AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Inżynierskie techniki obliczeniowe 2021/2022

Wykład nr 5

Dr inż. Przemysław Korohoda
E-mail: korohoda@agh.edu.pl
Tel.wewn.AGH: (012-617)-27-52
Pawilon C3 - p.506

Strona WWW:
home.agh.edu.pl/~korohoda/rok_2021_2022_lato/ITO_EL_1
UPeL: ITOEL2022

AGH

Plan wykładu

1. Generowanie pseudolosowe.
2. Modelowanie zależności - regresja liniowa.
3. Modelowanie zależności - regresja nieliniowa.

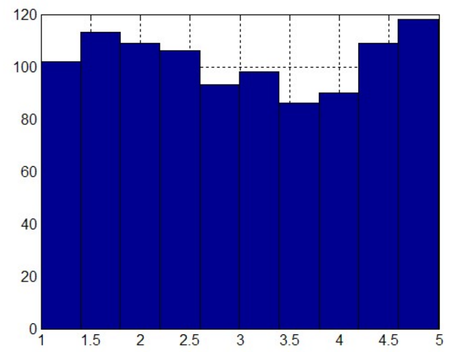
Rozkład (gęstości zmiennej losowej) równomierny o zadanych parametrach

Równomierny rozkład gęstości prawdopodobieństwa zmiennej losowej, określonej na odcinku $D=[x_0,x_1)$: $(x_1-x_0)=\Delta x$:

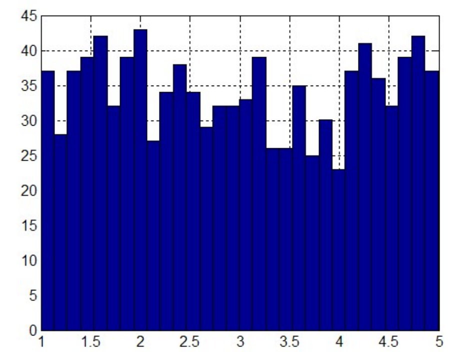
$$x_a, x_b \in D: \quad P\left(X \in [x_a, x_b)\right) = \int_{x_a}^{x_b} f(x)\,dx = \frac{(x_b - x_a)}{\Delta x}$$

```
…
  N=2^10;  u=3;  r=2;
  x=(rand(N,1)-0.5)*2*r+u;

  s  =  std(x),
  m  = mean(x),
  K=30;
  [hx,bx] = hist(x,K);

      figure(1); clf;
          bar(bx,hx,1);   grid on;
              xlim([u-r,u+r]);

…
```

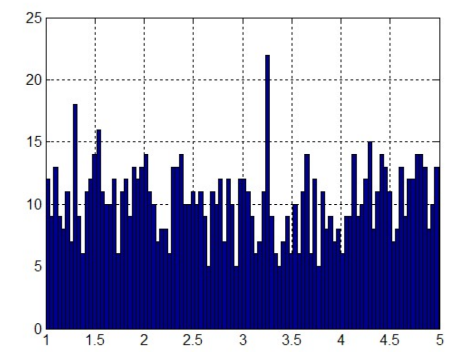
K=10:



K=30:



K=100:

