Elementy Modelowania Matematycznego
Wykład 3
Wykresy

Romuald Kotowski

Katedra Informatyki Stosowanej

PJWSTK 2009

Spis treści

Spis treści

Spis treści

Wykresy dla danych jakościowych

Tabele danych

Województwo mazowieckie											
2000						2006					
10-letnia grupa wiekowa	Ogółem	Mężczyzna		Kobieta		10-letnia grupa wiekowa	Ogółem	Mężczyzna		Kobieta	
		Miasto	Wieś	Miasto	Wieś			Miasto	Wieś	Miasto	Wieś
0 - 14	915883	266282	203102	253057	193442	0 - 14	797 058	237 844	170 252	226 364	162 598
15 - 19	419363	135330	78956	131931	73146	15 - 19	346 069	102 273	75 235	98 416	70 145
20 - 29	791637	268013	133739	269142	120743	20 - 29	846 081	276 514	150 734	281 937	136 896
30 - 39	642105	201277	123313	207446	110069	30 - 39	742 299	246 713	125 488	253 631	116 467
40 - 59	1401797	451414	223596	522431	204356	40 - 59	1 463 684	454 867	253 154	525 254	230 409
60 - 64	234227	67183	35903	89702	41439	60 - 64	223 020	67 318	31 457	87 790	36 455
65 i więcej	709998	167064	104439	277635	160860	65 i więcej	753 491	181 417	103 623	306 362	162 089
OGÓŁEM	5115010	1556563	903048	1751344	904055	OGÓŁEM	5 171 702	1 566 946	909 943	1 779 754	915 059

Rys. 1: Liczebności kobiet i mężczyzn w województwie mazowieckim w latach 2000 i 2006 (GUS)

Wykresy dla danych jakościowych

Tabele danych

Województwo mazowieckie					
2000					
10-letnia grupa	Ogółem	Mężczyzna		Kobieta	
		Miasto	Wieś	Miasto	Wieś
0 - 14	915 883	266 282	203 102	253 057	193 442
15 - 19	419 363	135 330	78 956	131 931	73 146
20 - 29	791 637	268 013	133 739	269 142	120 743
30 - 39	642 105	201 277	123 313	207 446	110 069
40 - 59	1 401 797	451 414	223 596	522 431	204 356
60 - 64	234 227	67 183	35 903	89 702	41 439
65 i więcej	709 998	167 064	104 439	277 635	160 860
OGÓŁEM	5 115 010	1 556 563	903 048	1 751 344	904 055

Rys. 2: Liczebności kobiet i mężczyzn w województwie mazowieckim w roku 2000 (GUS)

Wykresy dla danych jakościowych

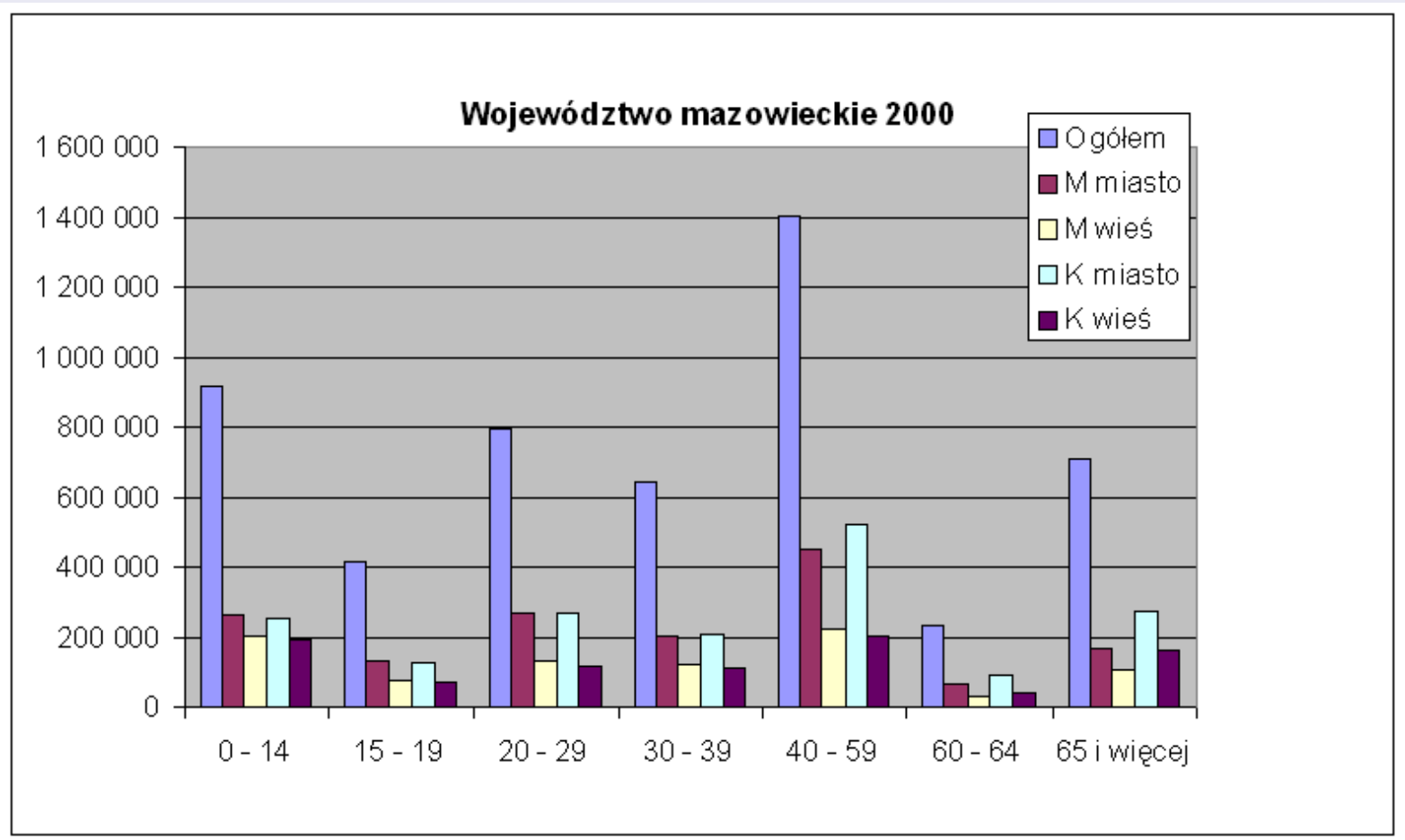
Tabele danych

Województwo mazowieckie					
2006					
10-letnia grupa	Ogółem	Mężczyzna		Kobieta	
		Miasto	Wieś	Miasto	Wieś
0 - 14	797 058	237 844	170 252	226 364	162 598
15 - 19	346 069	102 273	75 235	98 416	70 145
20 - 29	846 081	276 514	150 734	281 937	136 896
30 - 39	742 299	246 713	125 488	253 631	116 467
40 - 59	1 463 684	454 867	253 154	525 254	230 409
60 - 64	223 020	67 318	31 457	87 790	36 455
65 i więcej	753 491	181 417	103 623	306 362	162 089
OGÓŁEM	5 171 702	1 566 946	909 943	1 779 754	915 059

Rys. 3: Liczebności kobiet i mężczyzn w województwie mazowieckim w roku 2006 (GUS)

