Wizualizacja rozkładu zmiennej z wykorzystaniem pakietu Matlab 6.5

Praca została oparta na podręczniku autorstwa pani prof. Grażyny Wieczorkowskiej pt. *„Statystyka. Wprowadzenie do analizy danych sondażowych i eksperymentalnych”*. Przykłady napisane zostały w środowisku Matlab w wersji 6.5.

Zawarte w opracowaniu kody wystarczy skopiować i wkleić do pliku M-file w Matlab’ie. Wszystkie powinny się bez problemu skompilować i dać rezultaty identyczne (ewentualnie podobne) do przykładów z pracy.

mgr. inż. Anna Jurczyk

KOSZALIN 2005

I. Wizualizacja rozkładu zmiennej

1. Wykres słupkowy i kołowy (tortowy)

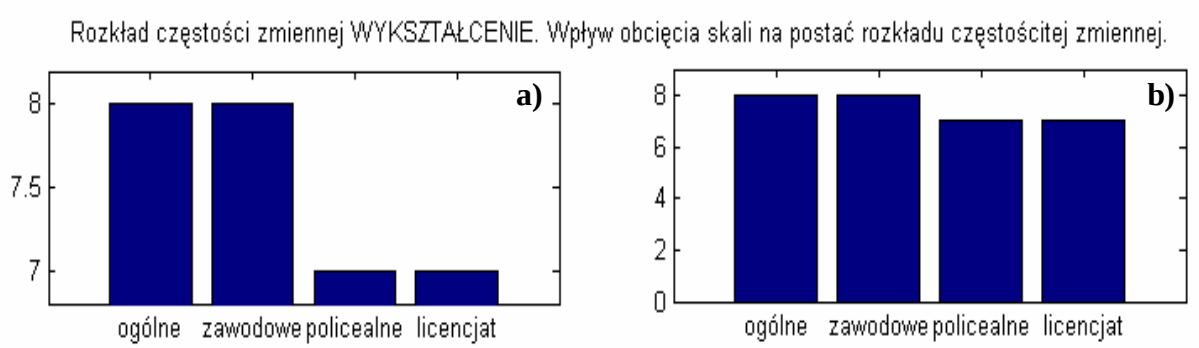
Wykresy te przekazują w formie graficznej informację, jaka daje nam zwykła tabela częstości. Wysokość słupków wykresu słupkowego zależy od liczebności danej kategorii zmiennej lub od jej procentowego udziału. Wykresy te są bardzo proste w interpretacji.

wykształcenie	liczebność	częstość	procent %
średnie ogólne	8	0.27	27
średnie zawodowe	8	0.27	27
policealne	7	0.23	23
licencjat	7	0.23	23
ogółem:	**30**	**1**	**100**

Tabela 1. Rozkład częstości zmiennej WYKSZTAŁCENIE

Wykres słupkowy

Trzeba uważać, aby siebie samego i innych nie zmylić wadliwie zrobionym wykresem słupkowym, ze źle dobraną skalą. Patrząc na rysunek 1a wydaje się, że różnica liczebności osób z wykształceniem policealnym i średnim zawodowym jest dość znaczna. Wystarczy jednak rzut oka na skalę, aby zrozumieć, że tak nie jest. Widać to na poprawnie zrobionym rysunku 2b, gdzie skala nie jest sztucznie obcięta.



Rysunek 1. Rozkład częstości zmiennej WYKSZTAŁCENIE. Wykres słupkowy

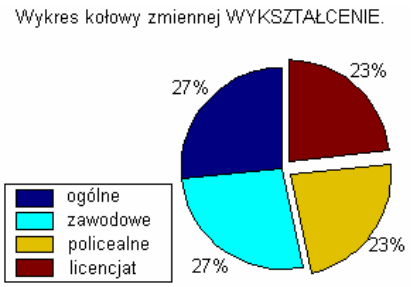
Implementacja w środowisku Matlab

```
function wyk_Slup()
%zamykanie wszystkich otwartych figur
close all
kategorie={'ogólne','zawodowe','policealne','licencjat'};
liczebnosci=[8 8 7 7];
%podział figury (okienka) na części (1 wiersz i 2 kolumny)
subplot(1,2,1)
%funkcja rysująca wykres słupkowy
bar(liczebnosci);
%ustawienie skali(ylim) oraz podpisanie kategorii(xticklabel)
```

```
set(gca,'ylim',[6.8 8.2],'xticklabel', kategorie);
subplot(1,2,2)
bar(liczebnosci);
set(gca,'ylim',[0 9],'xticklabel', kategorie);
%*************************************************
```

Wykres kołowy (tortowy)

Wykres ten ma dwa poważne ograniczenia. Gdy kategorii zmiennej jest wiele wówczas wykres staje się zupełnie nieczytelny. Drugie ograniczenie jest związane z percepcją i możliwościami ludzkiego oka. Ludzkie oko niezbyt dobrze porównuje powierzchnie. W rezultacie, gdy mamy więcej kategorii o podobnych liczebnościach to często trudno jest powiedzieć, która kategoria jest liczniejsza.



Rysunek 2. Wykres kołowy rozkładu zmiennej WYKSZTAŁCENIE

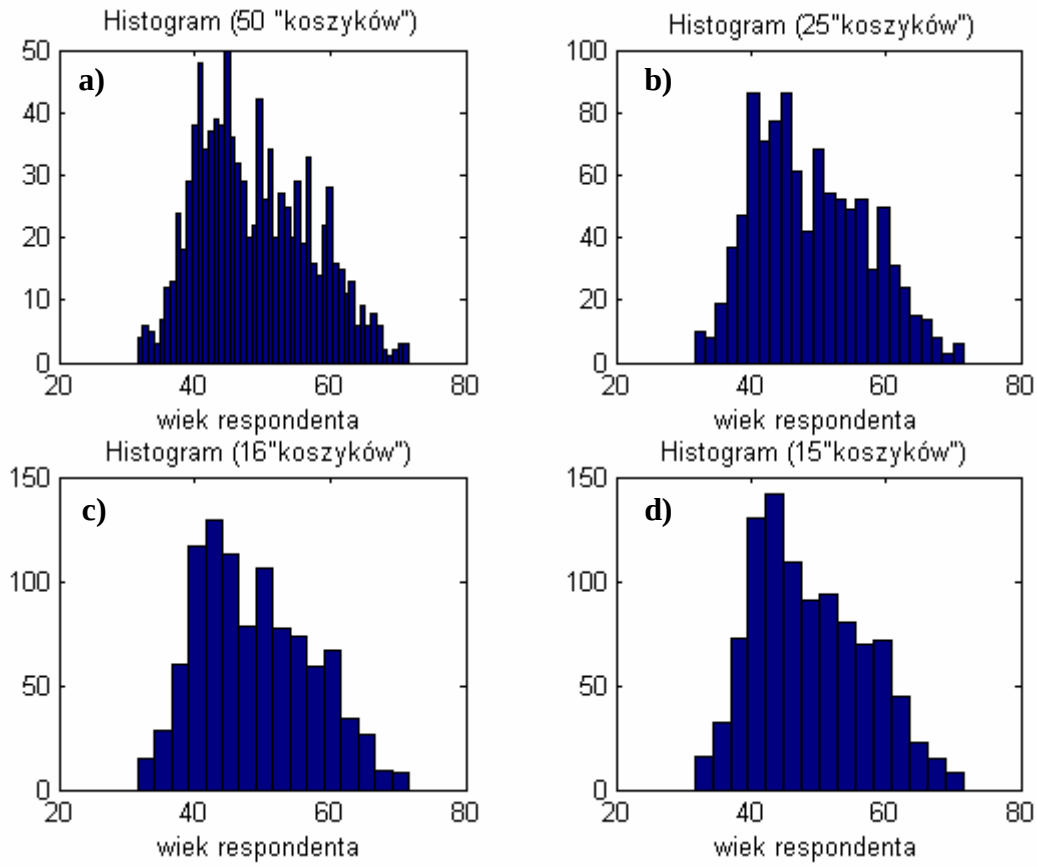
Implementacja w środowisku Matlab

```
function wyk_Kol()
%zamykanie wszystkich otwartych figur
close all
kategorie={'ogólne','zawodowe','policealne','licencjat'};
liczebnosci=[8 8 7 7];
%wektor, w którym należy wstawić 1 na pozycji liczebności, dla %której
wycinek koła ma zostać wysunięty
wycinek = [0 0 1 1 ];
%wyświetlenie wykresu
pie(liczebnosci,wycinek)
%kolory wycinków (np.: JET, HSV, COOL, HOT itp.)
colormap JET
%ustawianie czcionki na Arial Ce (ma polskie litery)
set(gca,'fontName','Arial Ce');
%wyświetlanie legendy przy czym (1 – prawy, górny róg; 2 – %lewy, górny
róg; 3 – lewy, dolny róg; 4 – prawy, dolny róg; %-1 – na zewnątrz)
legend(kategorie,3);
title('Wykres kołowy zmiennej WYKSZTAŁCENIE');
%*************************************************
```