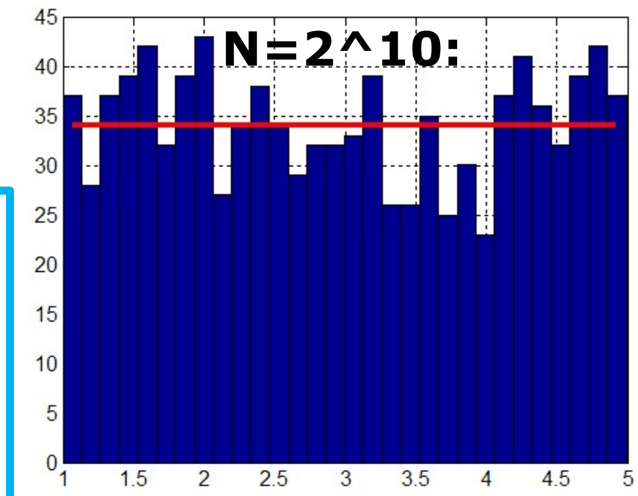


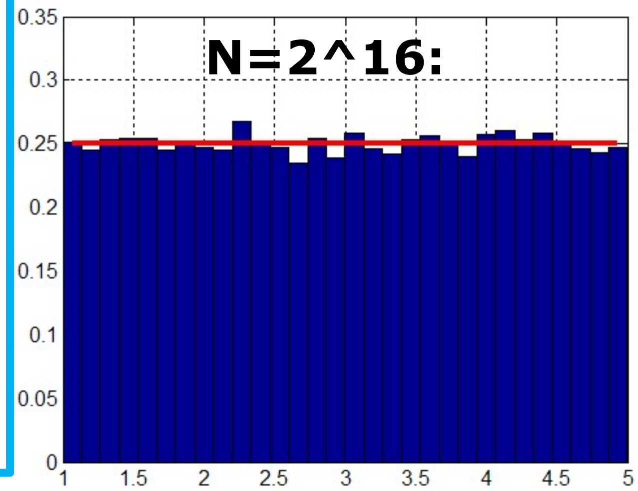
Inne pokrewne funkcje: histcounts, histogram.

Porównanie PDF z histogramem

Należy pamiętać o normalizacji.

```
...
  N=2^10;  u=3;  r=2;
  x = (rand(N,1)-0.5)*2*r+u;

  s   =  std(x),
  m  = mean(x),
  K=30;
  [hx,bx] = hist(x,K);     dx=bx(2)-bx(1);
  y1=1/(2*r);              q=sum(hx)*dx;

      figure(1); clf;
          bar(bx,hx,1);       grid on;    hold on;

          plot([bx(1),bx(end)],[y1,y1]*q,'r','linewidth',3);
                    xlim([u-r,u+r]);
...
```





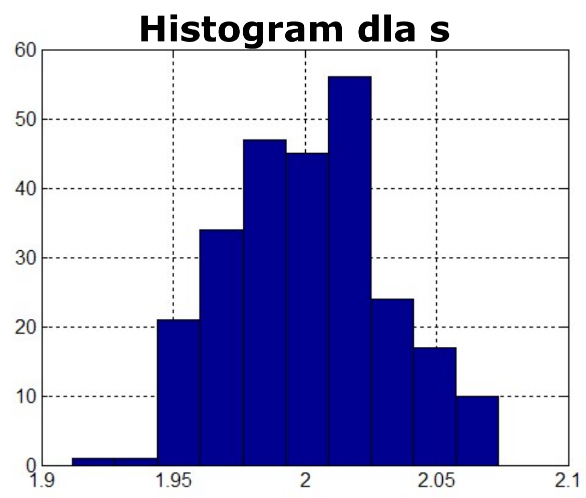
Uwaga – tu normalizacja do pdf'a

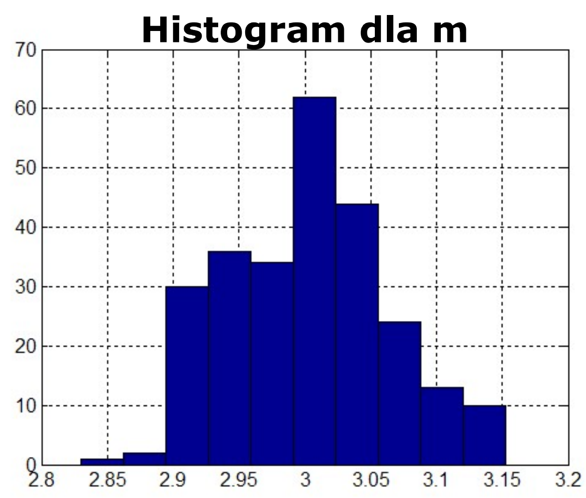
PDF – probability density function

Rozkłady z próby dla wartości średniej i odchylenia standardowego

Są to także zmienne losowe!
Zatem mają prawo mieć rozkłady (tzw. pdf'y)...

```
...
  M=2^8;  % liczba losowan;
  N=2^10;  u=3;  sigma=2;
  X=(rand(N,M)-0.5)*sqrt(12)*sigma+u;

  s   =  std(X),      % dziala na kolumnach macierzy;
  m  = mean(X),   % dziala na kolumnach macierzy;
    K=10;
    [hs,bs]=hist(s,K);
    [hm,bm]=hist(m,K);

...
```





