

Podstawy statystyki praktycznej

Lista 1

Laboratoria

- Zbiór danych `grades.txt` zawiera dane 78 uczniów siódmej klasy pewnej szkoły w USA. Dla każdego ucznia mamy informację o średniej ocen (ocenom F, D-, D, D+, C-, C, C+, B-, B, B+, A-, A, odpowiadają liczby od 0 do 11), wynik standardowego testu IQ, płeć oraz punktacja na teście psychologicznym (Piers-Harris Children's Self-Concept Scale).
 - Narysuj histogramy wszystkich zmiennych ilościowych. Opisz kształt, środek i rozrzut rozkładów. Podaj wartości odpowiednich statystyk (minimum, maximum, mediana, kwartyle, średnia, odchylenie standardowe, wariancja, współczynnik zmienności).
 - Powtórz punkt (a) osobno dla chłopców, a osobno dla dziewcząt. Porównaj wyniki.
- Dla wybranej zmiennej ilościowej ze zbioru `grades.txt` narysuj histogramy dla kilku różnych liczb klas (w tym bardzo małej liczby klas i bardzo dużej liczby klas). Skomentuj jaki efekt otrzymujemy wybierając zbyt mało klas, a jaki efekt, gdy wybieramy zbyt dużo klas.
- Zbiór danych `income.dat` zawiera wyniki ankiet przeprowadzonych przez Bureau of Labor Statistics (USA) na próbie 55 899 osób. Dla każdej osoby podany jest wiek (w latach), wykształcenie (1 = podstawowe, 2 = niepełne średnie, 3 = średnie, 4 = niepełne wyższe, 5 = wyższe (licencjat), 6 = wyższe (magisterium)), płeć (1 = mężczyzna, 2 = kobieta), roczne zarobki (w dolarach) oraz sektor zatrudnienia (5 = sektor prywatny, 6 = sektor publiczny, 7 = samozatrudnienie).
 - Narysuj histogram zarobków. Opisz kształt, środek i rozrzut. Podaj wartości odpowiednich statystyk (minimum, maximum, mediana, kwartyle, średnia, odchylenie standardowe, wariancja, współczynnik zmienności).
 - Powtórz punkt (a) osobno dla mężczyzn, a osobno dla kobiet. Porównaj wyniki.
 - W punkcie (a) wyznacz obserwacje odstające.

Ćwiczenia

- Dana jest średnia, mediana, rozstęp międzykwartylowy, współczynnik zmienności, odchylenie standardowe i wariancja dla pewnej zmiennej X . Wyznacz wartości tych samych statystyk dla nowej zmiennej X' zdefiniowanej następująco (a i b są pewnymi stałymi):

$$X' = aX + b$$

Podaj przykłady praktycznych sytuacji, gdy możemy się zetknąć ze zdefiniowaną w ten sposób liniową zamianą zmiennych.

- Uzasadnij, że średnia zbioru obserwacji $x_1, \dots, x_n$ to taka liczba c , dla której poniższa suma przyjmuje najmniejszą wartość:

$$(x_1 - c)^2 + \dots + (x_n - c)^2$$

Zinterpretuj ten wynik.

- Jeśli próbę $x_1, \dots, x_n \in \mathbb{R}$ potraktujemy jako element $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$, to wyznaczanie średniej próby sprowadza się do wyznaczenia rzutu ortogonalnego wektora x na wektor $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$. Wytlumacz i pokaż związek z poprzednim zadaniem.
- Uzasadnij następujący wzór na odchylenie standardowe:

$$s = \sqrt{\frac{\left(\sum_{i=1}^n x_i^2\right) - n\bar{x}^2}{n-1}}$$

- (Nierówność Czebyszewa) Udowodnij, że mniej niż $\frac{1}{k^2}$ wszystkich obserwacji znajduje się w odległości większej niż ks od średniej (gdzie s to odchylenie standardowe). W jakich sytuacjach osiągnięta proporcja obserwacji przekraczająca odległość ks jest zbliżona do $\frac{1}{k^2}$?