Analiza Danych

Jerzy Stefanowski

Instytut Informatyki
Politechniki Poznańskiej
Tel. 8782376
Jerzy.Stefanowski@cs.put.poznan.pl
http://www-idss.cs.put.poznan.pl/~stefan

Wykład dla kierunku „Informatyka”

Poznań, 2004

Wykład nr 1

Wprowadzenie do Analizy Danych
oraz
Statystyka Opisowa część 1

1. Uwagi wstępne

2. Cel przedmiotu

3. Statystyka i pojęcia z nią związane

4. Statystyka opisowa i wnioskowanie statystyczne

5. Pomiar cech i skale pomiarowe

6. Zbieranie danych

7. Opracowywanie materiału statystycznego (szeregi …)

8. Graficzne przedstawianie danych

9. Podsumowanie

Wprowadzenie

Po co analiza danych?

- Rozwój technologii informatycznych.
- Posiadanie dużych ilości danych nie jest równoznaczne z ich użytecznością.
- Potrzeba „inteligentnego” przetwarzania zebranych informacji.

Konieczna jest ocena jakości danych, ich analiza oraz uproszczenie opisu tak, aby można ułatwić interpretacje zjawisk, wyciągnąć użyteczne wnioski i podejmować lepsze decyzje.

Stosowane metody:

1. ***Metody statystyczne*** (tradycyjne rozumienie terminu analiza danych),
2. Metody eksploracji baz danych w celu poszukiwania użytecznej "wiedzy" (techniki „***data mining***” wywodzące się ze sztucznej inteligencji, systemy maszynowego uczenia się,...)

Dostępność wielu pakietów statystycznych (SAS, Statistica, SPSS, …).

Uwagi dydaktyczne

Statystyka matematyczna!

Czy to jest trudne?

Czy to jest nudne?

Do kogo kierowany jest ten wykład?

Cel:

Przedstawienie **przystępnego** wprowadzenia do metod statystycznej analizy danych i zdobycie umiejętności **PRAKTYCZNEGO** wykorzystania tych metod.

Metoda?

...

Świadomość poprawnego wykorzystania metod (jakie informacje są niezbędnie aby daną metodę użyć; jakie obliczenia należy wykonać) oraz interpretacji wyników.

Czy to jest trudne?

Tutaj będzie rysunek potwora statystycznego ☺)

Stosowany język, notacja, wzory, ….

Czy to jest trudne?

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \Big/ \sqrt{n}}$$

Co by było gdybyś zobaczył inny zapis:

„Pogłębiaj umiejętność docierania do najgłębszej prawdy we wszystkim co cię otacza”.

Cytat z „Księgi Pięciu Kręgów” (1645) Miyamoto Musashiego.

Literatura

- **Statystyka praktyczna, Starzyńska Wacława, PWN, 2000.**
- **Statystyka dla studentów kierunków technicznych i przyrodniczych, Koronacki Jacek, Mielniczuk Jan, WNT, 2001.**

-
- Przystępny kurs statystyki, Stanisz A., 1997.
- Po prostu statystyka, Clegg F., 1994.
- Metody statystyczne w badaniach sondażowych rynku, Kowal J., 1997.

Plan wykładów

1. Statystyka opisowa (przedstawianie danych, miary tendencji centralnej i rozproszonej).

2. Od rachunku prawdopodobieństwa do wnioskowania statystycznego (zmienne losowe i rozkłady; estymacja parametrów).

3. Weryfikacja hipotez statystycznych (testy parametryczne dla jednej zbiorowości, dla dwóch prób; testy nieparametryczne).

4. Analiza zależności między zmiennymi (korelacja, regresja, …).

5. Metodyka badań statystycznych.

? Extra niespodzianki ?

Statystyka i pojęcia z nią związane

Co to jest statystyka?

Zjawiska masowe – procesy powtarzające się dużą ilość razy.

Przykłady:

- Procesy gospodarcze (produkcja, konsumpcja, marketing,…),
- Zjawiska demograficzne (urodzenia, starzenie się ludności,…),
- Społeczne (preferencje polityczne,…).