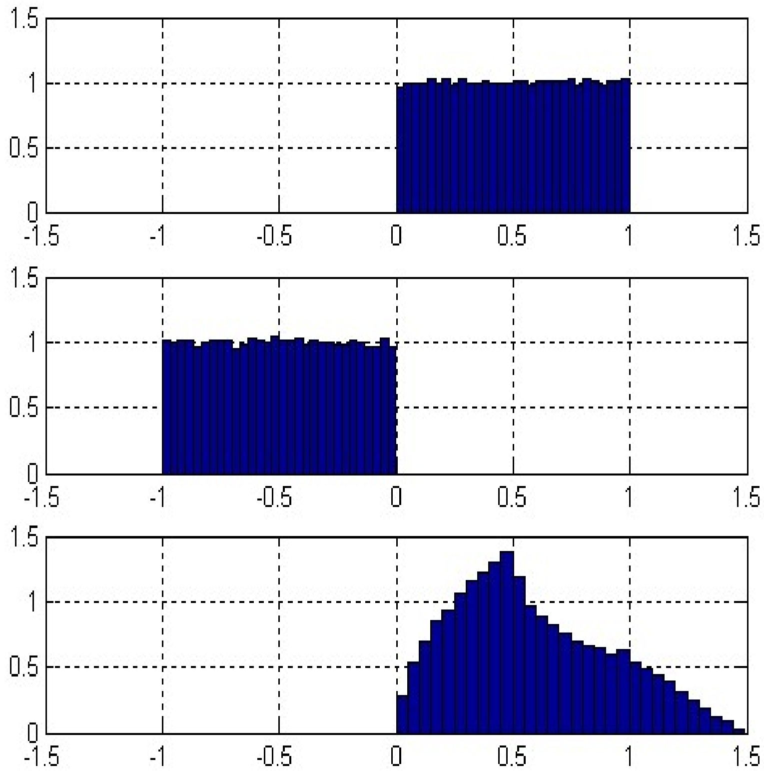
... i te rozkłady też mają swoje parametry!

Zmienna losowa otrzymana z innych zmiennych losowych (cd.)

$$Y(\mu_3, \sigma_3) = \frac{1}{2} \cdot X_1(\mu_1, \sigma_1) + [X_2(\mu_2, \sigma_2)]^2$$

```
...
  N=2^16;  % liczność próby losowej;
  x1=(rand(N,1)-0.5) + 0.5;
  x2=(rand(N,1)-0.5)  - 0.5;
  y=0.5*x1 + x2.^2;
  sx1 =  std(x1),   mx1  = mean(x1),
  sx2 =  std(x2),   mx2  = mean(x2),
  sy   =  std(y),    my    = mean(y),

    K=30;
    [hx1,bx1]  = hist(x1,K);
    [hx2,bx2]  = hist(x2,K);
    [hy,by]      = hist(y,K);
...
```

Rozkłady (pdf) otrzymane z próby:



Wtórne (pochodne) zmienne losowe

Wartość średnia z próby: $m = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \approx \mu$

Odchylenie standardowe z próby:

$$s_0 = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - m)^2} \approx \sigma$$

estymator obciążony

```
...
 s1  =  std(x),
 s2  =  std(x,1),
...
```

$$s_1 = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - m)^2} \cong \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu)^2}$$

estymator nieobciążony

<u>Transformacje rozkładów (to są tylko przykłady!):</u>

Logarytm ze zmiennej losowej (dodatniej!): $y = \log_e(x)$

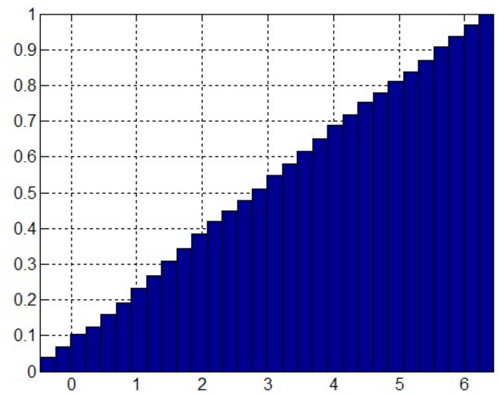
Funkcja wykładnicza (np. eksponenta) ze zmiennej losowej: $y = e^x$

Projektowanie generatorów o specjalnych rozkładach

Dystrybuanta (CDF – cumulative density function):

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \qquad \textbf{gdzie} \qquad f(x) = \mathrm{PDF}(x)$$

```
...
  N=2^10;  u=3;  sigma=2;
  x=(rand(N,1)-0.5)*sqrt(12)*sigma+u;

    K=10;
    [hx,bx]=hist(x,K);
    hcx=cumsum(hx);  % mozna zrobic samemu;

          figure(1); clf;
              bar(bx,hcx/hcx(end),1);   grid on
                            axis tight;
...
```



AGH

Projektowanie generatorów o specjalnych rozkładach (cd.)

Metoda ta może być stosowana dla rozkładów mających dystrybuantę, dla której można wyznaczyć funkcję odwrotną na przedziale, na którym rozkład jest niezerowy.

