Wstępna analiza danych

Kiedy po raz pierwszy spotykamy się z nowym zestawem danych, naszym zadaniem jest opis podstawowych ich cech. Główne cechy danych mówią nam o zasadniczych własnościach zjawisk lub eksperymentu, który badamy. Ponadto prawie zawsze potrzebny jest nam syntetyczny opis danych: bardzo trudno jest na przykład analizować "surowe" wyniki spisu powszechnego w Polsce. Konieczne jest dokonanie odpowiedniego ich przekształcenia i uproszczenia umożliwiającego analizę. Przede wszystkim musimy jednak ustalić, jaki jest typ danych. Jeśli mamy do czynienia z liczbami odpowiadającymi wartościom mierzonych wielkości, jak na przykład w przypadku pomiaru temperatury przy gruncie o godzinie ósmej rano na Śnieżce w kolejnych dniach listopada, to mówimy wtedy o danych ilościowych. W przypadku, gdy rejestrujemy cechę jakościową obiektów, na przykład płeć lub typ schorzenia pacjentów, mówimy o danych jakościowych. Oczywiście, jeśli dla jednego obiektu dokonujemy kilku pomiarów, to część z nich może być typu ilościowego, a część jakościowego. Możemy rejestrować jednocześnie wiek pacjenta (cecha ilościowa) i to, czy ma on lub nie problemy ze snem (cecha jakościowa). Określenie typu danych jest niezbędne przed przystąpieniem do ich wstępnej analizy.

Graficzne przedstawienie danych

Wykres zawiera znacznie więcej informacji niż jeden, a nawet kilka wskaźników liczbowych obliczonych na podstawie danych. Często jest tak, że wartość pewnego wskaźnika odpowiada dwóm zupełnie rożnym wykresom i dlatego opieranie się wyłącznie na wartości tego wskaźnika może być mylące. Zarazem, wykres też jest pewną redukcją informacji w stosunku do oryginalnych danych, ale jest to redukcja bez porównania mniej drastyczna.

http://en.wikipedia.org/wiki/Chart

Histogramy

http://en.wikipedia.org/wiki/Histogram

Określanie szerokości przedziałów (binów, słupków) histogramu

$$h = 2{,}64 \cdot \frac{IQR(x)}{n^{\frac{1}{3}}}$$

Wybór początku pierwszego przedziału – najlepszą metodą jest takie dobranie pierwszego przedziału, aby najmniejsza wartość występująca w zbiorze była środkiem pierwszego przedziału.

MODA (DOMINANTA) – WARTOŚĆ O NAJWIĘKSZYM PRAWDOPODOBIEŃSTWIE WYSTĄPIENIA. JEST TO WARTOŚĆ, KTÓRA WYSTĘPUJE NAJCZĘŚCIEJ W ZBIORZE DANYCH.

MEDIANA – WARTOŚĆ ŚRODKOWA (NIE ŚREDNIA!!) W ZBIORZE DANYCH. DRUGI KWARTYL.

KWARTYLE – WARTOŚCI ŚRODKOWE DLA 4 PRZEDZIAŁÓW W ZBIORZE.

ROZSTĘP MIĘDZYKWARTYLOWY (IQR) – RÓŻNICA WARTOŚCI POMIĘDZY PIERWSZYM I TRZECIM KWARTYLEM.

ŚREDNIA ARYTMENTYCZNA

$$\bar{a} = \frac{\sum_{i=1}^{n} a_i}{n}$$

ŚREDNIA UCINANA

$$\bar{a} = \frac{\sum_{i=k}^{n} a_i}{n-k}$$

WARTOŚCI „K” ORAZ „N” USTALAMY ARBITRALNIE, ALE ZAZWYCZAJ SORTUJEMY ZBIÓR WG WARTOŚCI I OBCINAMY WARTOŚCI:

- MIN I MAX
- ODSTAJĄCE OD INNYCH (RÓWNA LICZEBNOŚĆ Z POCZĄTKU I Z KOŃCA)
- 25% PO OBU KOŃCACH

ŚREDNIA WINSOROWSKA

LICZYMY DOKŁADNIE TAK SAMO JAK ŚREDNIĄ ARYTMETYCZNĄ. RÓŻNICA POLEGA NA TYM, IŻ WŚRÓD POSORTOWANYCH ELEMENTÓW WYBIERAMY OKREŚLONĄ LICZBĘ ELEMENTÓW SKRAJNYCH (TYLE SAMO NA POCZĄTKU I KOŃCU) I ZAMIENIAMY ICH WARTOŚCI NA WARTOŚĆ MIN I MAX Z POZOSTAŁYCH ELEMENTÓW.

ZADANIA

1. W STU KOLEJNYCH RZUTACH KOSTKĄ OTRZYMANO NASTĘPUJĄCE WYNIKI:
   5 2 2 6 3 2 5 3 1 2 5 3 6 2 5 4 4 6 1 6 4 5 5 2 4 6 1 4 4 3 4 2 4 2 4 4 1 1 4 5 3 1 5 6 5 6 1 5 6 2 4 5 5 2 5 4 5 5 1 1 2 2 5 5 2 6 3 5 5 4 1 4 5 5 1 4 3 2 1 2 6 1 2 1 6 5 1 3 6 1 5 6 6 2 2 3 5 5 2 4.

   - WYGENERUJ WYKRES ILOŚCIOWY;
   - OKREŚL:
     - MODĘ
     - MEDIANĘ
     - KWARTYLE

2. REJESTRUJEMY WIEK 20 PRACOWNIKÓW ZGŁASZAJĄCYCH SIĘ NA OKRESOWE BADANIA W PEWNYM ZAKŁADZIE PRACY. ZAOBSERWOWANE WIELKOŚCI WYNOSZĄ (W LATACH):
   36, 41, 33, 34, 38, 26, 33, 36, 30, 48, 39, 31, 38, 37, 22, 31, 25, 32.

   - WYZNACZ IQR (ROZSTĘP MIĘDZYKWARTYLOWY);
   - OKREŚL PRZEDZIAŁY I POCZĄTEK HISTOGRAMU;
   - UTWÓRZ HISTOGRAM;
   - OKREŚL:
     - MODĘ
     - MEDIANĘ
     - KWARTYLE
     - ŚREDNIĄ ARYTMETYCZNĄ

3. DLA PODANEGO ZBIORU DANYCH:
   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 26,60 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

   - OKREŚL:
     - MODĘ
     - MEDIANĘ
     - KWARTYLE
     - ŚREDNIĄ ARYTMETYCZNĄ
   - WYZNACZ IQR (ROZSTĘP MIĘDZYKWARTYLOWY);
   - OKREŚL PRZEDZIAŁY I POCZĄTEK HISTOGRAMU;
   - UTWÓRZ HISTOGRAM;
   - STWÓRZ WYKRES LINIOWY.

4. DLA ZAŁĄCZONEGO PLIKU Z DANYMI:

   - WCZYTAJ PLIK .TXT DO EXCELA ROBIĄC ODPOWIEDNIĄ KONWERSJĘ;
   - SFORMATUJ KOMÓRKI TAK, ŻEBY „COŚ” WIDZIEĆ;
   - SKOPIUJ DANE DO INNEJ KOLUMNY (WIERSZA) I POSORTUJ;
   - POLICZ MIN I MAX, MEDIANĘ, WYZNACZ MODĘ, KWARTYLE, IQR;
   - STWÓRZ WYKRES LINIOWY. TERAZ WIDAĆ MIN, MAX I MEDIANĘ.
   - POLICZ ŚREDNIĄ
   - OSZACOWAĆ LICZBĘ BINÓW I STWORZYĆ WIERSZ Z BINAMI DLA HISTOGRAMOWANIA