Statystyka – dyscyplina nauki zajmująca się metodami zbierania, opracowywania i analizy danych o zjawiskach masowych.

Funkcje metod statystycznych

Podstawowe funkcje metod statystycznych:

- opisowe (deskryptywne),
- diagnostyczne (wyjaśniające),
- prognostyczne (predykcyjne).

Przykładowo funkcje te odpowiadają uzyskaniu odpowiedzi na trzy pytania, które może postawić sobie zarządzający firmą:

1. czym charakteryzuje się interesujące nas zjawisko (np. popyt na określony towar),
2. co je spowodowało,
3. jak wspomniane zjawisko będzie wyglądać w przyszłości ?

Znalezienie odpowiedzi wymaga:

1. rzetelnych pomiarów,
2. właściwych wniosków.

Statystyka opisowa a wnioskowanie statystyczne

Statystyka opisowa a wnioskowanie statystyczne

Statystyka opisowa – prezentacja danych w sposób uporządkowany, prosty z wykorzystaniem, np., miar tendencji centralnej i miar rozproszenia. Pomagają one w redukcji dużej liczby danych do zbioru bardziej zwięzłych i "pojemnych" miar.

Wnioskowanie statystyczne – pozwala ustalać prawidłowości i podejmować decyzje na podstawie zredukowanej liczby danych (próby) przy zastosowaniu rachunku prawdopodobieństwa.

Rachunek prawdopodobieństwa – możliwe jest określenie jak błąd popełnia się uogólniając wyniki z **próby** na całą **zbiorowość**.

WS pełni funkcje wyjaśniające i prognostyczne (predykcja i przewidywanie).

Statystyka opisowa – przykłady

Przykład 1.

Dysponujemy informacjami o realizacji należności z tytułu sprzedaży produktów pewnego przedsiębiorstwa z różnymi terminami płatności:

Analiza pojedynczych zapisów transakcji?

Dane pogrupowane w tabeli (szereg rozdzielczy):

Terminy płatności - sprzedaż	% sprzedaży w latach			
	1994	1995	1996	1997
za gotówkę	70	69	67	71
z terminem płatności 30-dni	20	15	10	5
z terminem płatności 60-dni	7	10	11	13
z terminem płatności 90-dni	3	5	7	6
z terminem płatności 120-dni	-	1	5	5

Przykład 2.

Rejestracja ruchu samochodowego na autostradzie w oparciu o zapis z punktów pobierania opłat.

Różne formy prezentacji danych:

Tablice dzienne, tygodniowe, …

Lecz także inaczej?

„Typowa”, przeciętna liczba samochodów w danym okresie.

„Na autostradę wjeżdża ***średnio około*** 100 samochodów na godzinę”

średnia – miara tendencji centralnej

około – rozrzut / rozproszenie rozkładu wartości wokół wartości średnich.

Wnioskowanie statystyczne – przykłady

Badania marketingowe nowego produktu wśród konsumentów:

- wybór grupy osób; ocena i porównanie z dotychczasowymi produktami,
- obserwowane różnice oceny produktu,
- czy można postawić ogólniejsze wnioski wobec zbiorowości wszystkich klientów.

Inne przykłady:

Badania preferencji wyborczych na podstawie reprezentatywnej próby losowej.

Badanie skuteczności nowego leku.

Zbiorowość, populacja i próba

Zbiorowość statystyczna - zbiór elementów (osób, obserwacji, przedmiotów,...) podobnych do siebie pod względem określonych cech (ale nie identycznych) i objętych badaniem statystycznym.

Jednorodność badanej grupy - składa się z jednostek, które nie różnią się od siebie z punktu widzenia celu badania.

Populacja - zbiór elementów obejmujący wszystkie jednostki będące przedmiotem badań.

Badanie pełne vs. częściowe.

Próba - podzbiór populacji, obejmujący część jej elementów wybranych w określony sposób (losowy lub celowy).