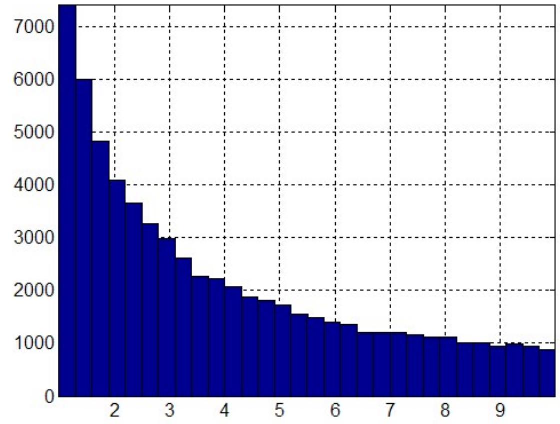
Algorytm:

1. **Za pomocą generatora *rand* losujemy liczby z przedziału [0,1], które traktujemy jak wartości z pionowej osi dystrybuanty.**
2. **Wyznaczamy (funkcją odwrotną do dystrybuanty) wartości na osi poziomej odpowiadające wylosowanym wartościom z osi pionowej.**
3. **Te wartości spełniają założony rozkład.**

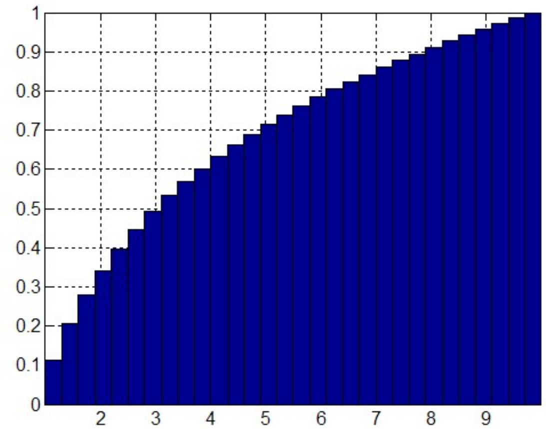
Projektowanie generatorów o specjalnych rozkładach (cd.)

```
...
  N=2^16;
  x=rand(N,1);
  y=10.^x;

    K=30;
    [hy,by]=hist(y,K);
    hcy=cumsum(hy);

        fs=12;
      figure(1); clf;
          bar(by,hy,1);                grid on;
             set(gca,'fontsize',fs);    axis tight;
      figure(2); clf;
          bar(by,hcy/hcy(end),1);  grid on;
             set(gca,'fontsize',fs);    axis tight;
...
```

Histogram z próby



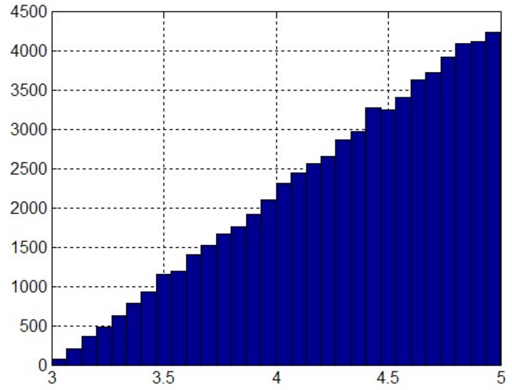
Dystrybuanta z próby



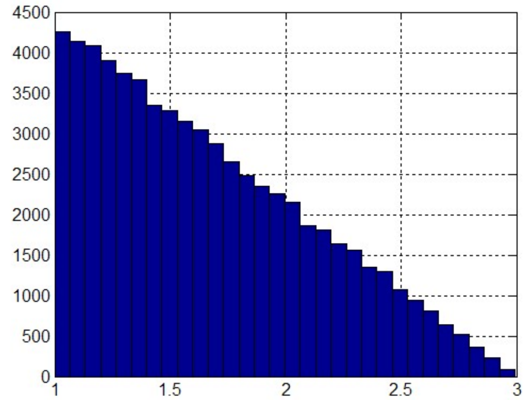
Projektowanie generatorów o specjalnych rozkładach (cd.)

Zadanko do przemyślenia: jak tą metodą można otrzymać rozkłady np. trójkątne o zadanych parametrach?