2. Histogram

Gdy zmienna porządkowa lub przedziałowa ma dużo kategorii, wówczas wykres słupkowy staje się nieczytelny, możemy wtedy posłużyć się histogramem. Jak wiemy, wykres słupkowy pokazuje liczebność kategorii zmiennych. Gdy kategorii jest dużo, to zamiast próbować rysować wszystkie, można połączyć te sąsiadujące ze sobą. W efekcie otrzymujemy mniejszą liczbę kategorii do narysowania. Tworzenie histogramu zaczyna się więc od przyporządkowania kategoriom zmiennych przedziału („koszyka"), do którego wpadną, następnie zliczamy liczbę przypadków w każdym koszyku i na tej podstawie tworzymy wykres.

W tym miejscu trzeba wskazać na poważną wadę histogramu. Otóż, o cym się zaraz przekonamy, jego wygląd zależy w znacznym stopniu od wyboru liczby przedziałów (koszyków), na które dzielimy naszą zmienną .



Rysunek 3. Histogramy tej samej zmiennej WIEK - dla różnej liczby „koszyków"

Popatrzmy na rysunek 3d, na którym mamy dane pogrupowane w 15 koszyków. Można z niego wywnioskować, że rozkład zmiennej jest jednomodalny i prawoskośny . Gdy zwiększymy liczbę koszyków do 25 (rys 3b), pojawia się druga modalna. Dalej, gdy koszyków jest jeszcze więcej (rys 3a), widzimy że są w zasadzie 3 mody. Trudno dać jednoznaczną odpowiedz na to pytanie *„Jaki wybór liczby koszyków jest prawidłowy?"*. Nie należy przy tym ulegać złudzeniu, że im więcej koszyków tym lepiej. Dodanie tylko jednego koszyka może spowodować dość znaczne różnice między histogramami.

Histogram jest przydatny, gdy możemy sprawdzić, jak będzie wyglądał w przypadku zmiany liczby koszyków. Wtedy, eksperymentując, możemy porównać wykresy otrzymane w różnych przypadkach i wyrobić sobie zdanie na temat rozkładu zmiennej.

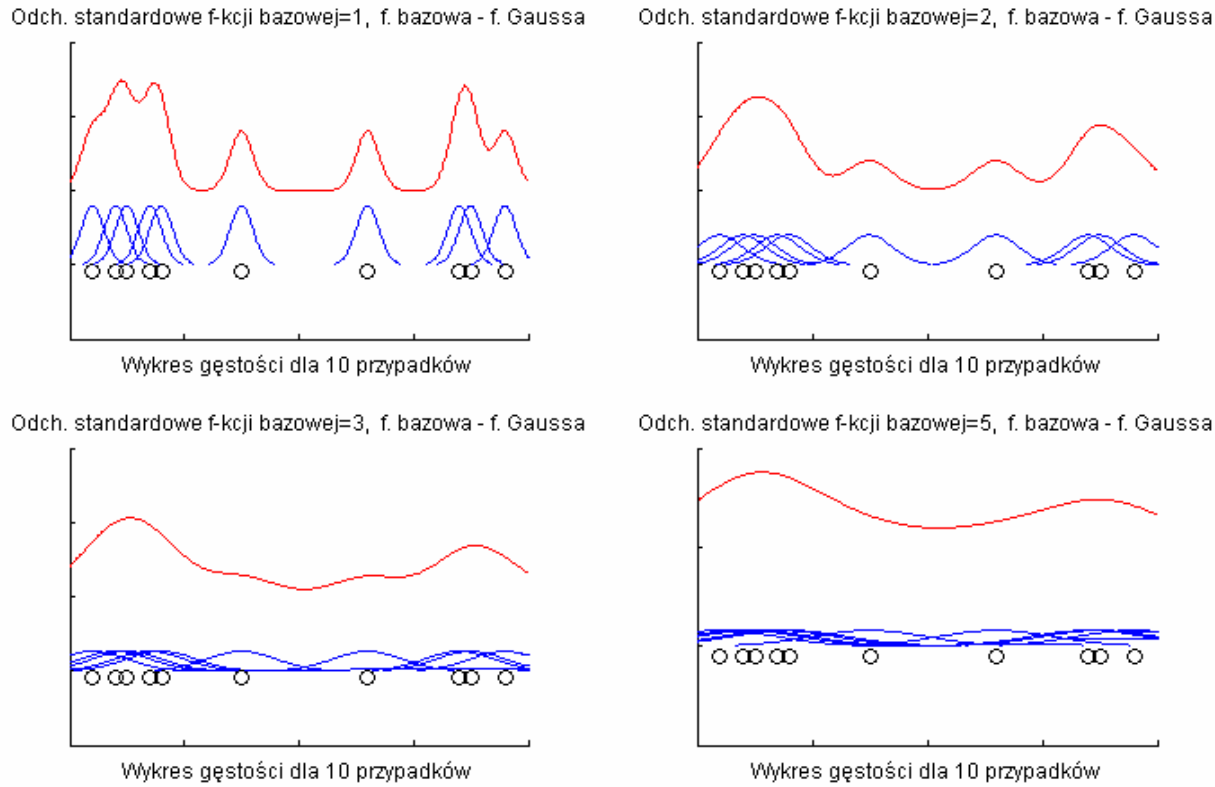
Implementacja w środowisku Matlab

```
function histogramy()
close all
clc
[rozklad,N] = generuj_rozklad; %funkcja opisana poniżej
odchylenie=std(rozklad);
srednia=mean(rozklad);
n_kosz=[50,25,16,15]; %wektor liczby „koszyków
for i=1:4
subplot(2,2,i); %podział figury na 2 wiersze i 2 kolumny
hist(rozklad,n_kosz(i)); %funkcja wyświetlająca histogram
%gtext - wypisanie tekstu na figurze w miejscu wskazanym
%myszką, strat – połączenie dwóch (i więcej) tekstów.
gtext(strcat('Odch. Std=',num2str(odchylenie,2)));
gtext(strcat('Średnia =',num2str(srednia,2)),...
'FontName','Arial Ce','FontSize',9);
gtext(strcat('N =',num2str(N,2)));
title(strcat('Histogram (',num2str(n_kosz(i)),' "koszyków")'));
xlabel('wiek respondenta');
end
%*************************************************

function [rozklad,N] = generuj_rozklad()
%generowanie rozkładu kilkumodalnego poprzez zsumowanie dwóch
%rozkładów normalnych o różnych wartościach średnich
%i odchyleniach
X=18:.1:100;
N=5000;
%normpdf(X,mu,s) zwraca wartości Y wykresu gęstości rozkładu
%normalnego o wartości średniej mu i odchyleniu standardowym s
Y_norm1=normpdf(X,mean(42),4);
Y_norm2=normpdf(X,mean(55),6);
p_r=Y_norm1+Y_norm2;
%zeskalowanie wektora p_r tak, aby suma jego wartości była %równa 1 (gdyż
ma to być wektor prawdopodobieństw)
r=1/sum(p_r);
p_r=p_r*r;
```

```
%randsrc(N,N,[wektorX; wektorP]) - generuje macierz N na N %wartości z
przedziału wektorX.Powtórzenia danej wartości z %wektorX zależne są od
wartości odpowiadającego jej %prawdopodobieństwa w wektorP.
%Np.: a = randsrc(1,10,[4 9;0.3 0.7]) - oznacza, że zmienna a
%jest wektorem 10 elementów z czego 3% to 4, a 7% to 9.
rozklad = randsrc(1,N,[X;p_r]);
%czyli zmienna rozkład jest wektorem 5000 elementów
%o wartościach z przedziału od 18 do 100 co 0.1 (wektor X), %które
powtarzają się według prawdopodobieństwa zapisanego
%w wektorze p_r
%*************************************************
```

