17042 1.6 3
Quercus Prinus L. Atlantic 7646 10.5 24
Quercus stellata Wang. Atlantic 19938 2.5 17
Quercus virginiana Mill Atlantic 7985 0.9 15
Quercus Michauxii Nutt. Atlantic 8897 6.8 .30
Quercus lyrata Walt. Atlantic 8982 1.8 24
Quercus Laceyi Small. Atlantic 233 0.3 11
Quercus Chapmanii Sarg. Atlantic 1598 0.9 15
Quercus Durandii Buckl. Atlantic 1745 0.8 23
Quercus Muehlenbergii Engelm Atlantic 17042 2.0 24
Quercus ilicifolia Wang. Atlantic 4082 1.1 3
Quercus incana Bartr. Atlantic 3775 0.6 13
Quercus falcata Michx. Atlantic 13688 1.8 30
Quercus laevis Walt. Atlantic 3978 4.8 9
Quercus laurifolia Michx. Atlantic 5328 1.1 27
Quercus marilandica Muenchh. Atlantic 18480 3.6 9
Quercus nigra L. Atlantic 10161 1.1 24
Quercus palustris Muenchh. Atlantic 8643 1.1 23
Quercus Phellos L. Atlantic 9920 3.6 27
Quercus rubra L. Atlantic 28389 8.1 24
Quercus velutina Lam. Atlantic 21067 3.6 23
Quercus imbricaria Michx. Atlantic 14870 1.8 18
Quercus myrtifolia Willd. Atlantic 2540 0.4 9
Quercus texana Buckl. Atlantic 829 1.1 9
```

```
Quercus coccinea Muenchh. Atlantic 8992 1.2 4
Quercus Douglasii Hook. & Arn California 559 4.1 18
Quercus dumosa Nutt. California 433 1.6 6
Quercus Engelmannii Greene California 259 2.0 17
Quercus Garryana Hook. California 1061 5.5 20
Quercus lobata Nee California 870 5.9 30
Quercus agrifolia Nee. California 803 2.6 23
Quercus Kelloggii Newb. California 826 6.0 26
Quercus Wislizenii A. DC. California 699 1.0 21
Quercus chrysolepis Liebm. California 690 17.1 15
Quercus vaccinifolia Engelm. California 223 0.4 1
Quercus tomentella Engelm California 13 7.1 18
```

One California species is an outlier in some analyses. This is the tree, Quercus tomentella Engelm, which is the only species of oak that grows on an island (Guadalupe), and not on the continent, and thus has its possible range restricted.

## 7.2 Efekty różnych czynników na pewną wielkość w jednej grupie

Czasami chcielibyśmy zbadać jakie korelacje istnieją pomiędzy różnymi czynnikami, a pewną wielkością obserwowaną i jakoś móc je pomiędzy sobą porównać (np. jak różne potencjalne czynniki karcerogenne wpływają na zachorowalność na raka, które z nich są bardziej, a które mniej istotne?). Do badania podobnych problemów służy wielowymiarowa regresja liniowa. Niech $y$ będzie wielkością obserwowaną a $X$ będzie zawierać wartości czynników, wtedy:

$$y = \beta X + \epsilon$$

### 7.2.1 Przykład: Smak cheddar'a

Reference: Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics.

Kiedy ser dojrzewa zachodzi wiele procesów chemicznych, które mają wpływ na jego ostateczny smak. Poniższe dane przedstawiają dla 30 próbek sera stężenia trzech substancji i subiektywną ocenę smaku próbki.

Zmienne:

1. Case: numer próbki 2. Taste: średnia ocena smaku 3. Acetic: logarytm stężenia kwasu octowego 4. H2S: Logarytm stężenia $H_2S$ 5. Lactic: stężenia kwasu mlekowego

Dane:

```
% Case taste Acetic H2S     Lactic
D=[ 1 12.3 4.543 3.135 0.86
    2 20.9 5.159 5.043 1.53
    3 39      5.366 5.438 1.57
    4 47.9 5.759 7.496 1.81
    5 5.6     4.663 3.807 0.99
    6 25.9 5.697 7.601 1.09
    7 37.3 5.892 8.726 1.29
    8 21.9 6.078 7.966 1.78
    9 18.1 4.898 3.85 1.29
    10 21      5.242 4.174 1.58
    11 34.9 5.74 6.142 1.68
    12 57.2 6.446 7.908 1.9
    13 0.7     4.477 2.996 1.06
    14 25.9 5.236 4.942 1.3
    15 54.9 6.151 6.752 1.52
    16 40.9 6.365 9.588 1.74
    17 15.9 4.787 3.912 1.16
    18 6.4     5.412 4.7     1.49
    19 18      5.247 6.174 1.63
    20 38.9 5.438 9.064 1.99
    21 14      4.564 4.949 1.15
    22 15.2 5.298 5.22 1.33
    23 32      5.455 9.242 1.44
    24 56.7 5.855 10.199 2.01
    25 16.8 5.366 3.664 1.31
    26 11.6 6.043 3.219 1.46
    27 26.5 6.458 6.962 1.72
    28 0.7     5.328 3.912 1.25
    29 13.4 5.802 6.685 1.08
    30 5.5     6.176 4.787 1.25];
% przeformatowujemy dane
przyp=D(:,1);
```