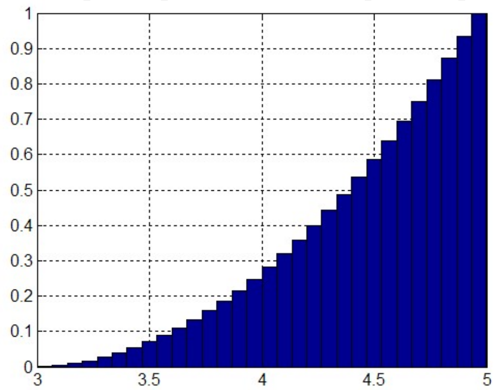
Histogram z próby



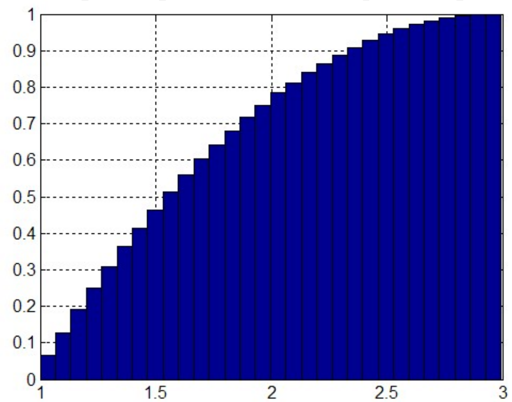
Histogram z próby



Dystrybuanta z próby



Dystrybuanta z próby

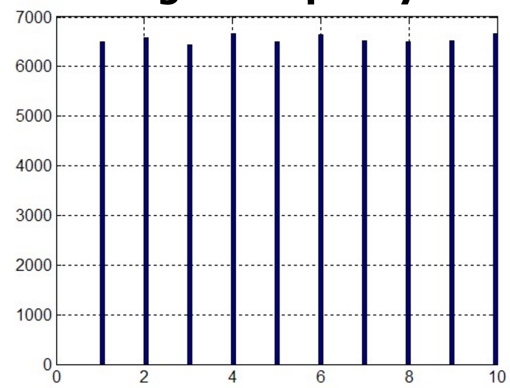


Rozkłady prawdopodobieństwa - dyskretne zmienne losowe

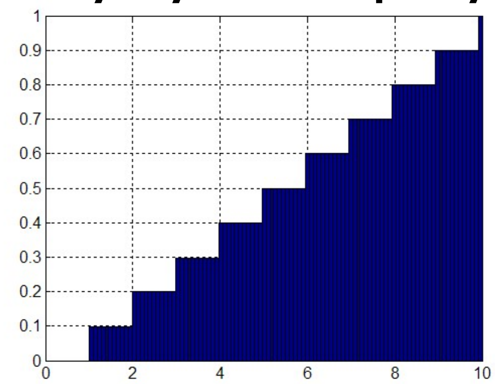
Np. równomierny rozkład prawdopodobieństwa dyskretnej zmiennej losowej: $k = 1,2,\ldots,n : P(X = x_k) = \frac{1}{n}$

```
...
  N=2^16;
  M=10;
  x=floor(rand(N,1)*M)+1;
  x(x>M)=M; % nadmiar ostroznosci?

    K=100;
    [hx,bx]=hist(x,K);
    hcx=cumsum(hx)/sum(hx);
...
```

Histogram z próby



Dystrybuanta z próby



Inne funkcje zaokrąglające: ceil, round.

Można też użyć funkcji: randi (byle poprawnie!).
Niekiedy przydatna może być też funkcja: randperm.

Klasyczne całkowitoliczbowe generatory pseudolosowe

Do generowania całkowitych, nieujemnych liczb losowych z przedziału [0, *M*) najczęściej stosowane są generatory kongruentne, liniowe działające według przepisu

$$x_{n+1} = \left(\sum_{m=0}^{k} a_m x_{n-m} \right) \bmod M$$

gdzie współczynniki kombinacji liniowej są liczbami całkowitymi wybranymi arbitralnie z zadanego przedziału [0, *M*)

Np.: generatory dwumianowe:

$$x_{n+1} = \left(a x_{n-1} + b \right) \bmod M$$

... i dalej np.:

a) generatory Fibonacciego:

$$x_{n+1} = \left(x_n + x_{n-1} \right) \bmod M$$

...lub:

b) generatory multiplikatywne :
(dla *b* różnego od zera!)

$$x_{n+1} = \left(b x_n \right) \bmod M$$

Matlab: mod, rem (uwaga – dla ujemnych argumentów wyniki mogą być różne)

Dlaczego jest to generowanie pseudolosowe?

Można sprawdzić:

```
...
N=4;
for n=1:3
    rng(2000); % random number generator: seed=2000;
    X1(n,1:N)=rand(1,N);
End

    rng('shuffle');  % ustawia poczatek sekwencji na podst. zegara;
for n=1:3
    X2(n,1:N)=rand(1,N);
end

X1
X2
...
```

AGH

Dlaczego rozkład normalny jest taki normalny?

