ALCUNI ELEMENTI DI TEORIA DELLA STIMA

Quando si vuole valutare un parametro θ (ad esempio: media, varianza, proporzione, coefficiente di regressione lineare, coefficiente di correlazione lineare, ecc) di una popolazione mediante un campione casuale, la **stima** del parametro può essere espressa mediante un unico valore (**stima puntuale)** desunto dal campione considerato oppure da un intervallo di valori (**stima intervallare**) entro cui, con un dato *livello di fiducia*, si ritiene cada il valore vero del parametro θ della popolazione.

Per ***Stimatore*** T di un parametro ϑ si intende una statistica calcolata sul campione (media campionaria $\overline{X}$, varianza campionaria S^2, percentuale campionaria P, ecc.) il cui valore si chiama ***stima***. Il valore della stimatore (stima) varia in corrispondenza del campione estratto.

Stima Puntuale

Stima puntuale di una media. La stima puntuale di una media consiste nel valutare, per mezzo di un campione, il valore μ (media) della popolazione. Come stimatore si prende la media campionaria $\overline{X}$ in quanto esso è uno corretto, consistente ed efficiente. Ad $\overline{X}$ si associa la sua deviazione standard nello spazio campionario, detta errore medio di campionamento o ***Errore Standard*** (ES): $ES = \frac{\sigma}{\sqrt{n}}$, dove σ^2 è la varianza della popolazione. L'ES è inversamente proporzionale alla radice quadrata della numerosità campionaria n, maggiore è n e minore è l'errore di campionamento.

Stima puntuale della varianza. Uno stimatore corretto della varianza σ^2 della popolazione è: $S^2 = \frac{\sum_i (X_i - \overline{X})^2}{n-1}$, essendo $(X_1, X_2, \ldots, X_n)$ il campione.

Stima puntuale di una proporzione. Uno stimatore corretto di una proporzione π della popolazione è la frequenza relativa $P = \frac{\sum_i X_i}{n}$ del campione, essendo $X_i = 1$ oppure $=0$ a seconda se l'evento di cui si vuole stimare la proporzione si è verificato o meno. La deviazione standard della proporzione (E.S.) risulta: $ES = \sqrt{\frac{p(1-p)}{n}}$ [1].

Stima Intervallare

Una volta note le distribuzioni di probabilità degli stimatori puntuali dei parametri di una popolazione, è possibile studiare la bontà della stima campionaria di un parametro incognito.

Stima intervallare di una media

Si supponga, ad esempio, di voler stimare il valor medio μ di una popolazione che si distribuisce normalmente con deviazione standard $\sigma = 2$. Si effettua un campionamento di n = 36 osservazioni indipendenti. È noto che in questo caso lo stimatore $\overline{X}$ è una variabile aleatoria distribuita normalmente,

[1] Si ricorda che, nel caso di campionamento da popolazione finita di numerosità N, la formula dell'errore standard va "corretta" moltiplicando per il fattore $\sqrt{\frac{N-n}{N-1}}$.

che ha valor medio μ e deviazione standard $\sigma/\sqrt{n} = 1/3$. Estratto un campione casuale, si ipotizzi di aver ottenuto un valore $\bar{x} = 13.8$ per la media campionaria.
Come si usa questa informazione? Una variabile aleatoria $\bar{X}$ distribuita secondo la normale ha una probabilità nota $1-\alpha$ di trovarsi in un intervallo centrato attorno al suo valor medio e di ampiezza data pari ad **a:**

$$Prob(\mu - a < \bar{X} < \mu + a) = 1 - \alpha .$$

Risolvendo rispetto a μ, si può anche scrivere:

$$Prob(\bar{X} - a < \mu < \bar{X} + a) = 1 - \alpha \qquad (*)$$

che esprime il fatto che l'intervallo casuale $(\bar{X} - a,\ \bar{X} + a)$ contiene al suo interno la media μ con probabilità pari a $1-\alpha$. Si ottiene, in tal modo, un ***intervallo di fiducia*** per la media μ.

