Probabilità e Statistica matematica 21. 01. 2008

1. Calcolare la media, la mediana, il primo e il terzo quartile, la varianza e la deviazione standard dei seguenti dati:

   (a) 0; 10

   (b)

6	$0, 5, 5, 8$
7	$2, 4, 4, 5, 7, 8$
8	$2, 4, 4, 5, 7, 8, 9$
9	$0, 2, 2, 4, 5, 5, 7$
10	$0, 2, 7, 8$
11	$0, 5, 5$
12	$2, 9$

   Qui i dati $60, 65, 65, \ldots, 129$ sono stati riportati in un diagramma *stem and leaf (ramo-foglia)*.

   (c) Il campione (b) ha un coefficiente di asimmetria positivo, negativo o zero?

2. Consideriamo una serie di n dati grezzi che possono assumere soltanto due valori, 0 e 1. Che cosa si può dire della varianza campionaria, se la media campionaria è pari a $1/2$?

3. Dato il campione $[x_1, \ldots, x_n]$ con la media $\overline{x}$ e la varianza s_x^2, si determinino la media e la varianza del "campione standardizzato" $y_i := \dfrac{x_i - \overline{x}}{s_x}$ $(i = 1, \ldots, n)$.

4. Dato il campione di dati numerici $[x_1,\ x_2,\ x_3]$, si calcoli la differenza tra la varianza e la varianza in base ai dati raggruppati $[[x_1, x_2],\ [x_3]]$ del campione (si intende la varianza calcolata con le medie e le frequenze delle classi).

5. Consideriamo i dati osservati (x_i, y_i), $i = 1, \ldots, n$, la loro retta di regressione $y = ax + b$ e i valori previsti $\hat{y}_i := ax_i + b$ dal modello di regressione lineare. Definiamo la *devianza totale*, $DT := \sum_{i=1}^n (y_i - \overline{y})^2$, la *devianza spiegata* dal modello, $DS := \sum_{i=1}^n (\hat{y}_i - \overline{y})^2$ e la *devianza dei residui*, cioè degli errori casuali, $DR := \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Definiamo infine il *coefficiente di determinazione* $R^2 := \frac{DS}{DT} = 1 - \frac{DR}{DT}$. Cio significa che R^2 è la parte della variazione nei dati delle risposte che si spiega con la diversità dei livelli di ingresso. In altre parole, R^2 è un indicatore di quanto bene il modello di regressione interpreti i dati. Si dimostri:

   (a) $\displaystyle\sum_{i=1}^{n} (\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 0$;

   (suggerimento: si usi $\overline{y} = a\overline{x} + b$, $a = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$)

   (b) $DT = DS + DR$;

   (c) R^2 è uguale al quadrato del coefficiente di correlazione r_{xy}.