Reprezentatywność – badanie, które przeprowadza się na części danych, może być również odniesione wszystkich elementów, które nie są badane.

Cechy statystyczne - obiekty (jednostki statystyczne, elementy zbiorowości) podlegające badaniu posiadają własności:

- wspólne – stałe dla wszystkich obiektów,
- **zmienne** – takie, które różnicują obiekty między sobą.

Pomiar cech i skale pomiarowe

Model – pominięcie części cech rzeczywistych badanego zdarzenia oraz „akcentowanie" tych aspektów, które są szczególnie użyteczne dla celu badania.

Pomiar wybranych zmiennych niezbędnym aspektem definiowania modelu.

Pomiar - przyporządkowanie liczb lub odpowiednich symboli obiektom zgodnie z określonymi regułami w taki sposób, aby odzwierciedlały one relacje zachodzące między tymi obiektami.

Mierzenie zmiennych

Rodzaje zmiennych:

1. Liczba obiektów lub zdarzeń
2. Natężenie lub intensywność występowania pewnej właściwości, którą wykazuje obiekt lub zdarzenie.
3. Częstość (lub częstotliwość) występowania właściwości lub zdarzeń.

Obiekty podlegające badaniu mają cechy/właściwości: jakościowe (nominalne), porządkujące i ilościowe.

Typy skal pomiarowych

Skale metryczne i niemetryczne

Wyróżnia się cztery podstawowe skale pomiarowe:

1. nominalna,
2. porządkowa,
3. przedziałowa,
4. ilorazowa.

Przykład różnych skal pomiarowych

Dostępne są dane o pracownikach pewnej uczelni.

Pracownik	**Specjalność**	**Płeć**	**Wiek**	**Znajomość j. ang.**	**Stanowisko**
A	bazy danych	M	45	b.dobra	profesor
B	zarządzanie	K	39	dobra	adiunkt
C	marketing	M	41	dobre	profesor
D	matematyka	M	47	słaba	wykładowca
E	sieci komp.	K	29	b.dobra	asystent
F	j.program.	M	36	dobra	adiunkt
G	ekonomia	M	34	słaba	adiunkt
…	…	…	…	…	…

Skala nominalna - zbiór dwóch lub więcej kategorii jakościowych, które umożliwiają zupełną i rozłączną klasyfikację wyników.

Relacje między kategoriami- równość i różność.

Liczby przypisywane kategoriom - nie mają innego znaczenia niż odróżnienie jednej kategorii od drugiej.

Nie są możliwe działania arytmetyczne na danych w tej skali.

Skalowanie binarne - cecha występuje 1,
w przeciwnym przypadku 0.

Skalowanie trychotomiczne (tak, nie, nie wiem).

Przykłady pomiarów wyrażonych w skali nominalnej:

- płeć,
- status małżeński,
- marka handlowa,
- gatunek towaru,
- typ posiadanej własności,
- branża przemysłowa,
- rodzaj lub profil firmy,
- rodzaj usług,
- typ potencjalnego klienta.

Skala porządkowa - wyznaczona przez relację porządkującą niektóre lub wszystkie elementy zbioru wyników.

Skala ta pociąga za sobą porządkowanie lub szeregowanie wartości zmiennej, np. uporządkowanie różnych marek handlowych: A, B, C, D.

1 - najbardziej preferowana marka D, 2 - C i 3 -B marki preferowane w dalszej kolejności, 4 - marka A najmniej preferowana.

--D----C--------------B--------A-

Pomiar porządkowy nie daje informacji o wielkości kolejnych różnic między elementami; nie można porównywać odległości ani dokonywać operacji arytmetycznych.

Przykłady:

- niektóre skale szacunkowe lub pozycyjne,
- status zawodowy,
- upodobanie do marki handlowej, gatunku towarów, itp.
- siły reakcji konsumentów na różne bodźce.

