

Stima della media di una v. statistica

Consideriamo una variabile statistica x la cui media μ (su tutta la popolazione) intendiamo stimare. Scegliamo in modo casuale un campione di n individui, cui sono associati altrettanti valori della variabile x , denotati $x_1, \dots, x_n$.

È ragionevole stimare μ con la media di questi valori, detta **media campionaria**.

$$\hat{\mu} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$\hat{\clubsuit} = \text{"stima di } \clubsuit \text{"}$

A questa stima, possiamo associare un **intervallo di confidenza** in due modi.

Metodo spiccio

Valido per campioni grandi e popolazioni grandi rispetto al campione.

Si stima la deviazione standard σ di x con la deviazione standard del campione

$$\hat{\sigma} = \sigma(\text{campione}) = \sqrt{\frac{(x_1 - \hat{\mu})^2 + \dots + (x_n - \hat{\mu})^2}{n}}$$

Ora, al variare di tutti i possibili campioni di n individui, le possibili medie campionarie si distribuiscono (indipendentemente dalla distribuzione di x) in modo circa gaussiano. La media di questa gaussiana è μ e la deviazione standard è l'errore standard della media

"Standard Error of the Mean"

$$SEM = \frac{\sigma}{\sqrt{n}} \quad \text{che stimiamo con} \quad \widehat{SEM} = \frac{\hat{\sigma}}{\sqrt{n}}$$

In pratica dobbiamo usare la stima $\widehat{SEM}$ perché σ non è nota.

Finalmente, siccome la distribuzione delle medie campionarie è gaussiana e ne sappiamo i parametri, possiamo calcolare con quale probabilità la nostra particolare media campionaria $\hat{\mu}$ si trovi in un dato intervallo attorno alla media vera μ . Per esempio.

$$\begin{array}{lcl} \hat{\mu} \text{ è in } \mu \pm \widehat{SEM} & \text{nel} & 68\% \text{ dei casi} \\ \hline \mu \pm 2 \widehat{SEM} & - & 95\% \\ \hline \mu \pm 3 \widehat{SEM} & - & 99.7\% \end{array}$$

Questi valori sono approssimati: cifre più precise si possono facilmente estrarre da una

tabella della distribuzione normale; per esempio al 95% corrisponde più precisamente l'intervallo $\pm 1.96 \widehat{SEM}$.

Per concludere, si osserva che dire

a è in $b \pm \delta$ equivale a b è in $a \pm \delta$
per cui

μ è in $\hat{\mu} \pm \widehat{SEM}$ nel 68% dei casi

_____ $\hat{\mu} \pm 2 \widehat{SEM}$ — 95% —

_____ $\hat{\mu} \pm 3 \widehat{SEM}$ — 99.7% —

Questi sono detti intervalli di confidenza e le percentuali sono i rispettivi livelli di confidenza. Valori tipici sono

l'i. di c. al 95% è $\hat{\mu} \pm 1.96 \widehat{SEM}$

_____ 99% — $\hat{\mu} \pm 2.58 \widehat{SEM}$

Metodo più generale

Valido anche per campioni e popolazioni piccole, a patto che la distribuzione di x non si allontani fortemente da una gaussiana.

Comunque, se la distribuzione di x è pazzesca, non c'è metodo che tenga.

Innanzitutto rimpiazziamo $\hat{\sigma} = \sigma(\text{campione})$ con

$$\hat{\sigma} = s(\text{campione}) = \sqrt{\frac{(x_1 - \hat{\mu})^2 + \dots + (x_n - \hat{\mu})^2}{n - 1}}$$

$s^2(\text{campione}) = \text{"varianza campionaria"}$

Stimiamo l'errore standard della media con

$$\widehat{SEM} = \frac{\hat{\sigma}}{\sqrt{n}} \cdot \sqrt{\frac{\# \text{ popolazione} - n}{\# \text{ popolazione}}}$$

"finite population correction"

si applica per campioni senza rimpiazzamento
se la popolazione è piccola rispetto al campione

Ora, la quantità $\frac{\hat{M} - \mu}{\widehat{SEM}}$ si distribuisce, al variare di tutti i possibili campioni, secondo la distribuzione detta *t di Student* con $n-1$ gradi di libertà (ce ne è una per ogni numero di gradi di libertà). Se il numero di gradi di libertà è alto, la *t di Student* coincide, sostanzialmente, con la gaussiana standard. Altrimenti è notevolmente diversa, e queste distribuzioni sono tabulate.

Student's curve, with degrees of freedom shown at the left of the table

The shaded area is shown along the top of the table

95%
2.5%
-t
t
is shown in the body of the table

Degrees of freedom	25%	10%	5%	conf. 95% 2.5%	1%	conf. 99% 0.5%
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
<u>9</u> = 10-1	0.70	1.38	1.83	<u>2.26</u>	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17

Per esempio, per un campione di $n = 10$ individui, vediamo che

$\frac{\hat{\mu} - \mu}{\widehat{SEM}}$ è in 0 ± 2.26 nel 95% dei casi

Per cui Anziché 1.96 secondo il metodo spiccio.

μ è in $\hat{\mu} \pm 2.26 \widehat{SEM}$ nel 95% dei casi

e questo è il nostro intervallo di confidenza della media.