Czyli: Centralne Twierdzenie Graniczne.
Jest kilka wersji tego twierdzenia, przy różnie sformułowanych założeniach.
Generalnie istotne jest, by rozkłady składowe były „skupione”.

$$\lim_{N\to\infty}\left(\sum_{n=1}^{N} X_n(\mu_n,\sigma_n)\right) = X_{norm}(\mu,\sigma)$$

W praktyce liczba zmiennych losowych jest zwykle skończona, co daje przybliżony – np. obustronnie obcięty - rozkład Gaussa.

np.: $$Y = \sum_{n=1}^{N} X_n(\mu_n,\sigma_n) \;\Rightarrow\; \mu_Y = \sum_{n=1}^{N}\mu_{X_n} \;\wedge\; \sigma_Y^2 = \sum_{n=1}^{N}\sigma_{X_n}^2$$

lub: $$Y = \frac{1}{N}\cdot\sum_{n=1}^{N} X_n(\mu_n,\sigma_n) \;\Rightarrow\; \mu_Y = \frac{1}{N}\cdot\sum_{n=1}^{N}\mu_{X_n} \;\wedge\; \sigma_Y^2 = \frac{1}{N^2}\cdot\sum_{n=1}^{N}\sigma_{X_n}^2$$

Efekt dla wartości średniej

Dla takich samych rozkładów, czyli przy powtarzanym *N* razy niezależnym losowaniu z tej samej populacji:

$$Y = \frac{1}{N} \cdot \sum_{n=1}^{N} X(\mu_X, \sigma_X) \quad \Rightarrow \quad \mu_Y = \mu_X \quad \wedge \quad \sigma_Y = \frac{\sigma_X}{\sqrt{N}}$$

zmienna losowa = średnia z próby

Rozkład Gaussa

Rozkład Gaussa zmiennej losowej *x*:

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

Standaryzacja zmiennej losowej:

$$X_0(0,1)=\frac{X(\mu,\sigma)-\mu}{\sigma}$$

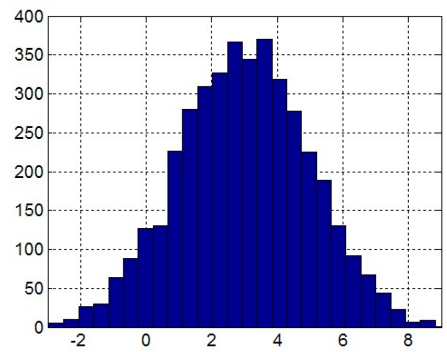
lub odwrotnie:

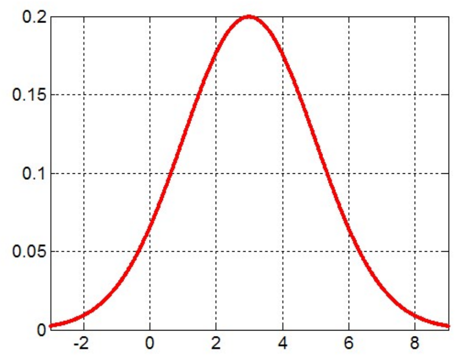
$$X(\mu,\sigma)=X_0(0,1)\cdot\sigma+\mu$$

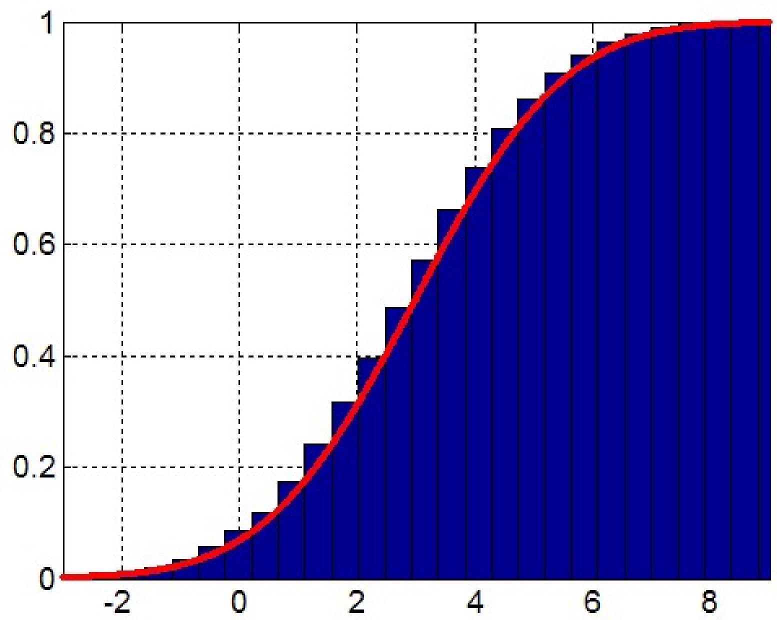
Matlab: randn, normpdf, normcdf, norminv.

