Appunti

Campione finito di una popolazione

Nelle prime lezioni abbiamo discusso delle connessione tra il calcolo delle probabilita', la statistica e la misura. Abbiamo discusso della metodologia che ci permetta, attraverso l'*inferenza statistica*, di dare una predizione al comportamento della popolazione sulla base dell'analisi del campione di misure[1]. I dati raccolti nel campione, necessariamente finito, di n misure, ci permettono di ricavare i momenti sperimentali, ed in particolare i due momenti principali della distribuzione campionaria:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad ; \quad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

(successivamente, ove non indicato la sommatoria si intendera' per $i = 1,,,n$) L'ipotesi associata alla popolazione e' di avere una distribuzione di probabilita' (o densita' nel caso di variabili continue) con i momenti principali definiti come:

$$\mu = \sum_j x_j F_j \quad ; \quad \sigma^2 = \sum_j (x_j - \mu)^2 F_j$$

con $j = 1,,,N$ i possibili valori della variabile x, ed F_j la probabilita' aleatoria che la variabile x abbia il valore x_j. Possiamo anche scrivere $n_j x_j^k = \sum_{i=1}^{n_j} x_i^k \ \forall \ j = 1,,,N$ (con $n_j = nF_j$ ed $x_i = x_j \ \forall \ i = 1,,,n_j$), che ci permette di definire μ e σ^2 in modo del tutto analogo a quanto fatto per i momenti sperimentali

$$\mu = \frac{\sum x_i}{n} \quad ; \quad \sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

Vogliamo ora dimostrare che i due valori sperimentali $\bar{x}$ e s^2 sono la migliore stima che noi possiamo dare dei valori corrispondenti della distribuzione aleatoria rappresentante la popolazione. Il metodo e' quello di calcolare il valore di aspettazione di queste due quantita'. E' pur vero che per quanto riguarda $\bar{x}$ l'evidenza e' immediata, visto che l'espressione per μ e' formalmente identica. Meno intuitivo e' il calcolo nel caso di s^2.

Sappiamo che $\mu = E[x]$ e che $\sigma^2 = Var[x] = E[(x-\mu)^2] = E[x^2] - E^2[x]$.

$$E[\bar{x}] = E\left[\frac{\sum x_i}{n}\right] = \frac{E[\sum x_i]}{n} = \frac{\sum E[x_i]}{n}$$

ma i valori misurati x_i appartengono alla popolazione e quindi i loro valori di aspettazione sono tutti uguali $E[x_i] = E[x] = \mu$ e quindi

$$E[\bar{x}] = \frac{n\mu}{n} = \mu$$

[1] Probabilita', Statistica e Simulazione, Cap.1 par.2

Come potevamo intuire facilmente, il valore di aspettazione di $\bar{x}$ e' proprio la media μ della distribuzione di probabilita' della popolazione. Possiamo ora valutare la Varianza di $\bar{x}$

$$\sigma^2_{\bar{x}} = Var[\bar{x}] = Var\left[\frac{\sum x_i}{n}\right] = \frac{Var[\sum x_i]}{n^2}$$

Poiche' le grandezze misurate x_i sono tra loro indipendenti (misure indipendenti), i termini di covarianza sono nulli e quindi

$$\sigma^2_{\bar{x}} = \frac{Var[\sum x_i]}{n^2} = \frac{\sum Var[x_i]}{n^2} = \frac{nVar[x]}{n^2} = \frac{\sigma^2}{n}$$

Questo risultato ci dice (per la prima volta!) che facendo piu' misure di una stessa grandezza, migliora la determinazione del valor medio della distribuzione di probabilita' della grandezza. Questo non significa che facendo piu' misure si ottengono via via valori sempre piu' vicini a $\bar{x}$, in quanto la dispersione dei valori della grandezza e' data da s^2.

Di questa grandezza valuteremo ora il valore di aspettazione.

$$E[s^2] = E\left[\frac{\sum(x_i - \bar{x})^2}{n-1}\right]$$

Per fare questo dobbiamo partire da un'espressione nota

$$\begin{aligned}
\sum(x_i-\mu)^2 &= \sum(x_i - \bar{x} + \bar{x} - \mu)^2 \\
&= \sum[(x_i-\bar{x}) + (\bar{x}-\mu)]^2 \\
&= \sum(x_i-\bar{x})^2 + \sum(\bar{x}-\mu)^2 + \sum 2(x_i-\bar{x})(\bar{x}-\mu) \\
&= \sum(x_i-\bar{x})^2 + \sum(\bar{x}-\mu)^2 + 2(\bar{x}-\mu)\sum(x_i-\bar{x})
\end{aligned}$$

ma l'ultimo termine e' nullo per la proprieta' del valor medio di rendere nulla la somma degli scarti. spostando i termini dell'equazione si ottiene

$$\sum(x_i-\bar{x})^2 = \sum(x_i-\mu)^2 - \sum(\bar{x}-\mu)^2$$

che ci permette di scrivere il valore di aspettazione di s^2 nel seguente modo

$$\begin{aligned}
E[s^2] &= E\left[\frac{\sum(x_i-\mu)^2 - \sum(\bar{x}-\mu)^2}{n-1}\right] \\
&= \frac{1}{n-1}\left(E[\sum(x_i-\mu)^2] - E[\sum(\bar{x}-\mu)^2]\right) \\
&= \frac{1}{n-1}\left(\sum E[(x_i-\mu)^2] - \sum E[(\bar{x}-\mu)^2]\right) \\
&= \frac{1}{n-1}\left(\sum \sigma^2] - \sum \sigma^2_{\bar{x}}\right) \\
&= \frac{n\sigma^2 - n\sigma^2/n}{n-1} \\
E[s^2] &= \sigma^2
\end{aligned}$$

La nostra intuitiva definizione di s^2, basata su considerazioni di opportunita‘ (ovvero evitare che con una singola misura potessimo definire una dispersione nulla della grandezza) si dimostra essere la migliore definizione di varianza campionaria. La scelta di $n-1$ al posto di n nasce dal fatto che nel calcolare s^2 abbiamo utilizzato $n-1$ gradi di liberta‘. Il numero dei “gradi di liberta‘ di una statistica” e‘ definito dal numero n di variabili campione meno il numero k di parametri stimati dai dati. Nel nostro caso per stimare $\bar{x}$ i gradi di liberta‘ erano tutti i nostri dati, mentre per il calcolo di s^2 abbiamo utilizzato un parametro stimato dai dati stessi, ($\bar{x}$ per l'appunto).