3. Wykres gęstości

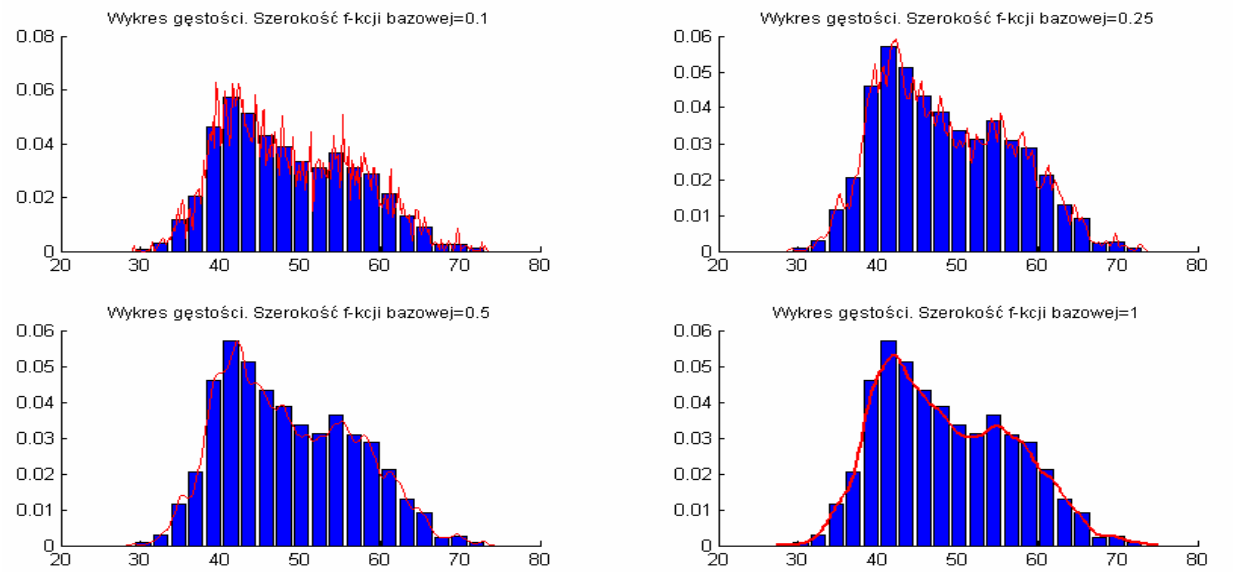
Innym wykresem, który daje nam informację analogiczną do histogramu jest wykres gęstości. Wykres gęstości powstaje w następujący sposób:
Wybieramy najpierw pewną symetryczną funkcję, którą będziemy nazywać funkcją bazową (nazywaną jądrem od ang. kernel). Następnie każdej obserwacji przyporządkowujemy tę funkcję tak, żeby oś jej symetrii pokrywała się z daną obserwacją. Po czym dodajemy do siebie wszystkie funkcje bazowe, otrzymując w ten sposób krzywą, która jest właśnie wykresem gęstości.



Rysunek 4. Wykres gęstości dla różnych szerokości h funkcji bazowej

Wykres gęstości podobnie jak histogram, wymaga ustalenia pewnych parametrów. Przede wszystkim musimy wybrać funkcję bazową. Tutaj była to funkcja Gaussa.

Drugi parametr, który musimy ustalić, stanowi szerokość funkcji bazowej h – w przypadku f. Gaussa jest to oczywiście odchylenie standardowe. Im szerokość ta jest większa tym gładszy wykres otrzymujemy, im mniejsza, tym bardziej wykres jest poszarpany (rysunek 4).



Rysunek 5. Histogram wraz z wykresem gęstości dla różnych szerokości h funkcji bazowej

Na rysunku 5 wykreślone są cztery histogramy zmiennej wiek wraz z wykresami gęstości o różnej szerokości funkcji bazowej h. Patrząc na serię rysunków widzimy, że dla szerokości 0.1 wykres zawiera dużo nieistotnych informacji, wykres otrzymany dla h=0.25 i h=0.5 wydaje się wiernie pokazywać podstawowe własności rozkładu zmiennej. Ostatni rozkład, dla h=1, jest zbyt wygładzony.

Warto zaznaczyć, że zarówno histogram, jak i wykres gęstości niezbyt dobrze radzą sobie z wartościami odstającymi – mogą one łatwo zostać niezauważone, a dla niektórych wartości parametrów (szerokość funkcji bazowej lub liczby koszyków) – w ogóle nie zauważone. Z tego względu zajmijmy się teraz wykresem skrzynkowym.