L'approccio classico al problema è dunque il seguente: fissato un valore di probabilità, ad esempio $1-\alpha$=0.95, si determina il valore a = $\mathbf{a_{0.95}}$ in modo che sia soddisfatta la (*). Così facendo possiamo affermare che il *"vero"* medio μ si trova, con probabilità $1-\alpha$ = .95, nell'intervallo $(\bar{X} - a_{0.95},\ \bar{X} + a_{0.95})$.

Una volta effettuato il campionamento e calcolato $\bar{x} = 13.8$, l'intervallo di confidenza associato a tale campione è $(13.8 - a_{0.95},\ 13.8 + a_{0.95})$: ma non è corretto dire che la probabilità che μ cada in tale intervallo è 0.95. In realtà, visto che la probabilità che la media appartenga a $(\bar{X} - a_{0.95},\ \bar{X} + a_{0.95})$ è 0.95, si ha ***fiducia*** che l'intervallo ottenuto della media campionaria $\bar{x} = 13.8$ contenga il valore *vero* e *incognito* della media μ.

In termini operativi ciò significa che estraendo ipoteticamente 100 campioni della stessa numerosità n, ci si aspetta che per 95 di essi la media μ appartenga all'intervallo $(\bar{x} - a_{0.95},\ \bar{x} + a_{0.95})$ della media $\bar{x}$ ottenuta in ciascun campione. Il grado di fiducia che si attribuisce alla stima è espresso dal livello di probabilità $1-\alpha$, detto ***livello di fiducia***.

Resta da calcolare il valore di $\mathbf{a_{0.95}}$**.** Per fare questo basta ricordare il procedimento di standardizzazione di una variabile normale. La variabile $Z = (\bar{X} - \mu)/\left(\sigma/\sqrt{n}\right)$ si distribuisce secondo la normale standard e la disuguaglianza $\bar{X} - a_{0.95} < \mu < \bar{X} + a_{0.95}$ è equivalente alla $-a_{0.95}\sqrt{n}/\sigma < z < +a_{0.95}\sqrt{n}/\sigma$. Quindi si ha: $a_{0.95} = z_c\ \sigma/\sqrt{n}$, essendo $\mathbf{z_c}$ il valore della variabile normale standard per cui Prob(-$\mathbf{z_c}$ < z < $\mathbf{z_c}$) = 0.95. In realtà il valore $\mathbf{z_c}$ (detto *valore critico*) è tale che le due code rispettivamente a destra di z_c e a sinistra di $-z_c$ abbiamo entrambe probabilità uguale a $(1-0.95)/2 = 0.025$, ossia del 2.5%. Utilizzando la tabella della distribuzione normale standard (Appendice Dispensa Distribuzione Normale) si ottiene che $\mathbf{z_c} \approx$ **1.96.** Si può allora concludere con i dati del nostro esempio che: il vero valore del valor medio μ si trova, con un livello di confidenza del 95%, nell'intervallo (13.8 - 1.96/$\sqrt{9}$, 13.8 + 1.96/$\sqrt{9}$) ≈ (13.15, 14.45).

Si osservi che per ogni estrazione di campione si ottiene un intervallo di confidenza diverso. Ognuno di essi è un intervallo di confidenza "lecito".

C'è un modo un pò diverso di interpretare lo stesso calcolo. Secondo questo punto di vista diremo che il valor medio μ appartiene all'intervallo appena costruito al *livello di errore del 5%.*. Questo significa che se si assume che il valor medio appartiene all'intervallo si può commettere un errore (cioè μ può anche non appartenere all'intervallo), ma la probabilità di sbagliare è solo del 5%.

Esempio1. Si vuole stimare il "vero" valore medio μ dell'uricemia in una popolazione maschile; è noto che in tale popolazione la dispersione dell'uricemia è σ = 1.1 mg/dl. Si assume un livello di confidenza del 95%.

Si supponga di estrarre un campione casuale di 40 soggetti dalla popolazione maschile, di determinare il valore di uricemia per ognuno dei 40 soggetti e di ottenere un valore della media pari a 5.55 mg/dl.
L'intervallo di confidenza della media al 95% è pari a: $(5.55 - 1.96*1.1/\sqrt{40}\ ,\ 5.55 + 1.96*1.1/\sqrt{40}) \approx (5.21, 5.89)$. Pertanto il parametro incognito μ è compero tra 5.21 e 5.89 e si ha *quasi* la certezza (confidenza del 95%) che ciò è vero. Naturalmente *l'affermazione potrebbe anche essere falsa* (infatti c'è una probabilità del 5% che l'intervallo non includa il parametro), ma si ritiene che tale deprecabile eventualità sia così poco probabile da non capitare.