Przykładowy kwestionariusz:

Skale metryczne

Skala przedziałowa (interwałowa, równomierna) – spełnia własności uporządkowania a ponadto zakłada, że porządkowany zbiór wartości cech składa się z liczb rzeczywistych. Skalę określa się wskazując stała jednostkę miary i relację przyporządkowującą każdemu wynikowi obserwacji liczbę (zero przyjęte „arbitralnie").

Działania arytmetyczne, takie jak dodawanie i odejmowania są możliwe.

Przykłady:

- temperatura w skali Celsjusza,
- wynik finansowy,
- czas kalendarzowy.

Skala ilorazowa – posiada własności skali przedziałowej; ponadto pomiary charakteryzują się stałymi stosunkami oraz istnieniem bezwzględnego zera (niearbitralnego).

Pomiary określane są na osi R_+ wraz z zerem.

Każdemu wynikowi obserwacji można przyporządkować liczbę o stałym ilorazie.

Można wykonywać operacje arytmetyczne (w tym mnożenie i dzielenie).

Przykłady:

- wiek,
- dochody,
- cena towaru,
- wielkość sprzedaży,
- upływający czas.

Jeszcze o zmiennych

Cechy **mierzalne** a cechy **niemierzalne**.

Zmienność skokowa:

Cecha przyjmuje skończoną liczbę wartości liczbowych.

Przykład:

Liczba sztuk produktu wyprodukowania w ciągu dnia może wynieść: 10, 11, 12, 13, 14, 15, 16 lub 17.

Zmienność ciągła:

Cecha przyjmuje dowolne wartości liczbowe z pewnego przedziału liczbowego.

Przykład:

Masa pewnego produktu wyrażona w kilogramach i częściach kilograma.

Zbieranie danych

Podstawowe źródła danych:

źródło pierwotne → dane pozyskane w trakcie specjalnie przeprowadzonego badania (np. badania ankietowe, eksperyment naukowy, spis powszechny, …),

źródło wtórne → materiał zaczerpnięty z innych zasobów danych (nie przeprowadzonych przez statystyka badań).

Efekt pozyskania danych → tzw. „surowy" materiał statystyczny.

Potrzeba wstępnego przetwarzania:

- kontrola materiału statystycznego,
- porządkowanie i grupowanie obserwacji,
- budowa szeregów (prezentacja tabelaryczna),
- sporządzanie wykresów (prezentacja graficzna).

Opracowanie materiału statystycznego

Po zebraniu materiału statystycznego – „*surowe*” *dane*.

Zebrany materiał → usystematyzowanie, przetworzenie i zestawienie w postaci odpowiednich tablic, tj. *szeregów statystycznych*.

Szereg statystyczny – ciąg wartości pomiarowych wzrastający lub malejący czy pogrupowany wg. określonych kryteriów statystycznych. Reprezentowany w postaci tablic jedno lub wielowymiarowych.

Stosuje się także macierze uporządkowania (skale porządkowe), tablice korelacyjne, czy wykresy (histogramy).

Klasyfikacja podstawowych szeregów statystycznych.

Prezentacja tabelaryczna materiału statystycznego

Zmienna skokowa

Przykład:

W stu kolejnych rzutach kostką otrzymano następujące wyniki:

Jak opisać powyższą próbę wyników?

Podajmy **rozkład wartości** jakie cecha przyjmuje w próbie.

Rozkład liczby oczek w próbie:

Wartość (liczba oczek)	1	2	3	4	5	6
Liczność (liczba wystąpień)	16	19	9	17	25	14
Częstość	0,16	0,19	0,09	0,17	0,25	0,14

Prezentacja tabelaryczna materiału statystycznego

Przykład 4

Rejestrujemy wiek 20 pracowników zgłaszających się na okresowe badania w pewnym zakładzie pracy.

