STATISTICA DESCRITTIVA

Singole variabili e classi

La statistica descrittiva considera un insieme di dati e li elabora. I dati raccolti rappresentano la realizzazione di variabili aleatorie. Si distinguono in variabili continue e discrete. Spesso le singole variabili vengono raggruppate in classi ovvero intervalli contigui. Una volta raggruppati gli n dati in classi si definiscono le frequenze. Esistono diversi tipi di frequenze:

- Frequenza assoluta di una classe f_a = è il numero di osservazioni che ricadono in quella classe
- Frequenza relativa di una classe f_r = è la sua frequenza assoluta divisa per il numero totale di osservazioni
- Frequenza percentuale di una classe f_p = è la sua frequenza relativa moltiplicata per 100
- Frequenza cumulativa di una classe F_a, F_r, F_p = è la somma delle frequenze della classe stessa e di tutte quelle che la precedono

Se i valori diversi osservati in un esperimento non sono troppo numerosi si può scegliere tutte le classi come singoli valori.

Spesso per rappresentare le frequenze si utilizzano gli istogrammi. La scelta delle classi è arbitraria; talvolta può essere conveniente trattare classi di ampiezza diversa. In tal caso solitamente è l'area del rettangolo ad essere proporzionale alla frequenza.

Indici di posizione

Defiliamo ora alcuni indici di posizione:

- Media = si definisce media il numero:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  La media di un insieme di dati si calcola utilizzando lo stimatore media campionaria; è quindi la stima puntuale del valore atteso della variabile aleatoria che modellizza la misura del dato in questione.
- Mediana = disponendo i dati in ordine crescente la mediana è il dato nella posizione centrale se n è dispare, oppure la media aritmetica dei due dati in posizione centrale se n è pari.
- Quantili, percentili e quartili = generalizzando il concetto di mediana cercando un valore q_p con la proprietà che almeno una frazione p dei dati sia non superiore a q_p ed almeno una frazione $1-p$ sia non inferiore a q_p. Si definisce dunque p-esimo quantile:
  - Se np non è intero allora $q_p = x_{k+1}$
  - Se $np = k$ con k interno allora $q_p = \frac{x_k + x_{k+1}}{2}$

  Il p-esimo quantile viene anche detto 100 p-esimo percentile. Il 25°,50°,75° percentile vengono detti anche primo, secondo e terzo quartile e indicati con Q_1, Q_2, Q_3. Il secondo quartile coincide con la mediana
- Moda = è il valore o più in generale la classe in corrispondenza del quale si ha la popolazione più numerosa. Se vi è un solo punto dove la frequenza è massima, si dice che la distribuzione delle frequenze è unimodale, se vi è più di un massimo si dice che la distribuzione delle frequenze è plurimodale

Indici di dispersione

Definiamo ora una serie di indici di dispersione:

- Range = se i dati sono $x_1, x_2, \dots, x_n$ il range è il numero reale:

$$r = max\{x_i : i = 1, \dots\} - min\{x_i : i = 1, \dots\}$$

- Varianza = si definisce varianza:

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

Una definizione alternativa di varianza per un insieme di dati è la seguente:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} {x_i}^2\right) - \overline{x}^2$$

Indici di forma

- La **skewness**

$$\gamma_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^3$$

  È una grandezza *adimensionale*. Può assumere valori sia positivi che negativi.
  Se è negativa denota una *coda* verso sinistra.
  Se è positiva denota una *coda* verso destra.
  se la distribuzione è simmetrica, allora la skewness è nulla, ma l'inverso non è vero.
  Per trasformazioni lineari $y_i = ax_i + b$ la skewness non cambia: $\gamma_3^y = \gamma_3^x$.

- La **curtosi**

$$\gamma_4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^4$$

  È una grandezza adimensionale e non negativa. Misura (in un certo senso) l'appiattimento della distribuzione delle frequenze, poiché assegna un peso elevato agli scarti grandi: valori elevati della curtosi segnalano distribuzioni significativamente diverse da $\overline{x}$ per grandi scarti, piccoli valori distribuzioni *appuntite* in corrispondenza di $\overline{x}$.
  Per trasformazioni lineari $y_i = ax_i + b$ la curtosi non cambia: $\gamma_4^y = \gamma_4^x$.