Rozkład Gaussa (cd.)

```
...
N=2^12;
   u=3; sigma=2;
x=randn(N,1)*sigma+u;

[hx,bx]=hist(x,30);
hcx=cumsum(hx)/sum(hx);

   xx=u-3*sigma:6*sigma/1000:u+3*sigma;

xpdf  =  normpdf(xx,u,sigma);
xcdf  =  normcdf(xx,u,sigma);
xx2   =  norminv(xcdf,u,sigma);

   test_xxx=max(abs(xx-xx2)),
...
```







AGH

Regresja liniowa

Próba losowa:

$$(x_n, y_n) \in \mathbf{R}^2 \quad : \quad n = 1,2,\dots,N$$

$$D = \{(x_1, y_1),(x_2, y_2),(x_3, y_3),\dots,(x_N, y_N),\}$$

$$\mathbf{D} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ . & . \\ x_N & y_N \end{bmatrix} = [\mathbf{x}, \mathbf{y}]$$

$$\mathbf{D}^T = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_N \\ y_1 & y_2 & y_3 & \dots & y_N \end{bmatrix} = \begin{bmatrix} \mathbf{x}^T \\ \mathbf{y}^T \end{bmatrix}$$

Regresja liniowa (cd.)

Zakładamy, że między wartościami *x* i *y* istnieje zależność matematyczna - model matematyczny tej zależności.
Zakładamy, że ma on postać funkcji liniowej:

$$l(a_1, a_2): \quad y = a_1 \cdot x + a_2$$

Jeżeli wartości *x* możemy uznać za dokładne, to prostą regresji wyznaczamy minimalizując sumę kwadratów błędów liczonych jedynie w kierunku *y*:

$$\min_{a_1, a_2}\left(\sum_{n=1}^{N}\left[y_n - (a_1 \cdot x_n + a_2)\right]^2\right)$$

Regresja liniowa (cd.)

$$a_1 \cdot x_n + a_2 = y_n - e_n \quad : \quad n = 1,2,\ldots,N$$

$$\mathbf{X} \cdot \mathbf{a} = \mathbf{y} - \mathbf{e}$$

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ . & 1 \\ x_{N-2} & 1 \\ x_{N-1} & 1 \\ x_N & 1 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_{N-2} \\ y_{N-1} \\ y_N \end{bmatrix} - \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ e_{N-2} \\ e_{N-1} \\ e_N \end{bmatrix}$$

Stosujemy macierz pseudoodwrotną:

$$\mathbf{a} = \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$$

Regresja liniowa, wielowymiarowa

$$a_0 + a_1 \cdot x_{1,n} + a_2 \cdot x_{2,n} + a_3 \cdot x_{3,n} + \ldots + a_K \cdot x_{K,n} = y_n - e_n$$

$$: \quad n = 1,2,\ldots,N$$

$$\mathbf{X} \cdot \mathbf{a} = \mathbf{y} - \mathbf{e}$$

N **wektorowych wartości x;** ***K*** **elementów każdego wektora x:**

$$\mathbf{x}_n = \begin{bmatrix} x_{1,n} \\ x_{2,n} \\ \cdot \\ x_{K,n} \end{bmatrix} \qquad : \quad n = 1,2,\ldots,N$$

Regresja liniowa, wielowymiarowa (cd.)

$$a_0 + a_1 \cdot x_{1,n} + a_2 \cdot x_{2,n} + a_3 \cdot x_{3,n} + \ldots + a_K \cdot x_{K,n} = y_n - e_n$$

$$: \quad n = 1,2,\ldots,N$$

$$\mathbf{X} \cdot \mathbf{a} = \mathbf{y} - \mathbf{e}$$

$$\begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \ldots & x_{K,1} \\ 1 & x_{1,2} & x_{2,2} & \ldots & x_{K,2} \\ 1 & x_{1,3} & x_{2,3} & \ldots & x_{K,3} \\ . & . & . & . & . \\ 1 & x_{1,N} & x_{2,N} & \ldots & x_{K,N} \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ . \\ a_K \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_{N-2} \\ y_{N-1} \\ y_N \end{bmatrix} - \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ e_{N-2} \\ e_{N-1} \\ e_N \end{bmatrix}$$