Zaobserwowane wielkości (w latach):

36, 41, 33, 34, 38, 26, 33, 36, 30, 48, 39, 31, 35, 36, 38, 37, 22, 31, 25, 32

Szereg uporządkowany:

22, 25, 26, 30, 31, 31, 32, 33, 33, 34, 35, 36, 36, 36, 37, 38, 38, 39, 41, 48

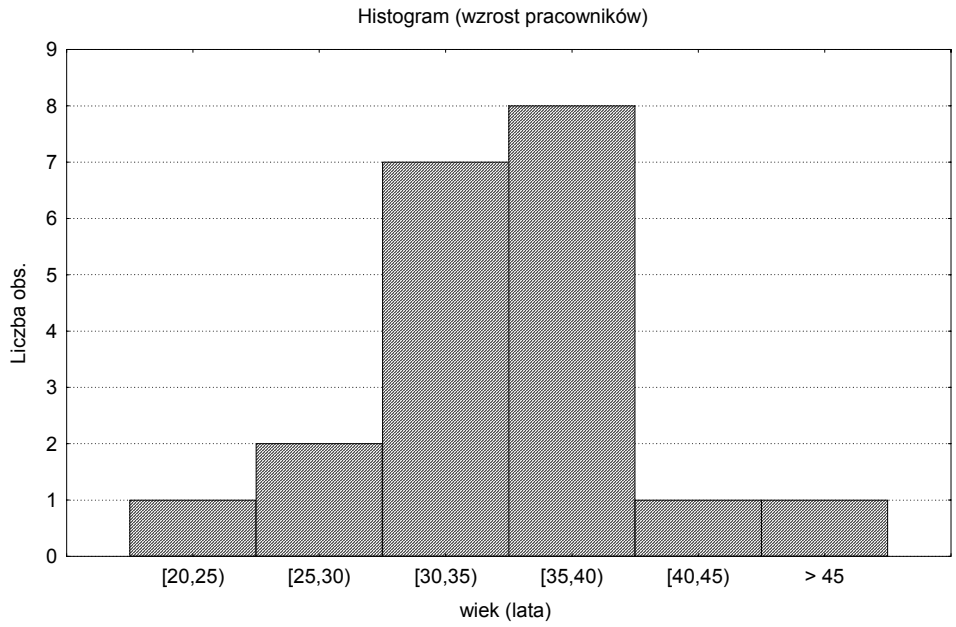
Liczba różnych wartości w próbie – 16.

Agregacja danych → grupowanie obserwacji z wykorzystaniem przedziałów klasowych.

Przykład 5 cd.

Rozpatrzmy następujące przedziały wiekowe:
[20,25), [25,30), [30,35), [35,40), [40,45), [45,50).

Przedział	Klasa	Liczność klasy	Częstość
[20,25)	22	1	1/20=0,05
[25,30)	25,26	2	2/20=0,1
[30,35)	30,31,31,32,33,33,34	7	7/20=0,35
[35,40)	35,36,36,36,37,38,38,39	8	8/20=0,4
[40,45)	41	1	1/20=0,05
[45,50)	48	1	1/20=0,05



Szeregi statystyczne

Grupowanie obserwacji – **przedziały klasowe**.

Jak dobierać przedziały klasowe?

Przedziały o równej szerokości:

Rozpiętość przedziału:

$$h = \frac{x_{\max} - x_{\min}}{k}$$

gdzie: $x_{\max}$ i $x_{\min}$ – maksymalna i minimalna wartość cechy, k – ustalona wcześniej liczba przedziałów.

Ale, skąd wiemy….?

Różne techniki heurystyczne, np.:

– liczba obserwacji niewielka → nie więcej niż 5 przedziałów,

– wielkość przedziału (Koronacki,Malenczuk str. 24)

$h_o = 2{,}64 \cdot IQR \cdot n^{-1/3}$ (IQR rozstęp międzykwartylowy),

– ….

Pojęcia powiązane:

środek przedziału klasowego – x_i.

Rozkład liczności – zbiór wartości zmiennej (lub jej wariantów) oraz liczba przypadków przyjmujących te wartości.

Liczności bezwzględne, częstości względne, częstości procentowe,

Częstości skumulowane – suma częstości wszystkich poprzednich przedziałów aż do danego przedziału włączenie.