Implementacja w środowisku Matlab

```
function Wyk_Gestosci()
close all
clc
[rozklad,N] = generuj_rozklad; %funkcja opisana poniżej
[liczeb,liczebX]= hist(rozklad,20)
szer_fB=[.1 .25 .5 1]; %wektor szerokośći funkcji bazowej

for i=1:4
subplot(2,2,i)
hold on %nałożenie kilku wykresów na 1
%ksdensity - zwraca wsp. punktów wykresu gęstości dla zmiennej rozklad
%gdzie parametr kernel to rodzaj funkcji bazowej a width to jej szerokość
[Yi,Xi] = ksdensity(rozklad,'kernel','normal','width',szer_fB(i));
```

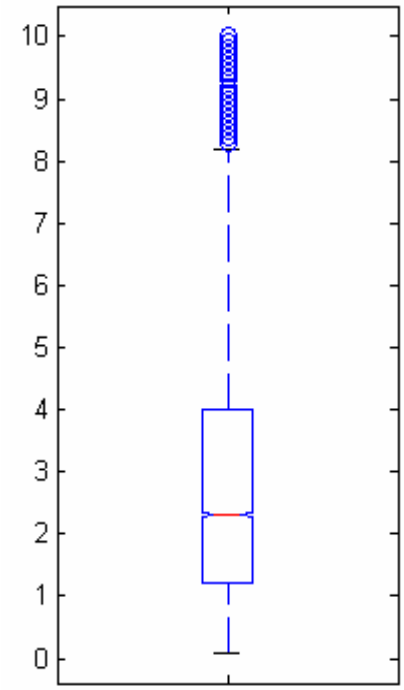
```
%skalowanie wysokości słupków histogramu do wykresu gęstości (nie zbyt
%precyzyjne)
r=max(liczeb)/max(Yi);
liczeb=liczeb/r;
bar(liczebX,liczeb,'b');
plot(Xi,Yi,'r','linewidth',2);
hold off
title(strcat('Wykres gęstości. Szerokość f-kcji bazowej=',...
num2str(szer_fB(i))), 'FontName','Arial Ce','FontSize',9);
end
%*************************************************
```

4. Wykres skrzynkowy

Wykres skrzynkowy jest kompromisem pomiędzy pokazywaniem wyłącznie przeciętnych wartości zmiennej a koncentrowaniem się na wartościach odstających.
Popatrzmy na przykład takiego wykresu na rysunku. Jest na nim przedstawiony wykres skrzynkowy fikcyjnej zmiennej.

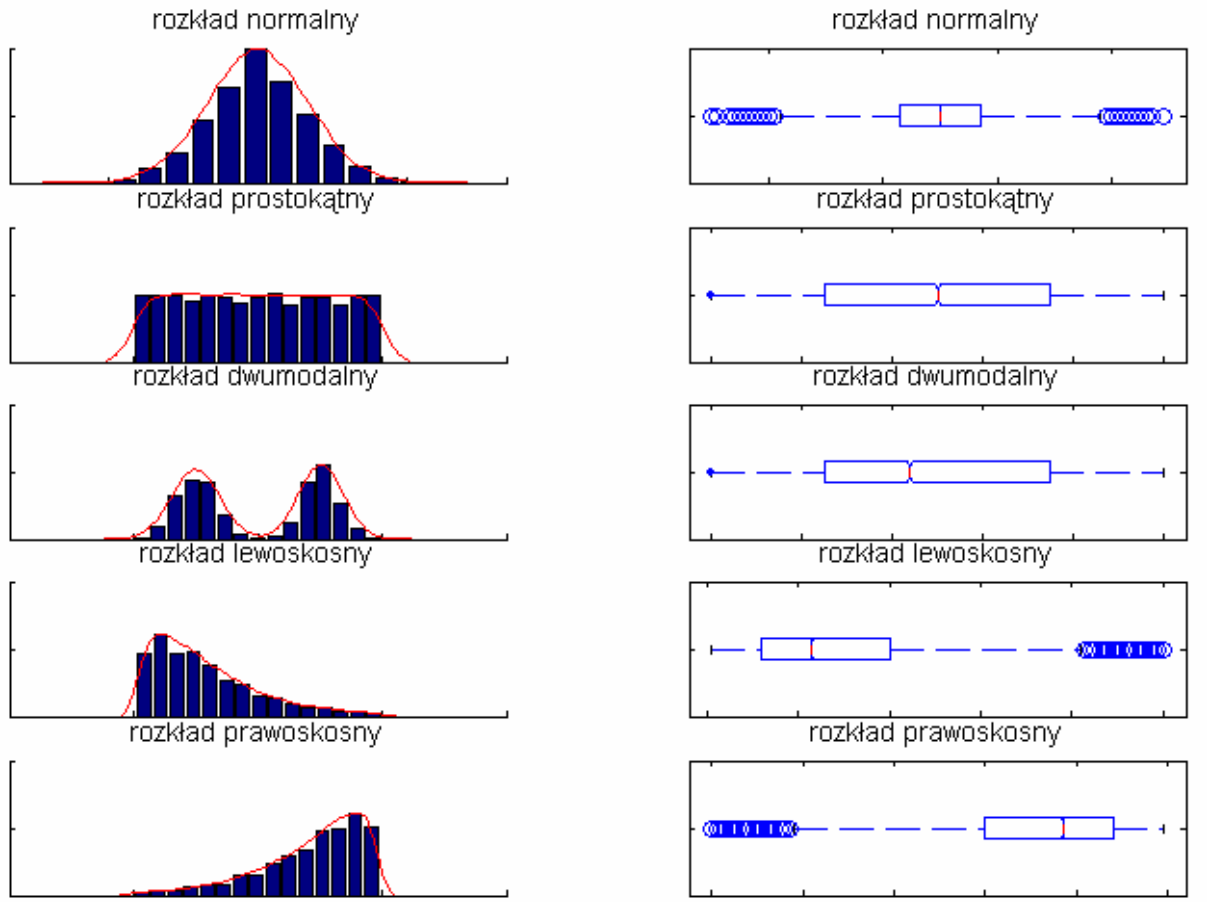
Dolna krawędź skrzynki odpowiada pierwszemu kwartylowi Q1, a górna trzeciemu kwartylowi Q3. Czyli długość skrzynki jest równa odstępowi międzykwartylowemu (IQR). Linia przechodząca przez w poprzek skrzynki odpowiada medianie. Skrzynka daje nam zatem informacje o tym, jak rozłożone są przeciętne wartości zmiennej.

Dodatkowo na wykresie mamy narysowane wąsy zaznaczone przerywaną linią i zakończone krótkimi kreskami. Wąsy wyznaczają najmniejszą i największą wartość zmiennej, o ile nie jest ona większa lub (mniejsza) od Q3 + 1.5 IQR (Q1 - 1.5 IQR). Obserwacje, które są większe od Q3 + 1.5 IQR lub mniejsze od Q1 - 1.5 IQR **nazywamy *obserwacjami odstającymi***; są one oznaczone na wykresie jako odrębne punkty (na rys 6 jako kółka).
Obserwacje, które leżą dalej niż 3 IQR od dolnej lub górnej krawędzi skrzynki nazywamy ***obserwacjami skrajnymi***.



Rysunek 6. Wykres skrzynkowy

Można łatwo porównać ze sobą histogram pewnej fikcyjnej zmiennej oraz wykres skrzynkowy. Warto zwrócić uwagę, że wartości odstające są prawie niewidoczne na histogramie – wskazują na nie jedynie kreski oznaczające przypadki, natomiast bardzo dobrze pokazuje je wykres skrzynkowy.



Rysunek 7. Histogramy i odpowiadające im wykresy skrzynkowe

Jak widzimy, wykres skrzynkowy nie pomoże nam wykryć rozkładu dwumodalnego.