```
y=D(:,2); % Smak jest nasza zmienna zależna

X=[ones(size(y)) D(:,3:5)]; % zestaw naszych zmiennych niezależnych powiększamy
% o kolumnę jedynek jako miejsce na wyrazy stałe w modelu
[b,bint,r,rint,stats] = regress(y,X);
stats
%Jakość dopasowania modelu musimy też zweryfikować przez zbadanie residuów.
subplot(221)
errorbar(przyp,r,rint(:,1)-r,rint(:,2)-r,'o')
line([1 30],[0 0],'Color',[1 0 0])
subplot(222)
normplot(r)

% dofitowane b i ich przedziały ufności
disp({'b' 'przedzial b'})
disp([b bint])
```

W wyniku otrzymujemy

```
stats =

    0.6518   16.2214    0.0000  102.6312
```

Pierwszy element `stats` zawiera $R^2$ czyli jaką część wariancji nasz model tłumaczy, następnie statystykę $F$, dla hipotezy zerowej, że wszystkie współczynniki `b` są zerami, i odpowiadające jej prawdopodobieństwo. Ostatnia wielkość to estymator variancji błędu $\epsilon$.

Jakość dopasowania modelu musimy też zweryfikować przez zbadanie residuów. Residua powinny pochodzić z rozkładu normalnego o średniej 0. Na rysunku odkładamy residua i ich przedziały ufności i sprawdzamy czy, przedziały zawierają zero. Jeśli nie to dany przypadek jest 'outlierem' i trzeba mu się oddzielnie przyjrzeć.

Wartości dopasowanych `b` i ich przedziały ufności są w kolejności są w kolejności, w której umieściliśmy je w wywołaniu `regress`. Jeśli przedział ufności dla któregoś `b` zawiera zero to, że czynnik ten nie ma istosnego wpływu na zmienną zależną. U nas takim nieistotnym czynnikiem jest stężenie kwasu octowego.

```
    'b'    'przedzial b'

  -28.8768  -69.4435   11.6900
    0.3277   -8.8394    9.4949
    3.9118    1.3457    6.4780
   19.6705    1.9333   37.4078
% zobrazujemy to na rysunku
subplot(223)
b(1,:)=[];
bint(1,:)=[];
errorbar(1:3,b,bint(:,1)-b,bint(:,2)-b,'o')
line([1 3],[0 0],'Color',[1 0 0])
```

# Rozdział 8

# Analiza czynników głównych

### 8.0.1 Przykład

Oglądamy dane:

```
load cities
whos
boxplot(ratings,0,'+',0)
set(gca,'YTicklabel',categories)
```

Ponieważ wariancje w dwóch grupach są duże znormalizujemy dane:

```
stdr=std(ratings);
sr=ratings./repmat(stdr,329,1);
boxplot(sr,0,'+',0)
set(gca,'YTicklabel',categories)
```

Teraz poszczególne grupy są bardziej porównywalne. Robimy analizę składowych głównych:

```
[vec, newdata, variances, t2]=princomp(sr);
```

`vec` zawiera współczynniki kombinacji liniowej z jakimi do nowych danych wchodzą stare zmienne

```
vec(:,1:3)
```

`newdata` zawiera nowe dane — stare dane rzutowane na osie wyznaczone przez składowe główne

`variances` — wariancja wyjaśniona przez kolejny czynnik główny

```
pr_exp=100*variances/sum(variances)
pareto(pr_exp)
xlabel('Czynnik glowny')
ylabel('Wyjasniona warjancja %')
```

Widać, że już pierwsze trzy czynniki wyjaśniają 2/3 wariancji zbioru.

`t2` — miara wielowymiarowej odległości każdego punktu danych od centrum zbioru — przydaje się do badania danych ekstremalnych

```
[st2,index]=sort(t2);
index(end)
names(index(end),:)
```

# Rozdział 9

# Analiza czynnikowa — Factor Analysis

W tej analizie zakładamy, że nasze dane są kombinacją liniową pewnej ilości wspólnych czynników głównych i specyficznej składowej szumowej. Zarówno FA jak i PCA prowadzą do redukcji wymiarowości danych ale FA daje modele lepiej tłumaczące strukturę korelacji pomiędzy danymi.

Z tą analizą zapoznamy się na przykładzie danych z giełdy. Przez 100 tygodni zapisywano procentowe zmiany notowań 10 spółek. Z tej dziesiątki pierwsze cztery spółki można zaklasyfikować jako technologiczne, następne trzy jako finansowe, a ostanie trzy jako handlowe detaliczne. Sensowne wydaje się założenie, że notowania spółek należących do tego samego sektora powinny podlegać podobnym zmianom pod wpływem zmiana warunków ekonomicznych. Możemy to pokazać przy pomocy FA.