Przykład 6.

Badaniu poddano pracowników przedsiębiorstwa ALFA z punktu widzenia ich wieku – możliwe poprzez przejrzenie kartoteki pracowników.

Dane zestawione wg. kolejności alfabetycznej:

Uporządkowany szereg statystyczny

Prosty rozkład częstości

Przykład 6. cd.

Różne szeregi rozdzielcze:

Graficzne przedstawianie danych

„Jeden rysunek jest więcej wart niż 100 słów”.

Wykresy statystyczne – służą do bardziej skondensowanego i poglądowego, w porównaniu do szeregów przedstawienia wyników.

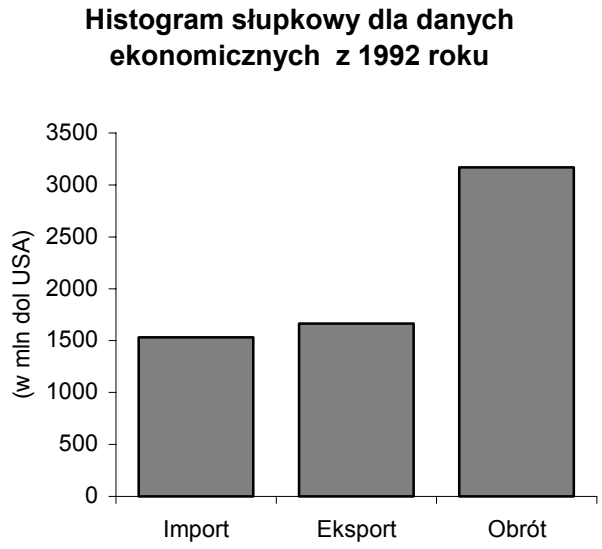
Wykresy dla danych jakościowych:

Przykład 7:

W poniższej tabeli przedstawiono dane dotyczące polskiego importu-eksportu w latach 1992-1995 wyrażone w mln dol. USA (Źródło: Rocznik Statystyczny, 1996, s. 467).

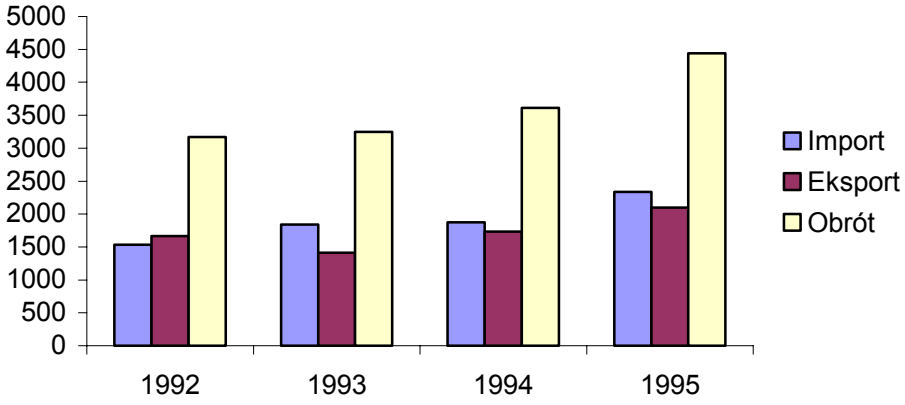
Pozycja	**1992**	**1993**	**1994**	**1995**
Import	1531,2	1837,9	1876,4	2339,2
Eksport	1665,7	1409,8	1735,5	2100,1
Obrót	3169,9	3247,7	3611,9	4439,3

Przykłady prezentacji za pomocą wykresów statystycznych:

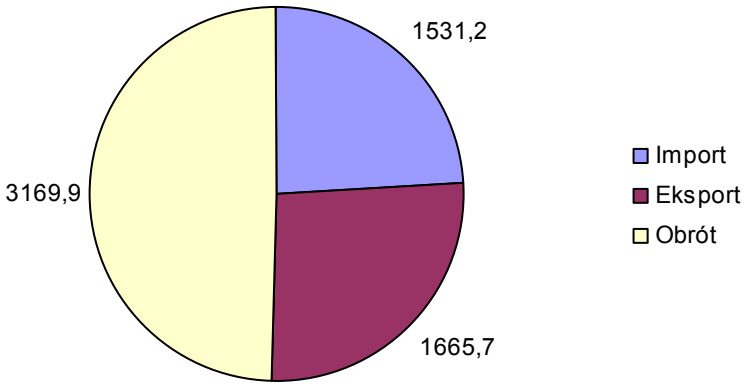


Inne formy prezentacji wykresów statystycznych

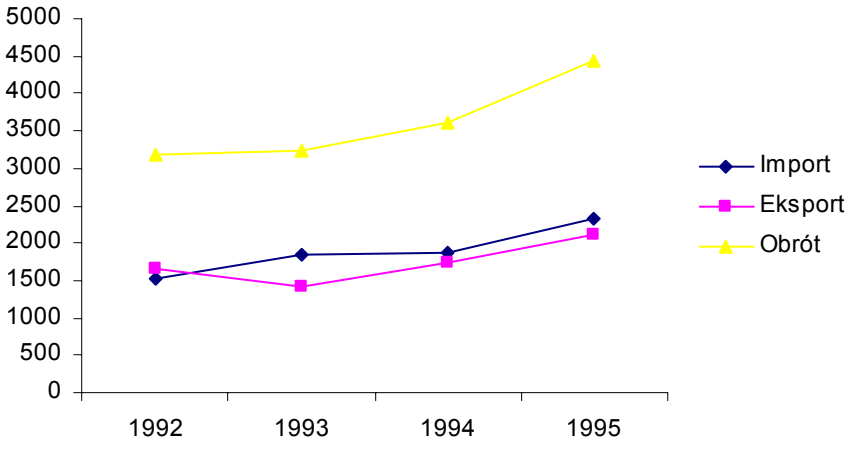
Histogram słupkowy dla danych ekonomicznych o wymianie z zagranicą



Wykres kołowy dla danych z 1992 roku



Wykres liniowy

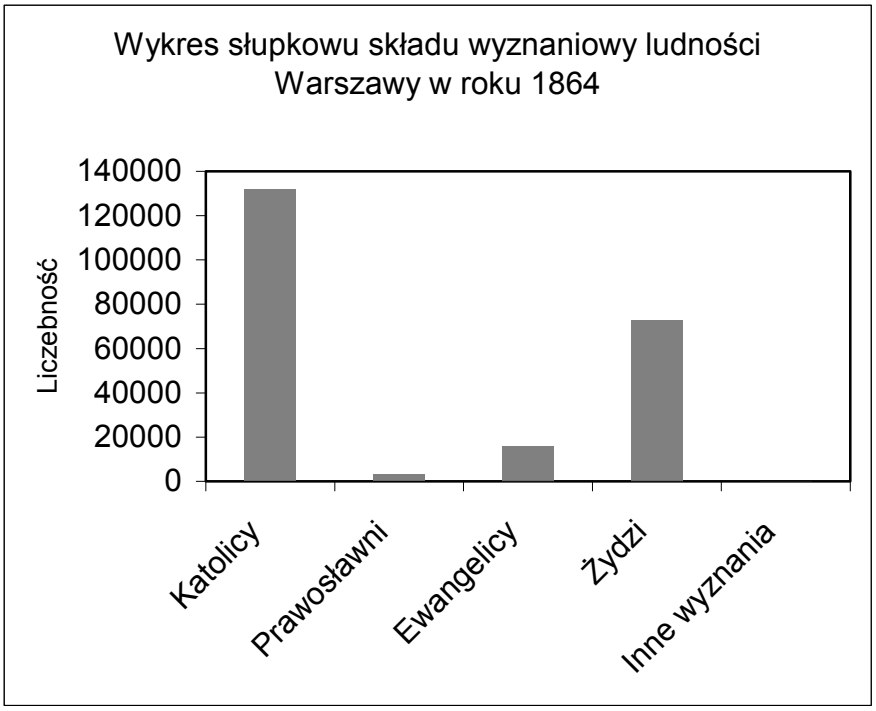


Zadanie domowe:

Za książką Koronacki, str 14 = dane dotyczące składu wyznaniowego ludności Warszawy w latach 1864 i 1917.