Wykresy dla danych jakościowych

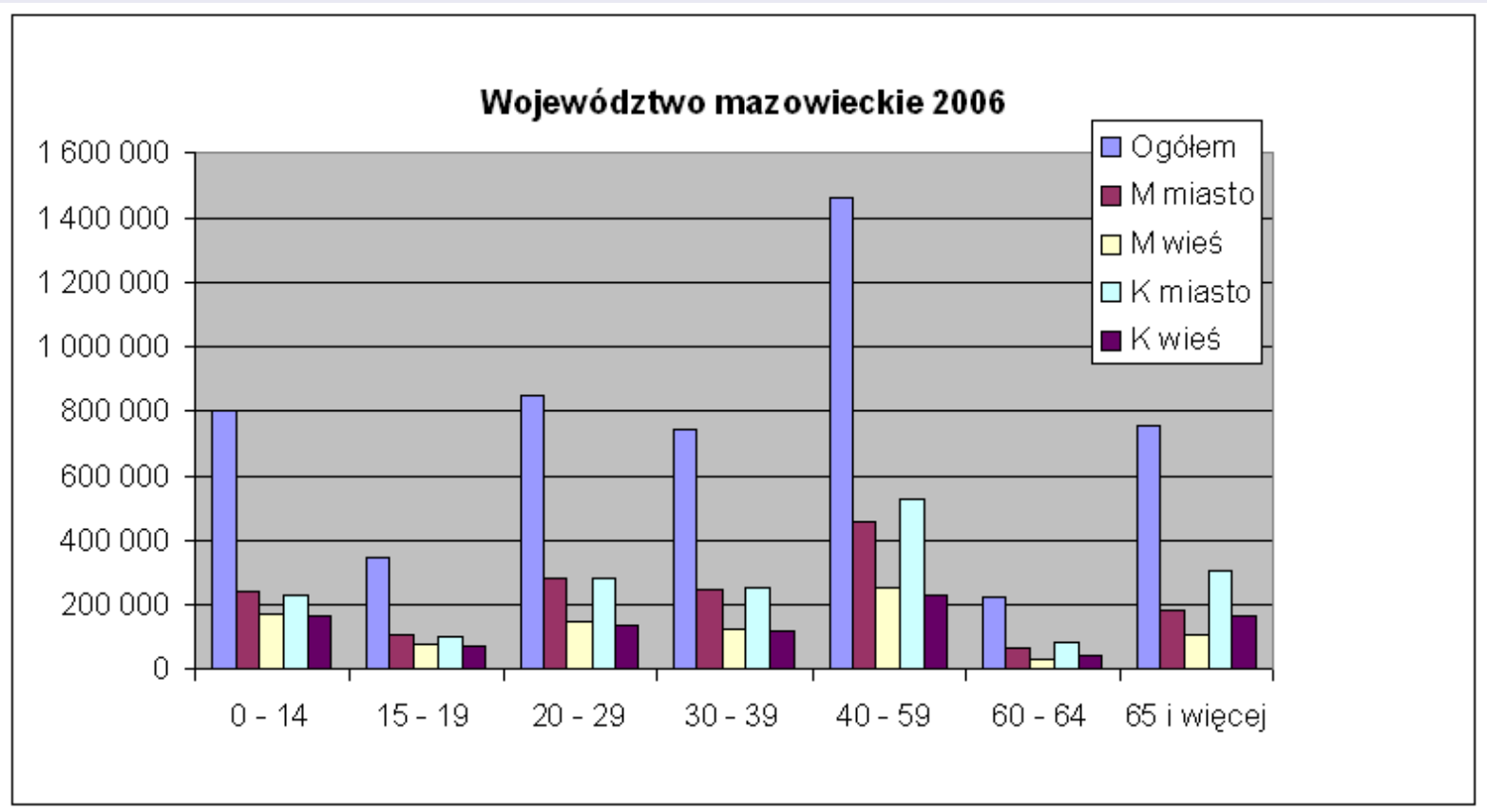
Wykresy słupkowe



Rys. 4: Liczebności kobiet i mężczyzn w województwie mazowieckim w roku 2000

Wykresy dla danych jakościowych

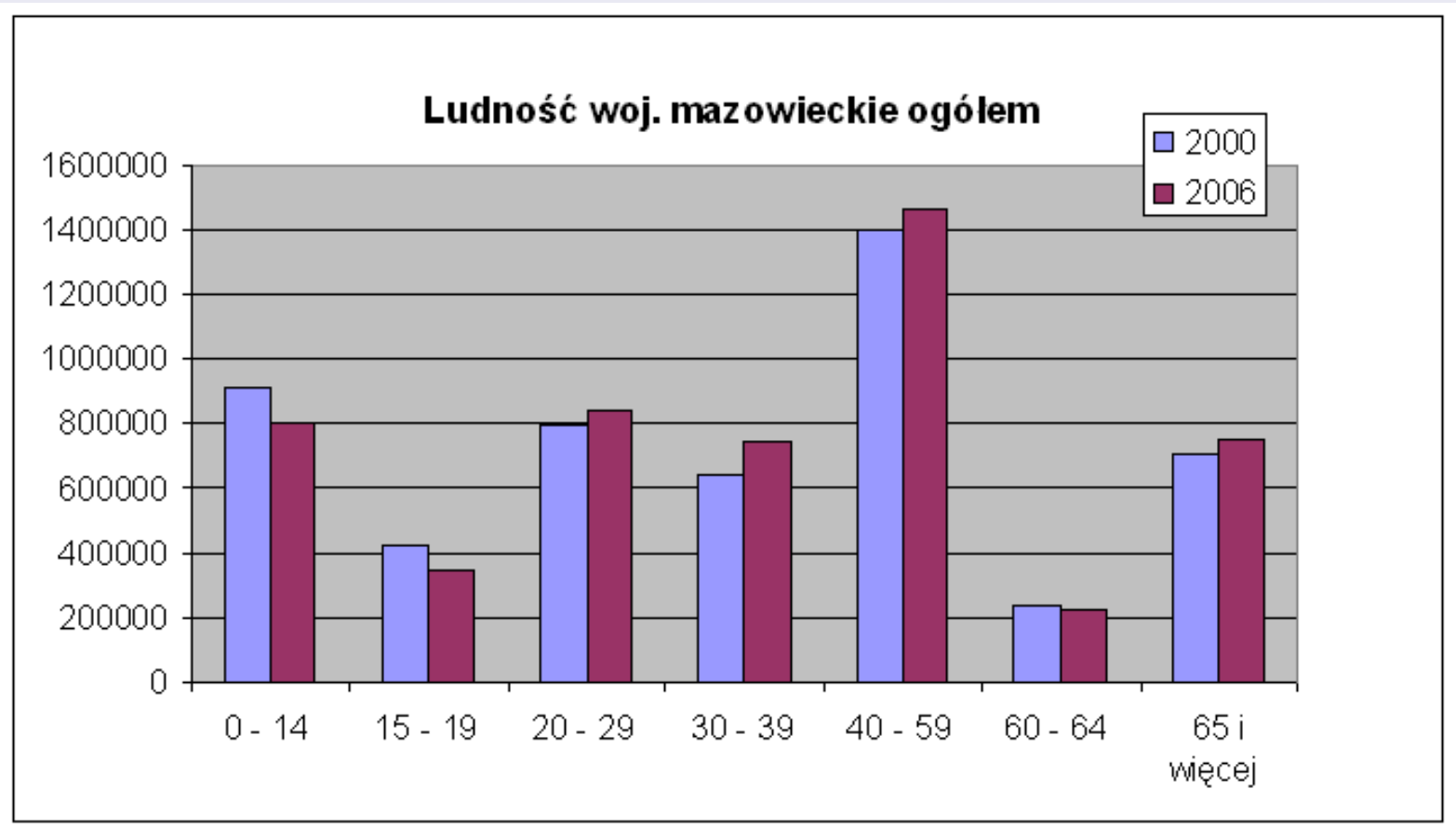
Wykresy słupkowe



Rys. 5: Liczebności kobiet i mężczyzn w województwie mazowieckim w roku 2006 (GUS)

