Przedmiot statystyki. Graficzne przedstawienie danych.

dr Mariusz Grządziel

Wykład 1; 6 lutego 2012

Przedmiot statystyki
Statystyka dzieli się na trzy części:
-zbieranie danych;
-opracowanie i kondensacja danych (analiza danych);
-wyciąganie wniosków na podstawie danych (wnioskowanie statystyczne).

Statystyka i gospodarka przestrzenna
Ceny mieszkań używanych we Wrocławiu (ceny proponowane przez sprzedających):
Ceny mieszkań w dzielnicy A (miasta Wrocławia): 65,80,139,180,355,158,240,205,265,305,200,155,209,
310,149,254,188,265,275,200,184,130,260,250,195
Ceny mieszkań w dzielnicy B (miasta Wrocławia): 420,350,275,277,327,223,198,275,350,327,260,306, 270,295,270,220,299
Średnia cena w dzielnicy A: $208{,}8$;średnia cena w dzielnicy B: $290{,}71$.
Czy ceny mieszkań w dzielnicy B są wyższe niż w dzielnicy A?

Analiza zależności cech
Powierzchnia (w metrach kwadratowych) mieszkań w dzielnicy B:
94,73,75,80,74,60,50,63,74,74,56,85,80,80,80,75,60
Dane nt. mieszkań z dzielnicy B można przedstawić jako tzw. „szereg dwucechowy”:

$$(94; 420), (73, 35), \ldots, (60; 299).$$

Wykres rozproszenia
Dane te można przedstawić przy pomocy wykresu rozproszenia:

Wykres rozproszenia+prosta MNK
Do „chmury punktów” na wykresie rozproszenia można dopasować prostą w następujący sposób. Oznaczmy dane („szereg dwucechowy”) przez $(x_1, y_1), \ldots, (x_n, y_n)$ Chcemy znaleźć prostą $y = b_0 + b_1 x$ taką,że

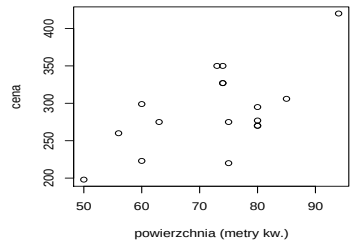
$$S(b_0, b_1) = \sum_{k=1}^{n} (y_i - b_1 x_i - b_0)^2 \text{ suma kwadratów odchyleń}$$

przyjmuje wartość minimalną (jeśli nie wszystkie wsółrzędne x-owe są równe jednej liczbie, to istnieje dokładnie jedna para liczb, dla których krytetium S przyjmuje wartość minimalną). Otrzymana prosta MNK (od **M**etody **N**ajmniejszych **K**wadratów)- odpowiada minmalnej wartości funkcji $S(b_0, b_1)$; wielkość $y_i - b_1 x_i - b_0$ można interpretować jako odchylenie $i - tej$ obserwacji y_i od wartości przewidywanej $b_1 x_i + b_0$
Dla danych nt. mieszkań w dzielnicy B prosta ta dana jest równaniem:

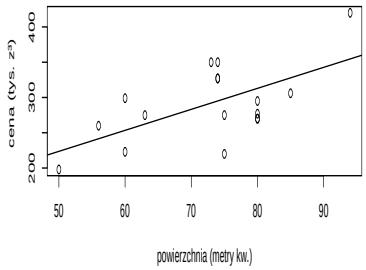
$$y = 74{,}78729 + 2{,}97698x$$

Wykres rozproszenia+prosta MNK
Problem: czy rozważana zależność liniowa między zmiennymi jest w jakimś sensie istotna statystycznie? adekwatna?

Wstępna analiza danych i wnioskowanie statystyczne
Schemat postępowania: dla posiadanego zbioru danych wykonujemy wstępną analizę: obliczamy wskaźniki sumaryczne (średnią itd.) oraz sporządzamy odpowiednie wykresy statystyczne; następnie przeprowadzamy odpowiednie analizy statystyczne (testujemy odpowiednie hipotezy itd.)



Rysunek 1: Wykres rozproszenia dla danych: powierzchnia i ceny mieszkań w dzielnicy B



Rysunek 2: Wykres rozproszenia dla danych: powierzchnia i ceny mieszkań w dzielnicy B+prosta MNK

Cechy ilościowe i jakościowe
Cena mieszkań w dzielnicy B- przykład cechy ilościowej (mamy tu do czynienia z liczbami odpowiadającymi wartościom mierzonych wielkości);
-cechy jakościowe:

- płeć;
- typ schorzenia;

Histogram i szereg rozdzielczy
Dla zbioru danych liczbowych $y_1, y_2 \ldots, y_n$ niech: $MIN1$ oznacza liczbę mniejszą od najmniejszej z liczb $y_1, y_2 \ldots, y_n$; $MAX1$ oznacza liczbę większą lub równą od największej z liczb $y_1, y_2 \ldots, y_n$; $MIN1 < MIN$ i $MAX1 \geqslant MAX$ mogą być odpowiednimi „zaokrągleniami" wartości, odpowiednio, minimalnej i maksymalnej naszego zbioru danych. (MIN i MAX oznaczają, odpowiednio, wartość minimalną i maksymalną dla $\{y_1, y_2 \ldots, y_n\}$). Podzielmy odcinek $(MIN1, MAX1]$ na k przedziałów (zwanych klasami) o równej długości:

$$(x_0, x_1], (x_1, x_2], \ldots, (x_{k-1}, x_k], \text{ gdzie } x_0 = MIN1, x_k = MAX1$$

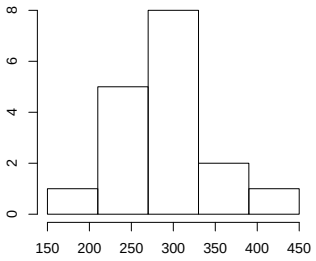
Funkcję przyporządkowującą poszczególnym przedziałom liczbę elementów naszego zbioru danych do nich należących będziemy nazywać szeregiem rozdzielczym.