Esempio2. Si vuole stimare la pressione arteriosa (PAS) di maschi di età 45-64. Si è misurata la PAS a 36 uomini nella fascia di età di interesse selezionati a caso a partire dalla lista dei pazienti di un medico di base. Si è trovato che la PAS media sul campione è pari a 144 mmHg. Si ipotizzi di conoscere che $\sigma = 24.0$ mmHg. Si intende calcolare l'intervallo di confidenza al 95% della PAS e stabilire se la PAS media dei maschi di età 45-64 possa essere pari a 150 mmHg.
L'intervallo di confidenza della media al 95% è pari a: $(144 - 1.96*24/\sqrt{36}\ ,\ 5.55 + 1.96*24/\sqrt{36}) \approx (136.16, 151.84)$. Pertanto la PAS media è compresa (con un livello di fiducia del 95%) tra 136.16 e 151.64; poiché il valore 150 è compreso nell'intervallo esso è uno dei valori plausibile per la pressione sistolica.

Esempio3. In un campione di 11 uomini estratti a caso da una data popolazione sono stati rilevati i seguenti valori di colesterolo (mg/100ml): 265, 208, 361, 143, 310, 252, 239, 225, 184, 220, 332. Assumendo che il livello di colesterolo abbia distribuzione normale con σ = 65 mg/100ml, si vuole determinare l'intervallo di confidenza al 99% per il valor medio del colesterolo nella popolazione di riferimento.
Utilizzando la tabella della distribuzione normale standard (Appendice) si ottiene che il valore $\mathbf{z_c}$ della variabile normale standard per cui Prob($-\mathbf{z_c} < z < \mathbf{z_c}$) = 0.99 è pari a **2.58**. Osservato che il valore medio campionario è pari a 249, un intervallo di confidenza per la media è: $249 \pm 2.58 \cdot 65/\sqrt{11} \approx (198.44, 299.56)$.

Se si effettua un campionamento a partire da una popolazione distribuita normalmente, ma di cui si ignora sia il valor medio che la deviazione standard, si può procedere come sopra semplicemente sostituendo la deviazione standard campionaria s a quella della popolazione σ e la distribuzione t di Student (con opportuni gradi di libertà) alla distribuzione normale standard. Di conseguenza, al posto dei valori critici z_c, si avranno valori critici $t_{c,n-1}$, dipendenti questa volta non solo dalla probabilità che l'intervallo deve avere di contenere il valor medio della popolazione, ma anche dalla dimensione n del campione.

Si supponga, ad esempio, di disporre di un campione di 16 valori e di aver ottenuto una media campionaria $\bar{x} = 3$ e una varianza campionaria $s^2 = 4$. Per costruire l'intervallo di confidenza al 95% per μ si deve sfruttare il fatto che la variabile aleatoria $\left(\bar{X} - \mu\right)/\left(s/\sqrt{n}\right)$ è distribuita secondo una t di Student con n−1 = 15 gradi di libertà e che, quindi, al 95% corrisponde il valore critico $t_{c,n-1}$ è pari a $t_{0.95,15} = 2.13$ (vedere tavola della distribuzione t di Student in Appendice alla Dispensa sul Campionamento per α=0.025). L'intervallo di confidenza è quindi dato da $\left(3 - 2.13 \cdot 2/\sqrt{16},\ 3 + 2.13 \cdot 2/\sqrt{16}\right)$, ovvero (1.935, 4.065).