Implementacja w środowisku Matlab

```
function Wyk_Skrzynk()
clc
close all

tab_rozklady=rozklady;
[ile_rozkladow,kol]=size(tab_rozklady);

sub_i=1;
for i=1:ile_rozkladow
[liczebnosc(i,:),x(i,:)]=hist(tab_rozklady{i,2},15);

subplot(5,2,sub_i)
hold on
[fks,xks]=ksdensity(tab_rozklady{i,2});
ileDz=max(liczebnosc(i,:))/max(fks);
```

```
bar(x(i,:),liczebnosc(i,:)./ileDz);
plot(xks,fks,'r','linewidth',1);
set(gca,'fontname','Arial Ce');
title(strcat('rozkład ',tab_rozklady{i,1}));
set(gca,'xticklabel','','yticklabel','');
hold off

subplot(5,2,sub_i+1)
boxplot(tab_rozklady{i,2},1,'o',0,1.5)

sub_i=sub_i+2;
set(gca,'fontname','Arial Ce');
xlabel('');
ylabel('');
set(gca,'xticklabel','','yticklabel','');
title(strcat('rozkład ',tab_rozklady{i,1}));
end
%*************************************************
function tab_rozklady=rozklady()
clc
close all
krok=0.1;
max_X=5;
min_X=-max_X;
N=20000;
X2=0:krok:2*max_X;
X=X2;
%Rozkład normalny
p_norm=normpdf(X,mean(X),1);
r=1/sum(p_norm);
p_norm=p_norm*r;
rozkl_norm = randsrc(1,N,[X;p_norm]);

%Rozkład prostoątny
p_prost=ones(1,length(X))*(1/length(X));
rozkl_prost = randsrc(1,N,[X;p_prost]);

%Rozkład dwumodalny
X01=0:krok:max_X;
X02=max_X+krok:krok:2*max_X;
r_norm1=normpdf(X01,mean(X01),0.8);
```

```
r_norm2=normpdf(X02,mean(X02),0.7);
p_dwumod=[r_norm1 r_norm2];
r=1/sum(p_dwumod);
p_dwumod=p_dwumod*r;
rozkl_dwumodalny = randsrc(1,N,[X;p_dwumod]);

%Rozkład lewoskośny
% p_lewo_skosny = fpdf(X2,5,3);
p_lewo_skosny = chi2pdf(X2,3);
r=1/sum(p_lewo_skosny);
p_lewo_skosny=p_lewo_skosny*r;
rozkl_lewo_skosny = randsrc(1,N,[X;p_lewo_skosny]);

%Rozkład prawoskośny
p_prawo_skosny = p_lewo_skosny(length(p_lewo_skosny):-1:1);
rozkl_prawo_skosny = randsrc(1,N,[X;p_prawo_skosny]);

tab_rozklady{1,1}=' normalny';
tab_rozklady{1,2}=rozkl_norm;
tab_rozklady{2,1}=' prostokątny';
tab_rozklady{2,2}=rozkl_prost;
tab_rozklady{3,1}=' dwumodalny';
tab_rozklady{3,2}=rozkl_dwumodalny;
tab_rozklady{4,1}=' lewoskosny';
tab_rozklady{4,2}=rozkl_lewo_skosny;
tab_rozklady{5,1}=' prawoskosny';
tab_rozklady{5,2}=rozkl_prawo_skosny;
%************************************************
```

5. Wykres kwantylowy

Wykres kwantylowy (lub centylowy) służy przede wszystkim do tego, aby sprawdzić, na ile rozkład badanej zmiennej odpowiada jakiemuś rozkładowi teoretycznemu, na przykład normalnemu. Bardzo często w statystyce taka informacja jest naprawdę istotna - wiele testów opiera się na założeniu, że pewna zmienna losowa ma rozkład normalny.

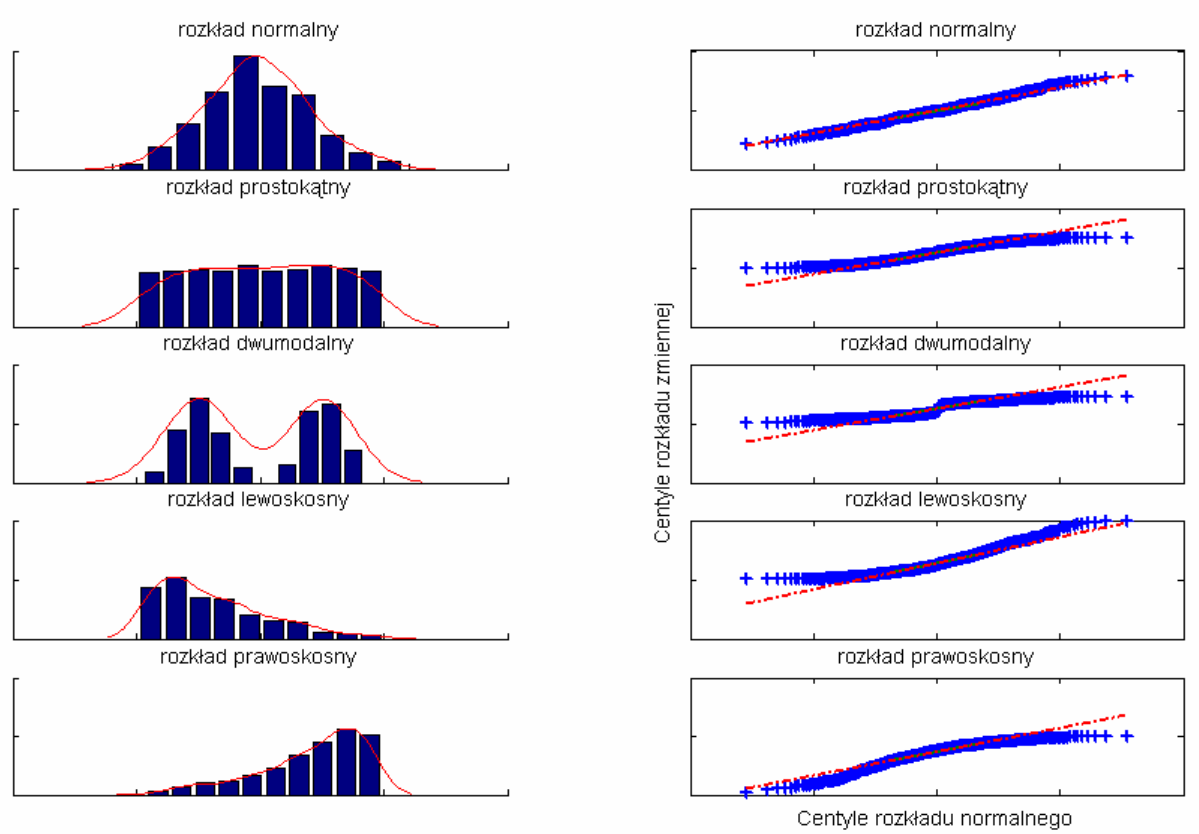
Idea wykresu kwantylowego bazuje na na bardzo prostej zasadzie. Oznaczamy przez X_i centyl i-tego rzędu zmiennej mającej rozkład normalny o średniej μ i odchyleniu standardowym σ. Centyl tego samego rzędu ***standaryzowanego*** rozkładu normalnego z_i wiąże się X_i w następujący sposób:

$$z_i = \frac{X_i - m}{s}$$

Ponieważ transformacja ta jest liniowa, to punkty (X_i, z_i) leżą na jesdnej prostej.