Wykresy dla danych jakościowych

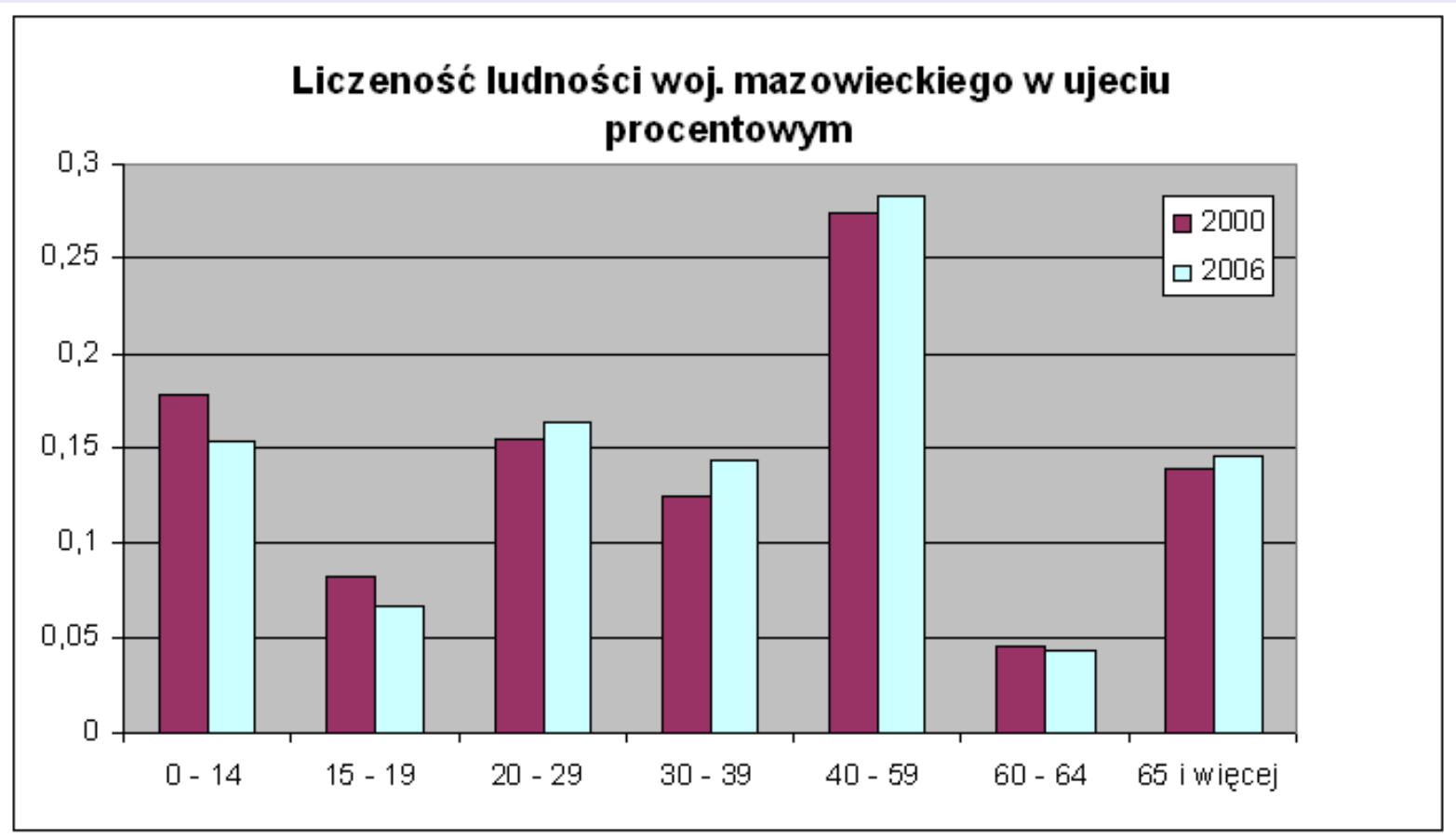
Wykresy słupkowe



Rys. 6: Liczebność ludności w województwie mazowieckim w latach 2000 i 2006 (GUS)

Wykresy dla danych jakościowych

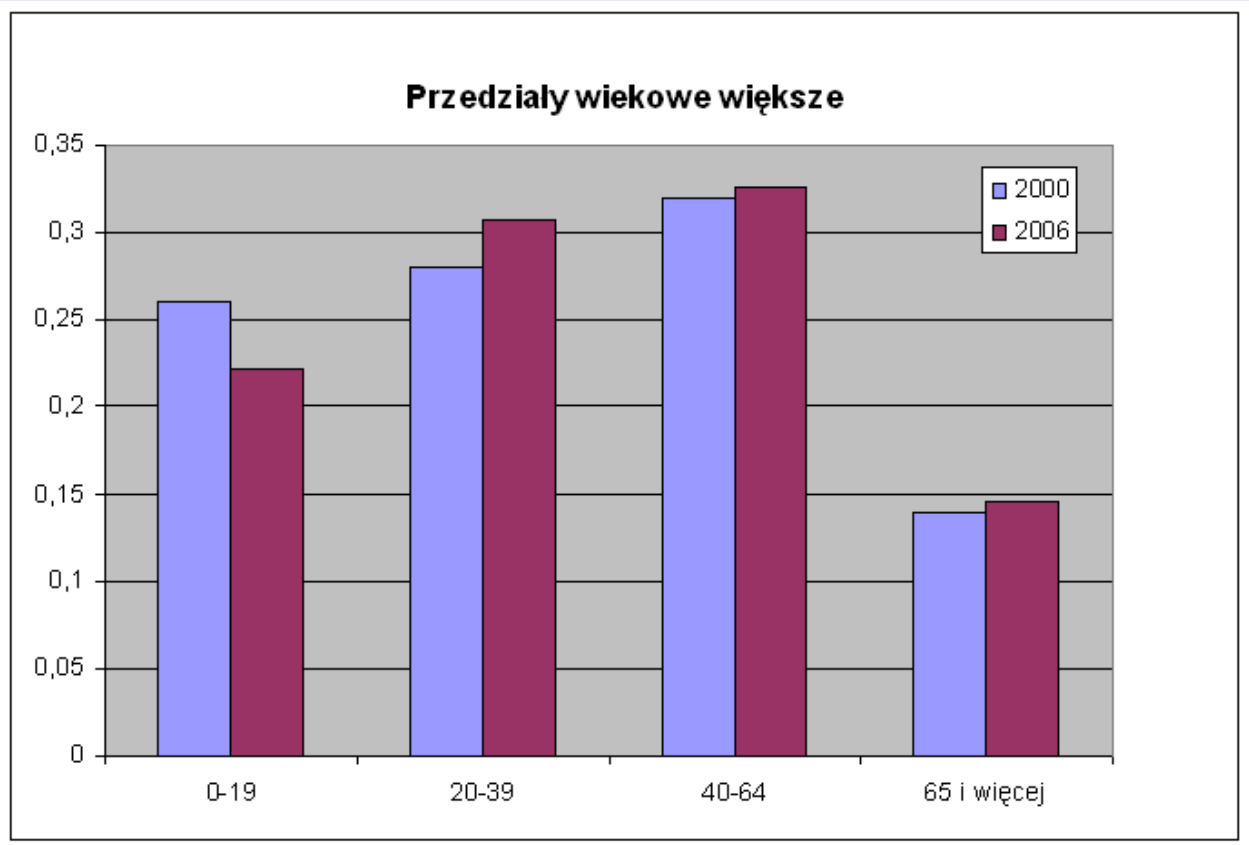
Wykresy słupkowe



Rys. 7: Liczebność ludności w województwie mazowieckim w latach 2000 i 2006 w ujeciu procentowym (GUS)