```
load stockreturns
[ Loadings, specificVar, T, stats]=factoran(stocks,3,'rotate','none');
```

Estymujemy model z trzema czynnikami wspólnymi, nie dokonujemy na razie rotacji w przestrzeni rozpiętej przez te czynniki.
`Loadings` — to współczynniki kombinacji liniowej z jaką czynniki wchodzą do tworzenia danej.
`specificVar` — to wariancja dla każdej z danych.

```
Loadings
Loadings =
```

```
    0.8885    0.2367   -0.2354
    0.7126    0.3862    0.0034
    0.3351    0.2784   -0.0211
    0.3088    0.1113   -0.1905
    0.6277   -0.6643    0.1478
    0.4726   -0.6383    0.0133
    0.1133   -0.5416    0.0322
    0.6403    0.1669    0.4960
    0.2363    0.5293    0.5770
    0.1105    0.1680    0.5524

specificVar
specificVar =

    0.0991
    0.3431
    0.8097
    0.8559
    0.1429
    0.3691
    0.6928
    0.3162
    0.3311
    0.6544
```

Specyficzna wariancja 1 wskazywałaby na to, że zmienna nie da się wyrazić przez czynniki wspólne, wartość 0 wskazywałaby na to, że zmienna wyraża się całkowicie przez owe czynniki.

W strukturze `stats` mamy pole `stats.p`. Jest to prawdopodobieństwo hipotezy zerowej, że nasze dane można wyrazić za pomocą zadanej ilości czynników 3 czynników.

```
stats.p

ans =

    0.8144
```

Czyli trzy czynniki są sensowną ilością. Możemy szybciutko podejrzeć co by było gdybyśmy zapostulowali dwa czynniki

```
[ Loadings2, specificVar2, T2, stats2]=factoran(stocks,2,'rotate','promax');
stats2.p
ans =

   3.5610e-06
```

Aby łatwiej było interpretować wyniki, trzeba poszukać takiego obrotu układu współrzędnych, w którym każda ze zmiennych będzie miała mało czynników wspólnych o dużej wartości. Metod obracanie jest wiele.

```
[ Loadings, specificVar, T, stats]=factoran(stocks,3,'rotate','none');
Loadings =

    0.9452    0.1214   -0.0617
    0.7064   -0.0178    0.2058
    0.3885   -0.0994    0.0975
    0.4162   -0.0148   -0.1298
    0.1021    0.9019    0.0768
    0.0873    0.7709   -0.0821
   -0.1616    0.5320   -0.0888
    0.2169    0.2844    0.6635
    0.0016   -0.1881    0.7849
   -0.2289    0.0636    0.6475
```

Narysujmy we współrzędnych wyznaczonych przez czynniki wspólne:

```
subplot(121)
plot(Loadings (:,1),Loadings (:,2),'b.' )
text( Loadings (:,1),Loadings (:,2),num2str((1:10)'))
xlabel('czynnik 1')
ylabel('czynnik 2')
grid
xlim([-1 1])
ylim([-1 1])
axis square
```

```
subplot(122)
plot(Loadings (:,2),Loadings (:,3),'b.' )
text( Loadings (:,2),Loadings (:,3),num2str((1:10)'))
xlabel('czynnik 2')
ylabel('czynnik 3')
grid
xlim([-1 1])
ylim([-1 1])
axis square
```

Widać, że dane układają się przy osiach i faktycznie tworzą grupy zgodnie z sektorami rynkowymi. Możemy więc otrzymane czynniki interpretować jako sektory rynkowe.

Jeśli zdecydujemy się na model trzy czynnikowy to możemy policzyć jak w kolejnych tygodniach wyglądały 'notowania sektorów'

```
clf
[ Loadings, specificVar, T, stats,F]=factoran(stocks,3,'rotate','promax');
t=1:length(F);
plot(t,F(:,1),t,F(:,2),t,F(:,3))
```

możemy też obejrzeć współzależność między sektorami:

```
plot3(F(:,1),F(:,2),F(:,3),'b.')
line([-4 4 NaN 0 0 NaN  0 0], [0 0 NaN -4 4 NaN  0 0] , [0 0 NaN  0 0 NaN -4 4]
grid on
view(-22.5,8)
axis square
xlabel('sektor finansowy')
ylabel(' sektor handlowy')
zlabel('sektor technologiczny')
```