Wykres kwantylowy tworzymy zatem, wykreślając na jednej osi centyle pochodzące z rozkładu normalnego, a na drugiej centyle badanej przez nas zmiennej (uprzednio standaryzowane, po to aby miały te same wartości średnie i odchylenia). Jeżeli punkty układają się na jednej prostej, to znaczy, że nasza zmienna ma rozkład normalny.

Jeżeli okaże się, że rozkład zmiennej nie jest normalny, to przy odrobinie wprawy możemy się przekonać, jakie są jego podstawowe właściwości: skośność oraz liczba modalnych. Przyjrzyjmy się następującym histogramom i odpowiadającym im wykresom kwantylowym.



Rysunek 8. Histogramy i wykresy kwantylowy rozkładów o różnych kształtach

Implementacja w środowisku Matlab

```
function Wyk_Kwant()
clc
close all

tab_rozklady=rozklady; %funkcja znajduje się powyżej
[ile_rozkladow,kol]=size(tab_rozklady);

sub_i=1;
for i=1:ile_rozkladow
[liczebnosc(i,:),x(i,:)]=hist(tab_rozklady{i,2},15);
```

```
subplot(5,2,sub_i)
hold on
[fks,xks]=ksdensity(tab_rozklady{i,2});
ileDz=max(liczebnosc(i,:))/max(fks);
bar(x(i,:),liczebnosc(i,:)./ileDz);
plot(xks,fks,'r','linewidth',1);
set(gca,'fontname','Arial Ce');
title(strcat('rozkład ',tab_rozklady{i,1}));
set(gca,'xticklabel','','yticklabel','');
hold off

subplot(5,2,sub_i+1)
qqplot(tab_rozklady{i,2});

sub_i=sub_i+2;
set(gca,'fontname','Arial Ce');
xlabel('');
ylabel('');
set(gca,'xticklabel','','yticklabel','');
title('');
      if(i==5)
      xlabel('Centyle rozkładu normalnego');
      ylabel('');
      elseif(i==3)
      xlabel('');
      ylabel('Centyle rozkładu zmiennej');
      end
title(strcat('rozkład ',tab_rozklady{i,1}));
end
%*************************************************
```

II. Wizualizacja zależności między zmiennymi

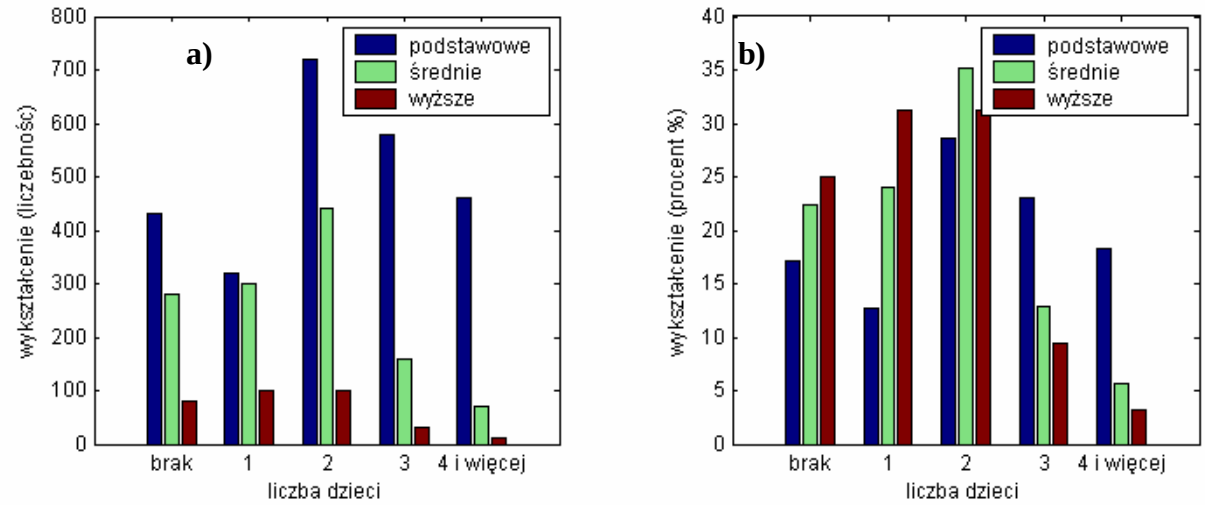
1. Wykres słupkowy zgrupowany

Gdy badamy dwie zmienne mierzone na skali nominalnej, to możemy zrobić zgrupowany wykres słupkowy – pokazuje on informację zawartą w tabeli krzyżowej (rys. 9).

Liczebności kategorii mogą być mylące. Na rys. 9a wykreślone są liczebności, a na rys. 9b procenty. Drugi rysunek jest o wiele lepszy, gdyż kategorie zmiennych nie są równoliczne (np. osób z wykształceniem podstawowym jest dużo więcej niż z wyższym). Trzeba o tym pamiętać, zwłaszcza badając dane sondażowe, gdzie kategorie rzadko są równoliczne.

Liczba dzieci	wykształcenie	podstawowe	średnie	wyższe	ogółem
brak	***Liczebność***	430,00	280,00	80,00	790,00
	Częstość	0,54	0,35	0,10	1,00
	Procent %	54,43	35,44	10,13	100,00
1	***Liczebność***	320,00	300,00	100,00	720,00
	Częstość	0,44	0,42	0,14	1,00
	Procent %	44,44	41,67	13,89	100,00
2	***Liczebność***	720,00	440,00	100,00	1260,00
	Częstość	0,57	0,35	0,08	1,00
	Procent %	57,14	34,92	7,94	100,00
3	***Liczebność***	580,00	160,00	30,00	770,00
	Częstość	0,75	0,21	0,04	1,00
	Procent %	75,32	20,78	3,90	100,00
4 i więcej	***Liczebność***	460,00	70,00	10,00	540,00
	Częstość	0,85	0,13	0,02	1,00
	Procent %	85,19	12,96	1,85	100,00
Ogółem:	**Liczebność**	**2510,00**	**1250,00**	**320,00**	**4080,00**
	Częstość	**0,62**	**0,31**	**0,08**	**1,00**
	Procent %	**61,52**	**30,64**	**7,84**	**100,00**

Tabela 2. Tabela krzyżowa



Rysunek 9. Zgrupowany wykres słupkowy, a)słupki przedstawiają liczebność, b) słupki przedstawiają Procent