Wykresy dla danych jakościowych

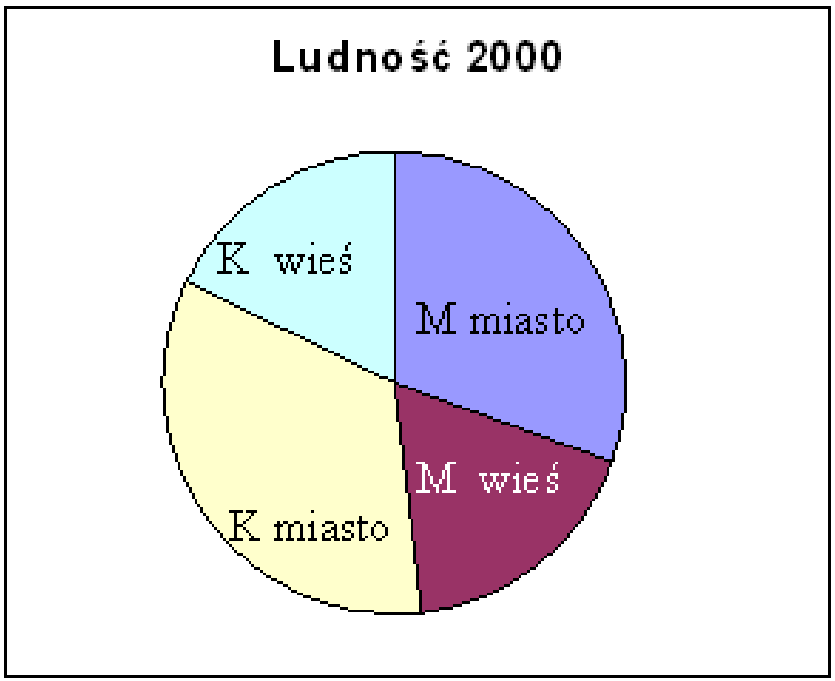
Wykresy słupkowe

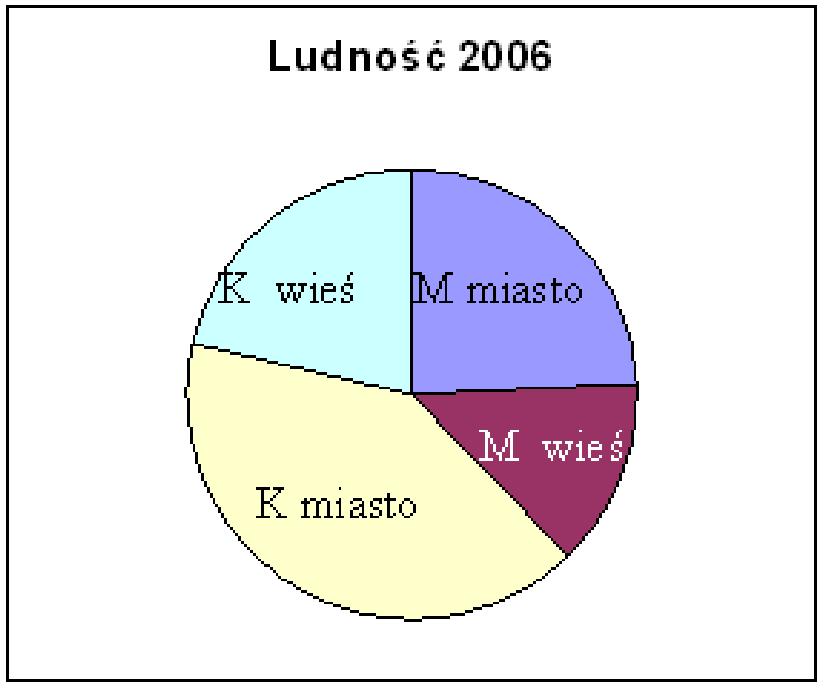


Rys. 8: Liczebność ludności w województwie mazowieckim w latach 2000 i 2006 w ujęciu procentowym w szerszych przedziałach wiekowych (GUS)

Wykresy dla danych jakościowych

Wykresy tortowe





Rys. 9: Ludność w województwie mazowieckim w latach 2000 i 2006 (GUS)

Wykresy dla danych jakościowych

Histogramy

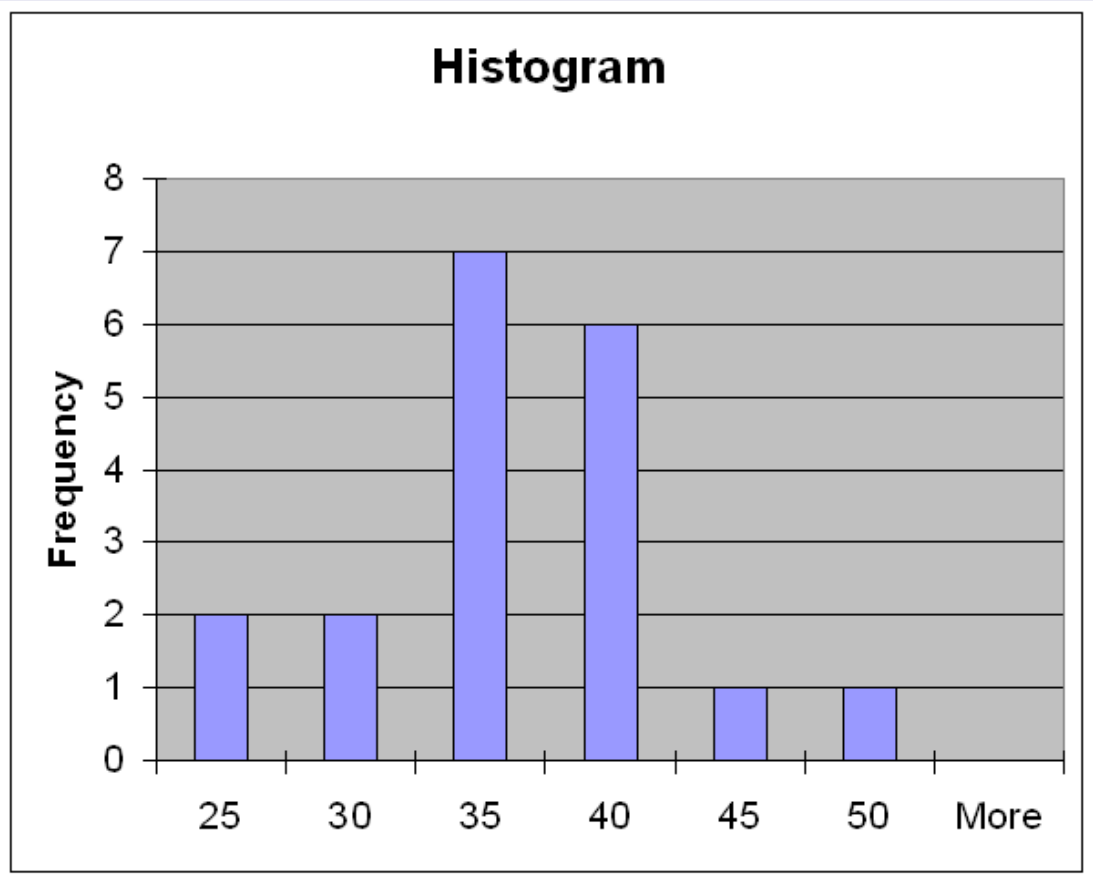
Przykład 1

Rejestrujemy wiek 20 pracowników zgłaszających się na okresowe badania w pewnym zakładzie pracy. Zaobserwowane wielkości wynoszą (w latach):
36, 41, 33, 34, 39, 26, 33, 36, 30, 49, 39, 31, 35, 36, 39, 37, 22, 31, 25, 32.

Najmłodszy z pracowników w próbie ma 22 lata, a najstarszy 49 lat, więc możemy na przykład rozpatrzeć następujące przedziały wiekowe:
$[20, 25), [25, 30), [30, 35), [35, 40), [40, 45), [45, 50)$.