Stosujemy macierz pseudoodwrotną:

$$\mathbf{a} = \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$$

Regresja nieliniowa

$$a_0 + a_1 \cdot g(x_n) = y_n - e_n \quad : \quad n = 1,2,3,\ldots,N$$

$$\mathbf{X} \cdot \mathbf{a} = \mathbf{y} - \mathbf{e}$$

$$\begin{bmatrix} 1 & g(x_1) \\ 1 & g(x_2) \\ 1 & g(x_3) \\ . & . \\ 1 & g(x_{N-2}) \\ 1 & g(x_{N-1}) \\ 1 & g(x_N) \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_{N-2} \\ y_{N-1} \\ y_N \end{bmatrix} - \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ e_{N-2} \\ e_{N-1} \\ e_N \end{bmatrix}$$

Stosujemy macierz pseudoodwrotną:

$$\mathbf{a} = \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$$

AGH

Regresja nieliniowa, wielowymiarowa

$$a_0 \cdot g_0(x_n) + a_1 \cdot g_1(x_n) + a_2 \cdot g_2(x_n) + \ldots + a_K \cdot g_K(x_n) = y_n - e_n$$

$$: \quad n = 1,2,3,\ldots,N$$

$$\text{np.:} \quad g_0(x) = 1; \quad g_k(x) = \cos(2 \cdot \pi \cdot k \cdot f_0 \cdot x - \varphi_k) \quad , k = 1,2,\ldots,K$$

$$\text{lub:} \quad g_k(x) = x^k \quad , k = 0,1,2,\ldots,K$$

Regresja nieliniowa, wielowymiarowa (cd.)

$$a_0 \cdot g_0(x_n) + a_1 \cdot g_1(x_n) + a_2 \cdot g_2(x_n) + \ldots + a_K \cdot g_K(x_n) = y_n - e_n$$

$$: \quad n = 1,2,3,\ldots,N$$

$$\mathbf{X} \cdot \mathbf{a} = \mathbf{y} - \mathbf{e}$$

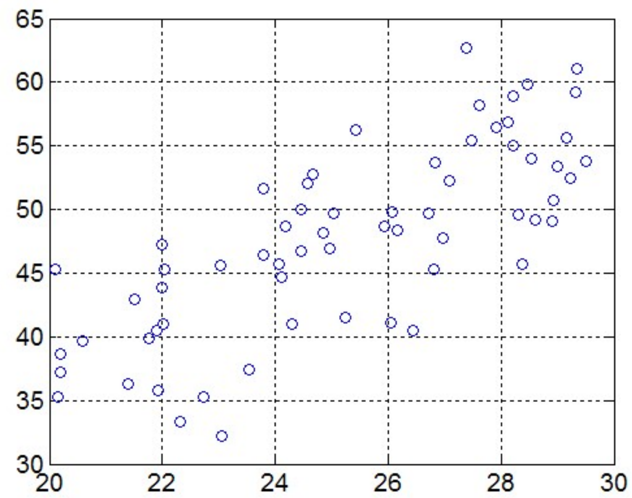
$$\begin{bmatrix} g_0(x_1) & g_1(x_1) & g_2(x_1) & \ldots & g_K(x_1) \\ g_0(x_2) & g_1(x_2) & g_2(x_2) & \ldots & g_K(x_2) \\ g_0(x_3) & g_1(x_3) & g_2(x_3) & \ldots & g_K(x_3) \\ . & . & . & . & . \\ g_0(x_N) & g_1(x_N) & g_2(x_N) & \ldots & g_K(x_N) \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ . \\ a_K \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_{N-2} \\ y_{N-1} \\ y_N \end{bmatrix} - \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ e_{N-2} \\ e_{N-1} \\ e_N \end{bmatrix}$$

Stosujemy macierz pseudoodwrotną:

$$\mathbf{a} = \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$$

Przykład generowania danych testowych

```
...
N=2^6;
x=rand(N,1)*10+20;

% wartości idealne:
a0=[-3, 2];
y0=a0(1)+a0(2)*x;          % wynik bez szumu
                            % “pomiarowego” ;

% zmierzone wartości y (zaszumione):
y=y0+randn(N,1)*5;         % dokladamy
                            % szum addytywny
                           % gaussowski
                            % o zerowej sredniej ;
...
```



AGH

Zapraszam do laboratorium ...