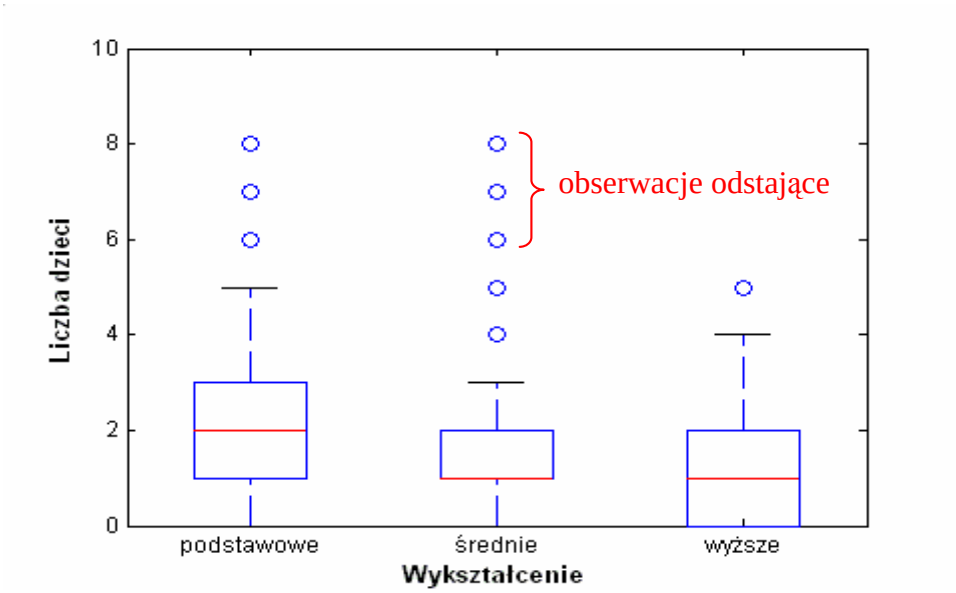
Implementacja w środowisku Matlab

```
function Wyk_SlpGrp()
kategorie={'brak','1','2','3','4 i więcej'};
legenda ={'podstawowe', 'średnie', 'wyższe'};
tekst={'liczebność', 'procent %'};

a=[430 320 720 580 460];
b=[280 300 440 160 70];
```

```
c=[80 100 100 30 10];

a_S=a./ sum(a)*100;
b_S=b./ sum(b)*100;
c_S=c./ sum(c)*100;

d{1}=[a' b' c'];
d{2}=[a_S' b_S' c_S'];

for i=1:2
subplot(1,2,i)
bar(d{i});
set(gca,'xticklabel',kategorie,'fontname','Arial Ce','fontsize',9);
xlabel('liczba dzieci');
ylabel(['wykształcenie (', tekst{i},')']);
legend(legenda);
end
%*************************************************
```

2. Wykresy skrzynkowe zgrupowane

Gdy przynajmniej jedna zmienna jest mierzona na skali porządkowej, to możemy analizować związek zmiennych za pomocą wykresów skrzynkowych. Jest to użyteczne zwłaszcza gdy zmienne mają więcej kategorii. Na rysunku 10 widzimy wykres liczby dzieci w zależności od wykształcenia.



Rysunek 10. Liczba dzieci w zależności od wykształcenia

Na wykresie pojawiają się obserwacje odstające (oznaczone kółkiem) oraz skrajne (Matlab nie umożliwia jednoczesnego oznaczenia obserwacji skrajnych i odstających - różnymi symbolami - na jednym wykresie)

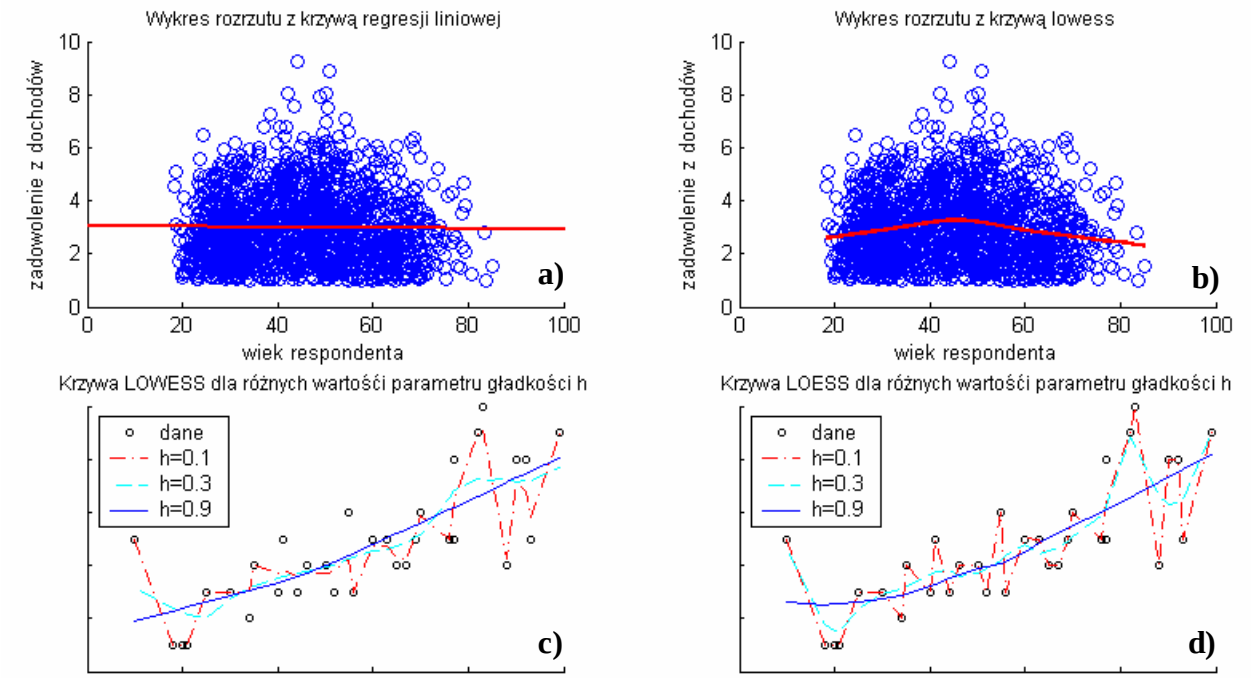
Implementacja w środowisku Matlab

```
function Wyk_SkrzGrup()
legenda={'podstawowe', 'średnie', 'wyższe'};
liczba_dzieci=[0 1 2 3 4 5 6 7 8];
podst=[250 400 1000 450 350 4 3 2 1];
sredn=[400 1300 500 200 50 7 1 1 1];
wyzsze=[700 800 600 300 50 10 0 0 0];

dane1=gest2dane(liczba_dzieci, podst);
dane2=gest2dane(liczba_dzieci, sredn);
dane3=gest2dane(liczba_dzieci, wyzsze);
dane=[dane1' dane2' dane3'];

hold on
boxplot(dane,0,'o',1,1.2)
ylim([0 10]);
set(gca,'xticklabel',legenda,'fontname','Arial Ce','fontsize',9);
xlabel('Wykształcenie','FontWeight','bold');
ylabel('Liczba dzieci','FontWeight','bold');
hold off
%*************************************************

function dane=gest2dane(x,y)
%zwraca wektor utworzony z wektorów x i y, gdzie
%x - kategorie; y – liczebności
%np. x=[1 3 4]; y=[2 1 3]; ->  dane=[1 1 3 4 4 4];
m=length(x);
dane=-999;
    for (j=1:m)
        dane_cz=-999;
        for (k=1:y(1,j))
            dane_cz(k)=x(j);
        end
            if dane==-999
            dane=[dane_cz];
            elseif dane_cz~=-999
            dane=[dane dane_cz];
```

```
            end
        end
%*************************************************
```

3. Wykres rozrzutu (korelacyjny)

Wykres ten jest bardzo prosty zarówno do zrobienia jak i do interpretacji. Jeżeli mamy dwie zmienne X i Y, to na wykres nanosimy punkty o współrzędnych (Xi,Yi) gdzie i jest numerem przypadku.