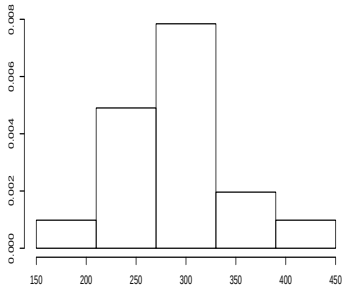
Ustalenie liczby klas w szeregu rozdzielczym
Istnieje kilka reguł ustalania liczby klas k szeregu rozdzielczego w zależności od liczby obserwacji n. Oto niektóre z nich:

$$k \approx \log_2 n + 1; \quad k \approx \sqrt{n}.$$

Szereg rozdzielczy dla danych: ceny mieszkań w B
Dla danych: „ceny mieszkań w B":
420,350,275,277,327,223,198,275,350,327, 260,306,270,295,270,220,299
znajdujemy: $MIN = 198$, $MAX = 420$. Przyjmujemy: $MIN1 = 150$; $MAX1 = 450$ oraz $k = 5$.
Otrzymujemy szereg rozdzielczy, przedstawiony w postaci tabeli:

klasa	$(150, 210]$	$(210, 270]$	$(270, 330]$	$(330, 390]$	$(390, 450]$
liczność	1	5	8	2	1





Histogram liczebności dla danych „ceny mieszkań w B”

Histogram częstości
Jeśliby histogram liczebności przeskalować w ten sposób, że wysokości słupków odpowiadałyby ilorazom liczebości klas i liczby wszystkich obserwacji n, wtedy otrzymalibyśmy histogram częstości. Wysokości słupków tego histogramu byłyby równe:

$$\frac{1}{17} \approx 0{,}06; \quad \frac{5}{17} \approx 0{,}29 \text{ itd.}$$

Histogram probabilistyczny
Jeśliby histogram przeskalować tak, aby suma pól wszystkich prostakątów („słupków”) była równa 1, otrzymamy tzw. histogram probabilistyczny (od *probability* (ang.) - prawdopodobieństwo).

$$h_i = \frac{n_i}{nh},$$

gdzie h_i oznacza wysokość i-tego słupka w histogramie probabilistycznym, n_i liczebność i-tej klasy, n liczebność próby, h szerokość klasy.
Histogram probabilistyczny: oszacowanie rozkładu jedności prawdopodobieństwa dla danej cechy.
Jeśli funkcja h-funkcja, odpowiadająca histogramowi probabilistycznemu, to „prawdopodobieństwo”, że wartość danej cechy X będzie się mieściła w $[a, b]$:

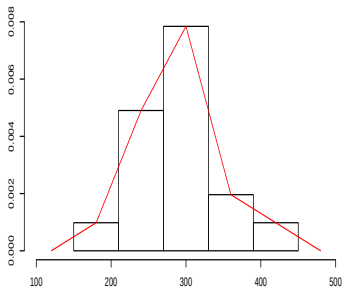
$$P(a < X < b) \approx \int_a^b h(x)dx$$

Histogram probabilistyczny dla „cen mieszkań w B”

Wielobok częstości
Oznaczmy długość klasy histogramu przez H. Jeżeli połączymy odcinakmi:

- punkt $(MIN1 - H/2, 0)$;
- środki boków słupków histogramu probabilistycznego leżących na przeciw podstaw tych słupków;
- punkt $(MAX1 + H/2, 0)$;

otrzymamy tzw. *probabilistyczny wielobok częstości* (por. Rys. 3)



Rysunek 3: Histogram probabilistyczny+probabilistyczny wielobok częstości dla danych „ceny mieszkań w B”

Literatura

[1] Bednarski, T. Elementy matematyki w naukach ekonomicznych. Oficyna ekonomiczna. Kraków 2004.

[2] Komsta, Ł, Wprowadzenie do środowiska R. Strona WWW http://r.meteo.uni.wroc.pl/doc/contrib/Komsta-Wprowadzenie.pdf

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

[4] Łomnicki, A., Wprowadznie do statystyki dla przyrodników. PWN. Warszawa 2003.

[5] Ostasiewicz, S., Rusnak, Z., Siedlecka, U. Statystyka. Elementy teorii i zadania. Wyd. Akadamii Ekonomicznej we Wrocławiu, 1999

[6] The R Project for Statistical Computing. Strona WWW http://www.r-project.org/

[7] Verzani, J. simpleR-Using R for Introductory Statistics. http://r.meteo.uni.wroc.pl/doc/contrib/Verzani-SimpleR.pdf