Wykresy dla danych jakościowych

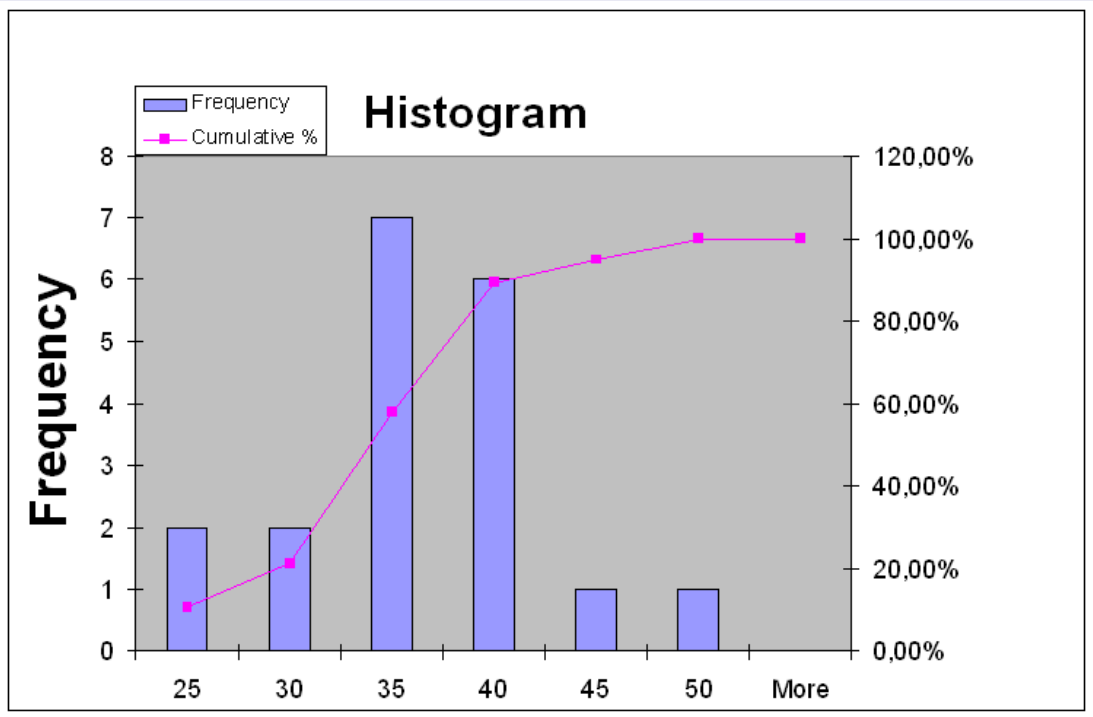
Histogramy



Rys. 10: Histogram wieku pracowników

Wykresy dla danych jakościowych

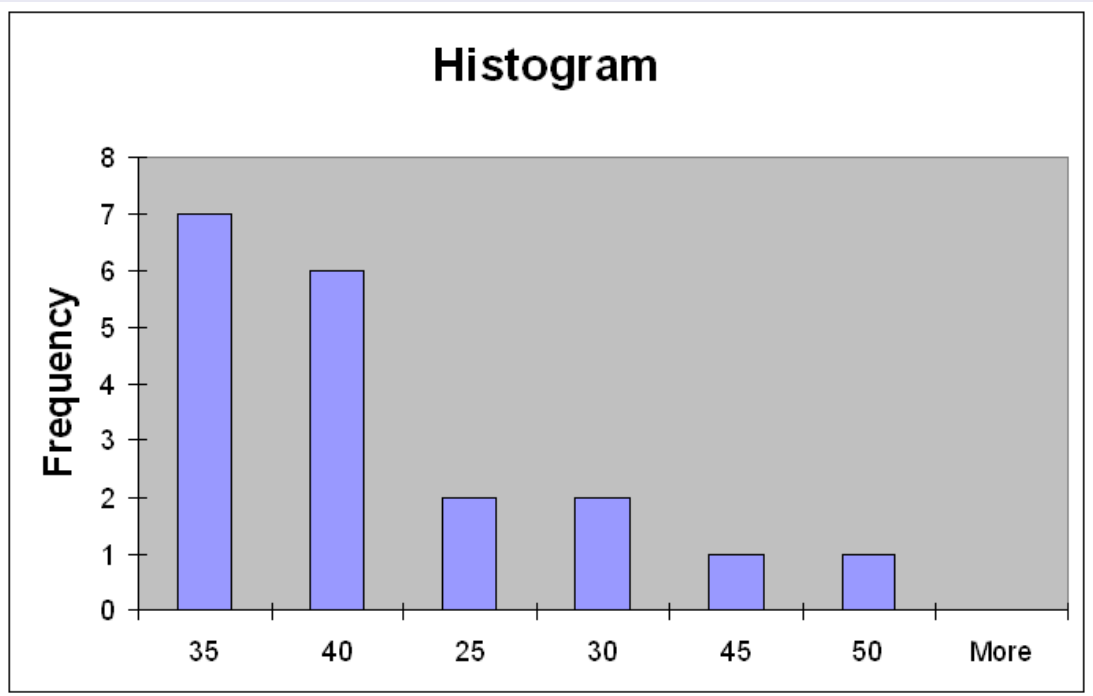
Histogramy



Rys. 11: Histogram wieku pracowników skumulowany

Wykresy dla danych jakościowych

Histogramy



Rys. 12: Histogram Pareto wieku pracowników

Wykresy dla danych jakościowych

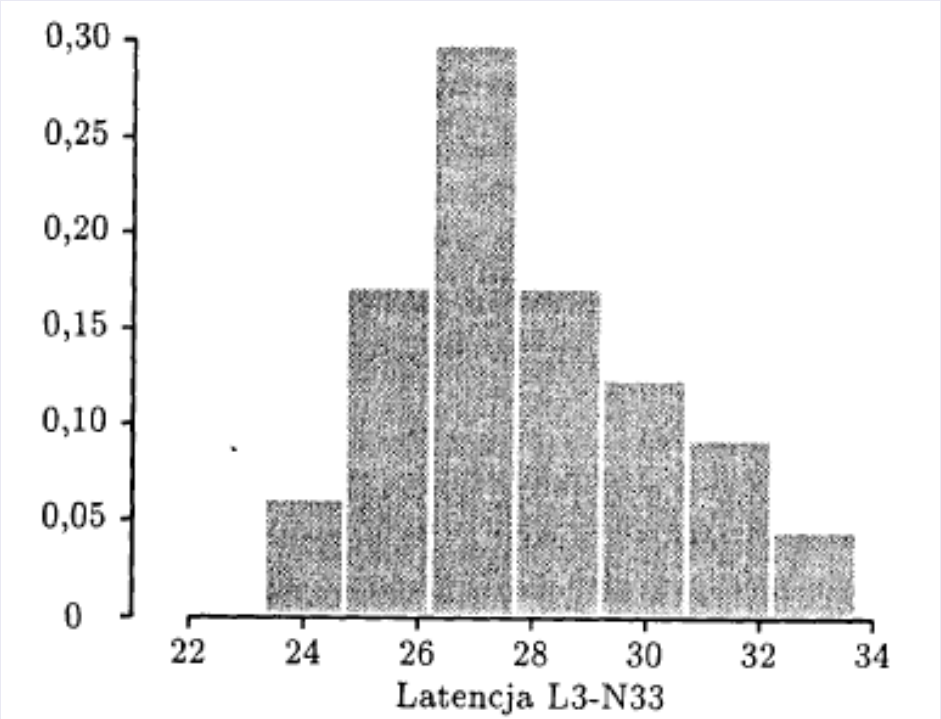
Histogramy

Przykład 2

W badaniu jest rejestrowany potencjał wzbudzony w kończynie lewej. Rozpatrywaną cechą jest jedna z charakterystyk potencjału zwana latencją L3-N33: jest to czas od momentu wzbudzenia potencjału w tzw. korzeniu L3 do osiągnięcia przez potencjał pierwszego maksimum lokalnego. Dane zebrane dla 62 pacjentów (w milisekundach) są następujące:
26,40 31,60 29,60 28,20 24,80 26,50 25,85 26,10 26,90 26,05 31,40
28,00 25,55 29,70 26,80 28,80 26,50 28,30 30,50 24,70 25,30 30,20
29,20 28,40 26,90 25,50 26,40 33,00 25,20 26,60 27,50 25,10 24,60
31,80 29,80 27,90 30,20 26,50 31,60 25,60 26,50 27,50 28,40 27,10
30,90 30,30 30,10 28,70 27,60 27,60 28,70 32,90 26,30 26,30 27,40
26,80 24,20 28,70 31,50 26,00 32,60 24,60

Wykresy dla danych jakościowych

Histogramy



Rys. 13: Histogram częstości dla danych z Przykładu 2

Wykresy dla danych jakościowych

Histogramy

Zbudowaliśmy histogram oparty na przedziałach o długości 1,5 milisekundy, rozpoczynający się od punktu 23,25 milisekundy. Histogram ma wyraźną modę: jest nią przedział wartości $[26.25, 27.75)$. Oznacza to, że dla największej liczby osobników ich czasy latencji L3-N33 były zawarte między 26.25 a 27.75 milisekundy. W odróżnieniu od histogramu z poprzedniego przykładu nie jest on w przybliżeniu symetryczny: wartości histogramu po prawej stronie mody maleją znacznie wolniej niż po jej lewej stronie. Czasami mówimy w tej sytuacji, że prawy ogon histogramu jest znacznie dłuższy i maleje wolniej niż jego lewy ogon. Taki histogram, a zarazem rozkład cechy w próbie, dla której jest on skonstruowany jest nazywany prawostronnie skośnym (dodatnio skośnym lub prawostronnie asymetrycznym). Gdy sytuacja po obu stronach mody jest odwrotna mówimy o (ujemnej) lewostronnej skośności lub lewostronnej asymetrii.