Kategoria wyznanowiona	Rok 1864		Rok 1917	
	Liczebność	%	Liczebność	%
Katolicy	131808	59,1	387069	46,2
Prawosławni	3026	1,4	3961	0,5
Ewangelicy	15909	6,7	12147	1,5
Żydzi	72772	32,6	329535	39,3
Inne wyznania	287	0,2	104500	12,5

Proszę przedstawić graficzne powyższe dane.



Wykresy dla danych ilościowych

Uporządkowany szereg wartości cechy:

$$x_1, x_2, \ldots, x_k \quad (x_1 < x_2 < \ldots < x_k \; ; \; k \leq n)$$

Rozkład wartości:

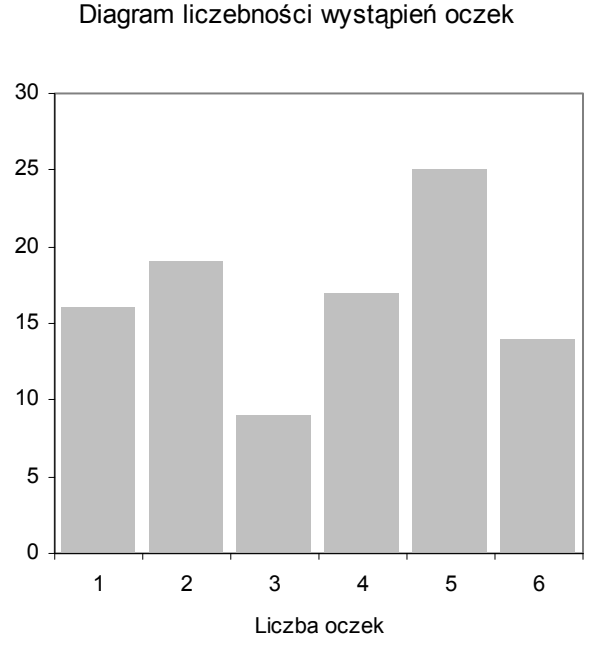
$$(x_1,n_1), (x_2,n_2) \ldots, (x_k,n_k)$$

Histogram – wykres słupkowy umieszczony w układzie współrzędnych; oś rzędnych reprezentuje częstości / liczności a oś odciętych różne wartości zmiennej, wysokość słupków jest proporcjonalna do częstości.

Przykłady:

Rozkład liczby oczek w próbie:

Wartość (liczba oczek)	1	2	3	4	5	6
Liczność (liczba wystąpień)	16	19	9	17	25	14
Częstość	0,16	0,19	0,09	0,17	0,25	0,14



Wykresy szeregów czasowych

Wizualizacja zdarzeń w następujących po sobie momentach czasowych. Wykres w funkcji czasu.

Podsumowanie

Omówiono:

1. Pojęcia związane ze statystyczną analizą danych, w tym:
    - cel i zakres badań statystycznych
    - rozróżnienie metod na statystykę opisową i wnioskowanie statystyczne,
    - zbiorowość, populacja, próba, cechy statystyczne.
2. Pomiar cech i skale pomiarowe.
3. Opracowywanie danych w postaci szeregów statystycznych, metody doboru przedziałów klasowych
4. Podstawowe wykresy do graficznego przedstawiania danych.

Więcej o wizualizacji danych:

późniejsze wykłady + literatura.

Warto **rozszerzyć wiedzę**, np.

- zapoznać się z zagadnieniami gromadzenia danych i metodyce prowadzenia badań statystycznych,
- tworzeniem prób,
- …