Rysunek 11. Wykres rozrzutu, a) z krzywą regresji liniowej, b) z krzywą Lowess, c) i d) porównanie krzywych Lowess i Loess przy różnej wartości parametru gładkości h

Badanie trendu. Krzywa regresji. Krzywa LOWESS (LOESS)

Najprostszym i często stosowanym sposobem badania trendu jest dopasowanie do danych lini prostej, pochodzącej z regresji liniowej (metoda najmniejszych kwadratów). Na rysunku 11a mamy wykreśloną linię regresji. Kąt nachylenia linii wskazuje na to, że nie istnieje zależność liniowa między wiekiem respondenta a jego zadowoleniem z dochodów. Ponieważ często brak jest liniowej zależności między zmiennymi wymyślono lepszą metodę wizualizacji trendu – krzywe LOWESS i LOESS.

Sposób robienia tego wykresu jest bardzo prosty. W pobliżu każdego punktu na wykresie (w pobliżu każdej obserwacji) dopasowujemy do danych pewien wielomian niskiego stopnia (prosta - krzywe Lowess lub parabola - krzywe Loess). Robimy to jednak w szczególny sposób, otóż nie bierzemy przy dopasowaniu pod uwagę wszystkich obserwacji a tylko te, które są odpowiednio blisko. Ponadto te obserwacje, które są dalej od punktu estymacji mają mniejszą wagę niż te, które są bliżej. O tym ile obserwacji będzie wzięte pod uwagę decyduje parametr gładkości h, który dobieramy samodzielnie.

Rysunek 11b przedstawia ten sam wykres rozrzutu , tym razem z dopasowaną do naszych danych krzywą Lowess. Widzimy, że wskazuje ona na istnienie zależności między zmiennymi, tyle że krzywoliniowej.

Gdy wspomniany parametr h przyjmuje minimalną wartość 0 – wówczas przy dopasowaniu krzywej pod uwagę wzięte są wszystkie obserwacje – otrzymujemy w ten sposób bardzo gładką krzywą, która zazwyczaj słabo pokazuje trend. Gdy h jest większe, to do dopasowania krzywej używany jest odpowiedni ułamek danych (np. 0.3 - 30% danych ignorowanych, pod uwagę branych jest 70% danych). Zależność wyglądu krzywej od doboru tego parametru jest zobrazowana na rysunku 11c i 11d.

Jedyną wadą krzywej LOWESS stanowi to, że jej dopasowanie wymaga dość dużego zbioru danych (obserwacje muszą być blisko siebie)

Implementacja w środowisku Matlab

```
function WykrKorelac()
close all
clc
Z=1:0.01:10
X=18:.01:100;
N=500;

p_normX1=normpdf(X,45,6);
r=1/sum(p_normX1);
p_normX1=p_normX1*r;

p_normZ1=normpdf(Z,3,1.8);
r=1/sum(p_normZ1);
p_normZ1=p_normZ1*r;

p_normX2=normpdf(X,30,5);
r=1/sum(p_normX2);
p_normX2=p_normX2*r;

p_normZ2=normpdf(Z,2,1.7);
r=1/sum(p_normZ2);
p_normZ2=p_normZ2*r;

p_normX3=normpdf(X,60,9);
r=1/sum(p_normX3);
p_normX3=p_normX3*r;

p_normZ3=normpdf(Z,2,1.7);
```

```
r=1/sum(p_normZ3);
p_normZ3=p_normZ3*r;

rozkladX1 = randsrc(1,N,[X;p_normX1]);
rozkladZ1 = randsrc(1,N,[Z;p_normZ1]);

rozkladX2 = randsrc(1,N,[X;p_normX2]);
rozkladZ2 = randsrc(1,N,[Z;p_normZ2]);

rozkladX3 = randsrc(1,N,[X;p_normX3]);
rozkladZ3 = randsrc(1,N,[Z;p_normZ3]);

rozkladX=[rozkladX1,rozkladX2,rozkladX3];
rozkladZ=[rozkladZ1,rozkladZ2,rozkladZ3];
%---------------------------------------
subplot(2,2,1)
hold on
plot(rozkladX,rozkladZ,'o')
h=lsline
set(h,'color','r','linewidth',2);
set(gca, 'FontName','Arial CE','fontSize',9);
title('Wykres rozrzutu z krzywą regresji liniowej');
xlabel('wiek respondenta');
ylabel('zadowolenie z dochodów');
hold off
%---------------------------------------
subplot(2,2,2)
hold on
plot(rozkladX,rozkladZ,'o')
[a,b]=sortuj2kol(rozkladX,rozkladZ)
yy = smooth(a,b,0.5,'lowess')
plot(a,yy,'r','linewidth',2);
set(gca, 'FontName','Arial CE','fontSize',9);
title('Wykres rozrzutu z krzywą lowess');
xlabel('wiek respondenta');
ylabel('zadowolenie z dochodów');
hold off
%---------------------------------------
a1=[2 1 3 5 2.1 3.5 6 7 4 5.6 1.8 6.3 7.7 8.3 9 6.5 9.9 7.7 8.8 4.4 5.2 6.7
7.6 8.2 9.2 3.4 2.5 4.6 4.1 5.5 6.9 9.3];
b1=[1 5 3 4 1 4 5 6 3 3 1 5 5 10 8 4 9 8 4 3 3 4 5 9 8 2 3 4 5 6 5 5];
```

```
[a,b]=sortuj2kol(a1,b1)

subplot(2,2,3)
hold on
plot(a,b,'ok','markersize',4)
yy1 = smooth(a,b,0.1,'lowess')
yy2 = smooth(a,b,0.3,'lowess')
yy3 = smooth(a,b,0.9,'lowess')
plot(a,yy1,'-.r','linewidth',1);
plot(a,yy2,'--c','linewidth',1);
plot(a,yy3,'-b','linewidth',1);
set(gca, 'FontName','Arial CE','fontSize',9);
legend('dane','h=0.1','h=0.3','h=0.9',2);
title('Krzywa LOWESS dla różnych wartości parametru h');
hold off
%--------------------------------------
subplot(2,2,4)
hold on
plot(a,b,'ok','markersize',4)
yy1 = smooth(a,b,0.1,'loess')
yy2 = smooth(a,b,0.3,'loess')
yy3 = smooth(a,b,0.9,'loess')
plot(a,yy1,'-.r','linewidth',1);
plot(a,yy2,'--c','linewidth',1);
plot(a,yy3,'-b','linewidth',1);
set(gca,'FontName', 'Arial CE', 'fontSize', 9);
legend('dane', 'h=0.1', 'h=0.3', 'h=0.9', 2);
title('Krzywa LOESS dla różnych wartości parametru h');
hold off
%**************************************************

function [x,y]=sortuj2kol(x,y)
[x pos]=sort(x);
  for i=1:length(y)
    z(i)=y(pos(i));
  end
y=z;
%**************************************************
```

III. Podsumowanie

Przekonaliśmy się, że możliwości graficznej analizy i prezentacji danych są bardzo duże. Trzeba z nich mądrze korzystać, pamiętając, jakie są wady i zalety poszczególnych typów wykresów. Najwięcej informacji przydatnych do analitycznego badania danych dostarczają wykres skrzynkowy oraz wykres kwantylowy – najlepiej więc użyć ich obu.
Dla zmiennych o małej liczbie kategorii niezawodny jest wykres słupkowy.
Gdy chcemy otrzymane przez nas wyniki pokazać osobom, które nie zajmują się statystyką, to nie należy oczywiście prezentować im wykresu skrzynkowego czy kwantylowego, a na przykład dobrze zrobiony histogram czy wykres słupkowy.

Spis treści