Wykresy dla danych jakościowych

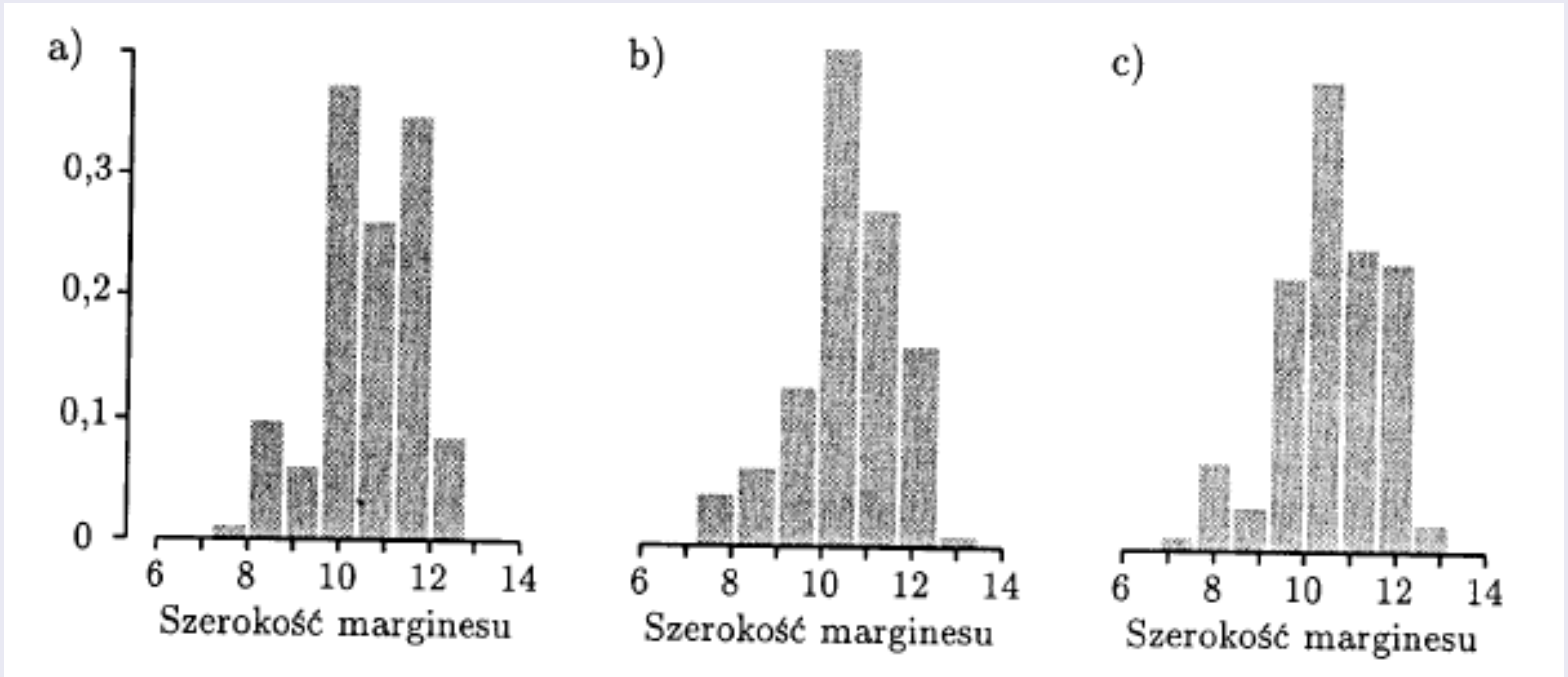
Histogramy

Przykład 3

Dane dotyczą szerokości (w milimetrach) dolnego marginesu 100 fałszywych banknotów dwudziestofrankowych franka szwajcarskiego. Przy przyjęciu początku pierwszego przedziału jako 7.2 mm i jego długości $h = 0.8$ mm otrzymamy histogram, mający 3 mody (przedział drugi, czwarty i szósty na Rys. 15a). Gdy zachowamy początek pierwszego przedziału i zmienimy długość na $h = 0.9$ mm histogram 'straci' pierwszą i trzecią modę (Rys. 15b). Z kolei zmiana początku histogramu na 6.8 mm przy zachowaniu pierwszej długości przedziału $h = 0.8$ mm prowadzi również do zmniejszenia liczby mód, ale tym razem tylko o jedną (rys. 15c).

Wykresy dla danych jakościowych

Histogramy



Rys. 14: Histogramy dla danych z Przykładu 3

Wykresy dla danych jakościowych

Histogramy

Wybór początku histogramu i długości przedziału mogą mieć duży wpływ na jego kształt. Zauważmy, że często dysponujemy dodatkową informacją pomagającą wybrać właściwy kształt spośród wielu zbudowanych dla różnych początków i długości przedziału. Na przykład trzy mody na rys. 15a mogą odpowiadać trzem różnym miejscom fałszowania banknotów. Jeśli wiemy, ze banknoty pochodziły faktycznie od 'producentów' z trzech źródeł, to jest to istotny argument przemawiający za wyborem histogramu trójmodalnego. Ogólnie zauważmy, że histogram o kilku modach może wskazywać na to, że obserwacje pochodzą z kilku istotnie różnych populacji.

Wykresy dla danych jakościowych

Histogramy – modelowanie

Wybór długości przedziału

Zakładamy, ze histogram ma rozkład zbliżony do normalnego. Możemy skorzystać ze wzoru

$$h_0 = 2.64 \times IQR \times n^{-1/3} \tag{1}$$

IQR – rozstęp międzykwartylowy, n – liczebność próby (nie ma jednej metody)

Wybór początku histogramu

Godny polecenia wydaje się wybór początku tak, aby najmniejsza wartość była środkiem pierwszego przedziału histogramu. Skuteczną metodą uniezależnienia się od wpływu początku histogramu na otrzymany kształt jest uśrednienie pewnej liczby histogramów, których początki są nieznacznie przesunięte względem siebie (metoda ASH; D. Scott (1992): *Multivariate density estimation*. Wiley, New York).

Wykresy dla danych jakościowych

Wykresy przebiegu

Jeśli dane ilościowe są zbierane w następujących po sobie momentach czasowych, dobrym pomysłem na ich wizualizację jest sporządzenie ich wykresu w funkcji czasu. Dane tego typu noszą nazwę **szeregu czasowego**, a odpowiedni wykres nazywamy **wykresem przebiegu**. Na jego podstawie można się przekonać, czy wartości zebrane w różnych odcinkach czasowych zachowują się podobnie i czy istnieje zależność między wartościami obserwowanymi w sąsiednich momentach czasowych. Tego typu informacji nie można uzyskać po przeanalizowaniu histogramu, który rejestruje tylko zagregowane w przedziały wartości cechy, pomijając momenty czasowe, w których się one pojawiły.

Wykresy dla danych jakościowych

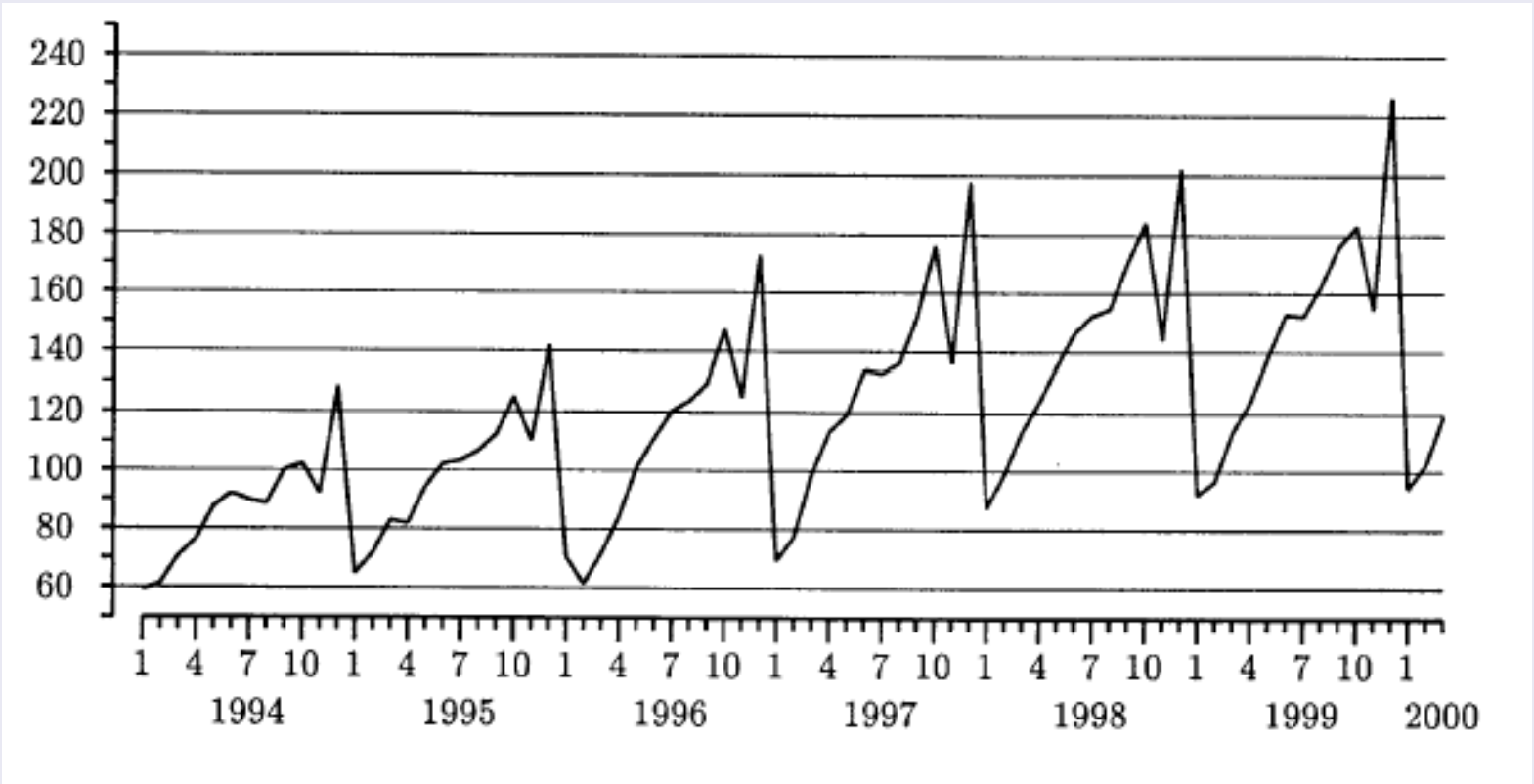
Wykresy przebiegu

Przykład 4

Rozpatrzmy wykres przebiegu produkcji sprzedanej budownictwa od stycznia 1994 do stycznia 2000 roku (na podstawie danych GUS-u). Wartości są rejestrowane co miesiąc przy przyjęciu średniej produkcji miesięcznej w 1995 roku jako 100. Obserwacje dla kolejnych momentów czasowych połączono odcinkami i otrzymano wykres w postaci linii łamanej. Dwie cechy wykresu są łatwo zauważalne: powolna, ale wyraźna ogólna tendencja wzrostu oraz powtarzający się cyklicznie kształt wykresu w poszczególnych latach. Produkcja sprzedana jest najniższa w styczniu i lutym każdego roku, później rośnie do października, po czym następuje późnojesienny zwrot powodujący spadek w listopadzie, a następnie pojawia się zwrot w przeciwnym kierunku, którego rezultatem jest największa (w skali roku!) produkcja sprzedana w grudniu (na co niepośledni wpływ miała tak zwana ulga podatkowa na budowę oraz remont i modernizację mieszkań).

Wykresy dla danych jakościowych

Wykresy przebiegu



Rys. 15: Wykres przebiegu z Przykładu 4

Wartości średnie

Wartość średnia w próbie

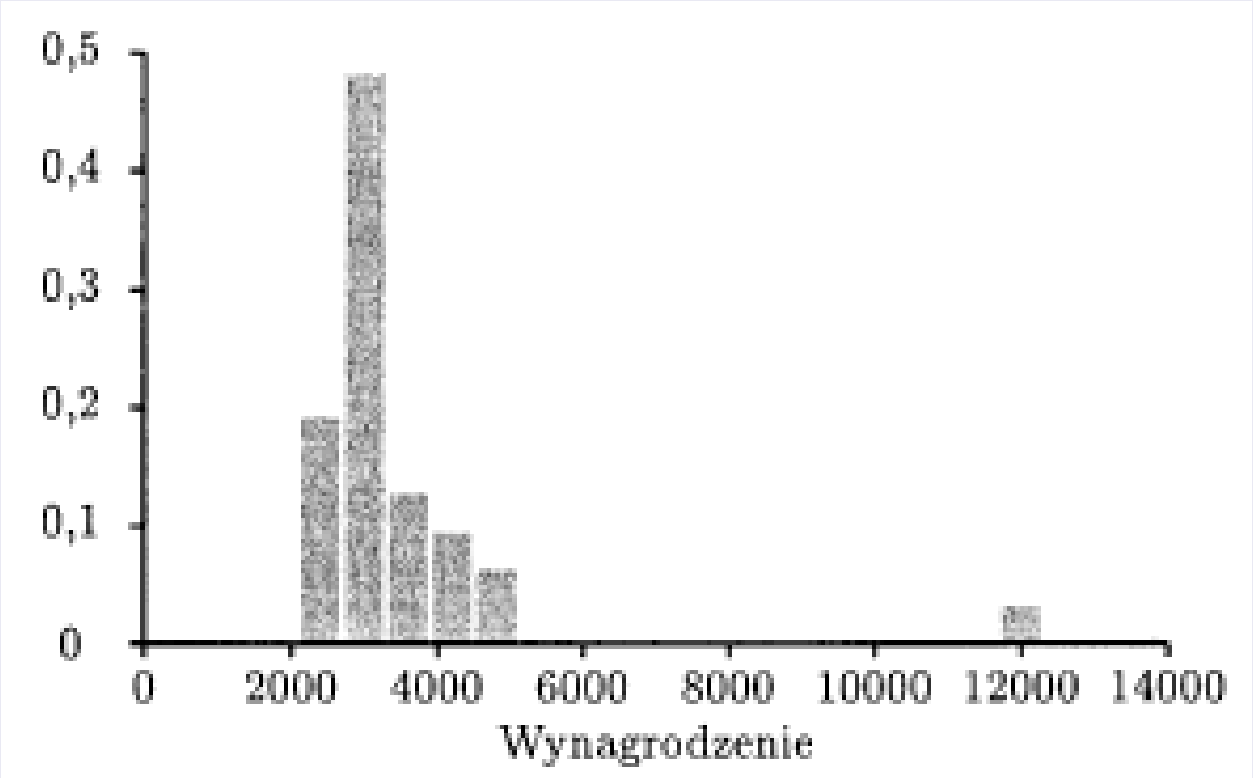
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

Mediana w próbie

$$x_{med} = \begin{cases} x_{((n+1/2))} & \text{gdy } n \text{ nieparzyste} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{gdy } n \text{ parzyste} \end{cases} \tag{3}$$

Wartości średnie

Mediana w próbie



Rys. 16: Histogram częstości dla wynagrodzeń ze średnią 3506 zł.

Wartości średnie

Mediana w próbie

Dla rys. 16 mediana wynosi $x_{((31+1)/2)} = x_{(16)} = 3100$ zł i znacznie lepiej oddaje zarobkowe perspektywy nowo zatrudnianego kandydata niż wartość średnia $\bar{x} = 3506$ zł.

Istotną cechą mediany jest jej brak wrażliwości na wartości odstające, czyli wartości bardzo wyraźnie oddalone od innych wartości w próbie i w tym sensie zdecydowanie nietypowe dla zaobserwowanego rozkładu pozostałych wartości cechy w próbie. Przez brak wrażliwości, zwany dalej odpornością na obserwacje odstające rozumiemy to, że obserwacje takie wcale lub tylko nieznacznie wpływają na wartość danego wskaźnika (w tym przypadku mediany).

Wartości średnie

Średnia ucinana

$$\bar{x}_{tk} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)} \tag{4}$$

Jeżeli rozkład cechy w próbie jest w przybliżeniu symetryczny oraz gdy nie występują w niej obserwacje odstające, średnia $\bar{x}$ i średnia ucinana powinny mieć bliskie wartości.
Ucinanie wartości skrajnych ma na celu pozbycie się wpływu ewentualnych wartości odstających na wartość wskaźnika położenia.
W przypadku średniej ucinanej musimy zdecydować jaką wartość k zastosować. Wartość ta powinna być nie mniejsza niż liczba wartości odstających na każdym z dwóch krańców rozkładu próby.
t – trimmed = ucinanie

Wartości średnie

Średnia winsorowska

$$\bar{x}_{wk} = \frac{1}{n}\left[(k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}\right] \qquad (5)$$

Średnia winsorowska wykorzystuje $n-2k$ 'środkowych' elementów próby, otrzymanych w wyniku pominięcia k najmniejszych i k największych jej elementów. Aby uwzględnić fakt pojawienia się w próbie oryginalnej k wartości nie większych niż $x_{(k+1)}$ oraz k wartości nie mniejszych niż $x_{(n-k)}$, przy obliczaniu średniej postępuje się tak, jakby $x_{(k+1)}$ i $x_{(n-k)}$ wystąpiły dodatkowo k razy (te dodatkowe wystąpienia wymienionych dwóch statystyk niejako zastępują wartości $x_{(1)}, \ldots, x_{(k)}$ oraz $x_{(n-k+1)}, \ldots, x_{(n)}$.

Ten wskaźnik zaproponował C.P. Winsor

Wskaźniki rozproszenia

Rozstęp w próbie

$$R = x_{(n)} - x_{(1)} \tag{6}$$

$x_{(n)}$ – największy element w próbie; $x_{(1)}$ – najmniejszy element w próbie;

Wskaźniki rozproszenia

Wariancja w próbie

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{7}$$

$\bar{x}$ – średnia w próbie.

Odchylenie standardowe

$$s = \sqrt{s^2} \tag{8}$$

Wskaźniki rozproszenia

Odchylenie przeciętne

$$d_1 = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \tag{9}$$

Wskaźniki rozproszenia

Kwartyle

Analizując wskaźniki położenia, zauważyliśmy, że mediana może być uważana za lepszy wskaźnik niż średnia w próbie, gdy rozkład cechy w próbie jest asymetryczny. W przypadku takiego rozkładu na wartość podanych wskaźników rozproszenia (zwłaszcza wariancji) zbyt duży wpływ mogą mieć wartości skrajne, pochodzące z długiego ogona rozkładu. Wartości takich nie jest zwykle zbyt wiele w próbie, ale są to wartości bardzo odległe od średniej i stąd mające istotny wpływ na wariancję. Dlatego, gdy mamy do czynienia z rozkładami asymetrycznymi, rozproszenie cechy w próbie warto określać na podstawie elementów położonych w centralnej części tej próby, nie uwzględniając zachowania się cechy w ogonach jej rozkładu. Wskaźnikiem opartym na pomiarze rozproszenia centralnej części próby jest **rozstęp międzykwartylowy**.

Wskaźniki rozproszenia

Kwartyle – definicja

(Pierwszym) **dolnym kwartylem** próby nazywamy medianę podpróby, składającej się ze wszystkich elementów próby o wartościach mniejszych od mediany całej próby.

(Trzecim) **górnym kwartylem** próby nazywamy medianę podpróby, składającej się ze wszystkich elementów próby o wartościach większych od mediany całej próby.

Medianę całej próby nazywamy również **drugim kwartylem** całej próby.

Oznaczenia: Q_1 – dolny kwartyl; Q_3 – górny kwartyl; Q_2 – mediana.

Wskaźniki rozproszenia

Rozstęp międzykwartylowy

$$IQR = Q_3 - Q_1 \tag{10}$$

Rozstęp międzykwartylowy jest rozstępem odniesionym do centralnej połowy wartości cechy w próbie.

IQR – interquartile range.

Wskaźniki rozproszenia

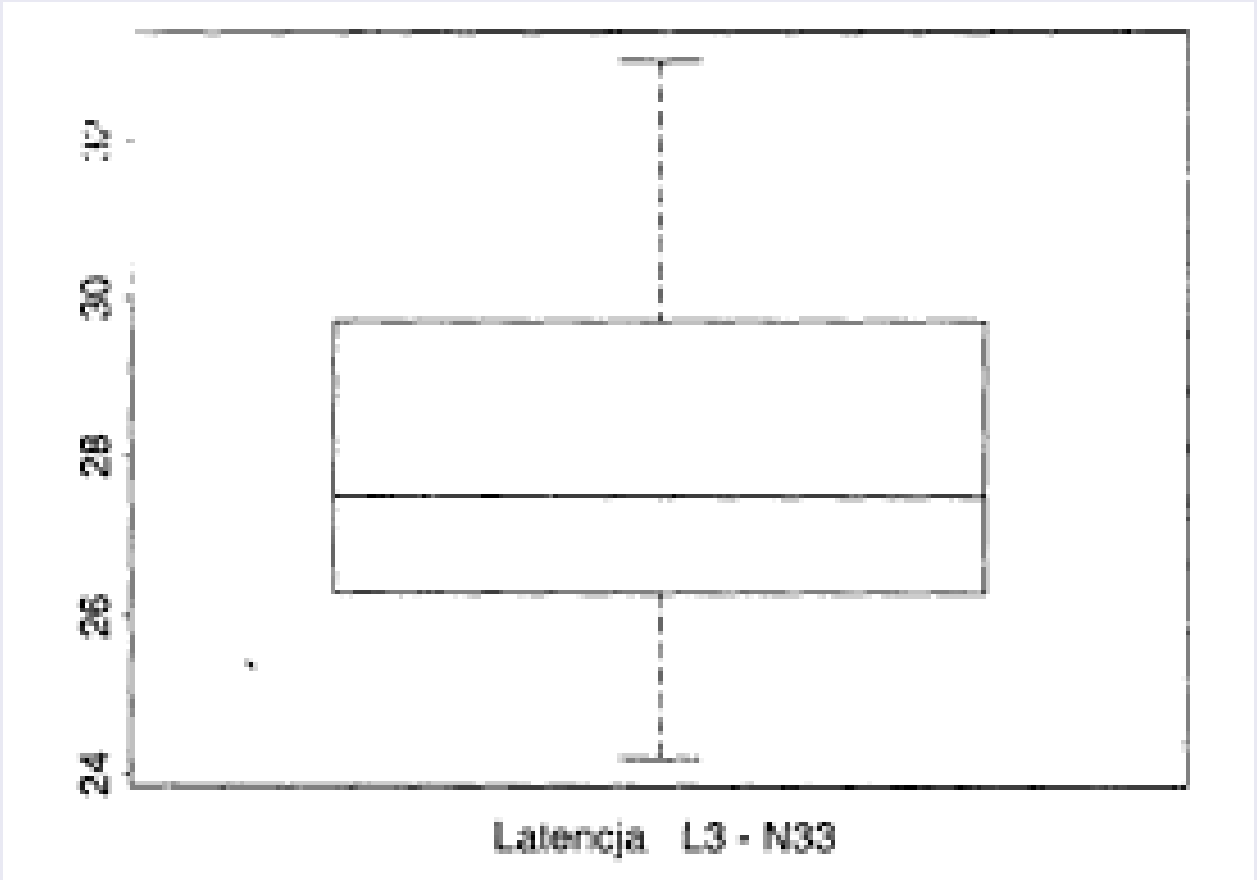
Wykres ramkowy (pudełkowy)

Wykres ramkowy dla danych z przykładu 2 jest pokazany na rys. 17. Skala na osi pionowej odpowiada wartościom obserwacji. Na wykresie, współrzędna y dolnej podstawy ramki jest równa pierwszemu kwartylowi Q_1. Współrzędna y górnej podstawy ramki jest równa trzeciemu kwartylowi Q_3. Długość boku (wysokość) ramki jest zatem równa rozstępowi międzykwartylowemu IQR. Poziomy odcinek wewnątrz ramki, niekiedy zastępowany przez mały kwadracik, wyznacza medianę cechy w próbie. Odcinek wychodzący z górnej podstawy ramki kończy się poziomą linią, wyznaczając4 największą obserwację (w próbie), spełniającą dodatkowy warunek, iż jest nie większa niż

$$Q_3 + 1.5 \times IQR \tag{11}$$

Wskaźniki rozproszenia

Wykres ramkowy (pudełkowy)



Rys. 17: Wykres ramkowy dla Przykładu 2

Wskaźniki rozproszenia

Wykres ramkowy (pudełkowy)

Podobnie do górnego wąsa tworzy się dolny wąs, sięgający od dolnej podstawy ramki do najmniejszej zaobserwowanej wartości, spełniającej dodatkowy warunek, iż jest nie mniejsza niż

$$Q_1 - 1.5 \times IQR \tag{12}$$

Wąs nie może być dłuższy niż półtora rozstępu międzykwartylowego, obserwacje zaś o wartościach mniejszych niż $Q_l - 1.5 \times IQR$ (o ile występują w próbie) są nanoszone na wykres indywidualnie.

Literatura

[1] J. Koronacki, J. Mielniczuk, Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT, 2001

Koniec?

Koniec wykładu 3