Stima intervallare di una frequenza o proporzione

Il problema della stima intervallare di una frequenza relativa di una modalità di un carattere consiste nell'individuazione, sulla base di un campione, di un intervallo reale entro cui il valore della frequenza nella popolazione di riferimento (non noto) cada con un dato ***livello di fiducia***. Come stima puntuale della frequenza relativa π della popolazione si prende la frequenza relativa p del campione, in quanto stima *corretta* di π.
Per costruire l'intervallo di confidenza per π occorre conoscere la distribuzione della frequenza relativa campionaria $P = \frac{F}{n}$. Per una numerosità campionaria "abbastanza" grande, P si distribuisce secondo una

curva normale con media π e deviazione standard $DS(P) = \sqrt{\frac{\pi(1-\pi)}{n}}$. Pertanto, attraverso l'utilizzo della distribuzione normale standard, può essere determinata la probabilità che la frequenza relativa standardizzata: $Z = \frac{P-\pi}{DS(P)}$ appartenga ad un dato intervallo reale: (-a, a). Se allora, si rileva una frequenza relativa sul campione pari a p, si può costruire l' ***intervallo di fiducia*** per la frequenza relativa della popolazione: si è "fiduciosi al $(1-\alpha)\%$ " che tale frequenza cada nell'intervallo:

$$p - z_c\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_c\sqrt{\frac{p(1-p)}{n}}.$$

essendo $\mathbf{z_c}$ il valore della variabile normale standard per cui Prob($-\mathbf{z_c} < z < \mathbf{z_c}$) = $(1-\alpha)$.

Esempio. In un campione casuale semplice di 400 individui estratto da una popolazione di fumatori, risulta che 80 sono "forti" fumatori (fumano più di un pacchetto di sigarette al giorno). Si vuole stimare la proporzione π di "forti" fumatori nella popolazione di riferimento mediante un intervallo di confidenza al 95%.

Osservato che p = 80/400 = 0.2, $\alpha = 0.05$ e $z_c = 1.96$, l'intervallo di confidenza è dato da:

$$0.2 - 1.96\sqrt{\frac{0.2\cdot 0.8}{400}} < \pi < 0.2 - 1.96\sqrt{\frac{0.2\cdot 0.8}{400}}$$

ovvero: (0.16; 0.24) = (16%; 24%).

QUESITI

1) In una popolazione di uomini infartuati al miocardio, il livello di colesterolo medio è pari a 240 mg/dl con una deviazione standard di 40 mg/dl. Estraendo casualmente un campione di 100 soggetti si è trovata una media di 235 mg/dl. Qual è la probabilità che il livello medio di colesterolo sia maggiore o uguale a 260 mg/dl? Qual è l'intervallo di confidenza per la media μ della popolazione al livello del 95% ?

Risposte: $\Pr ob(\bar{x} > 260) = \Pr ob\left(z > \frac{260-240}{40/\sqrt{100}}\right) = P\left(z > \frac{20}{4}\right) = P(z > 5) = 0$;

$235 \pm 1.96\cdot 40/\sqrt{100} \approx (227.16, 242.84)$.

2) L'indice di massa corporea BMI (kg/m^2) misura il grado di soprappeso di un soggetto. Per la popolazione di uomini di mezza età che svilupperanno diabete mellito, la distribuzione di BMI ha forma approssimativamente normale con media μ non nota è deviazione standard σ=2.7 kg/m^2. Un campione casuale di 58 soggetti selezionati da questo gruppo ha fatto registrare una media di 25 kg/m^2. Determinare l'intervallo di confidenza al 99% per la media della popolazione di diabetici.

Risposta: $25 \pm 2.58\cdot 2.7/\sqrt{58} \approx (24.1, 25.9)$.

3) Si supponga di aver calcolato l'intervallo di confidenza al 95% della media di una popolazione. L'intervallo al 99% sarà più grande o più piccolo?

4) Il peso medio di un campione di 81 adulti è risultato pari a 80 kg. Sapendo che la deviazione standard della popolazione è pari a 5, costruire l'intervallo di confidenza al 90% e 95% per la media della popolazione.

5) L'età media di un gruppo di 10 studenti che hanno appena conseguito un diploma di laurea triennale è 22 anni. Costruire un intervallo di confidenza al 95% per la media della popolazione degli studenti iscritti al corso di laurea sapendo che tale popolazione si distribuisce normalmente con *varianza* pari a 45.

6) In un campione di 250 individui a cui era stato somministrato un vaccino anti-influenzale, 35 individui contraggono la malattia. Calcolare l'intervallo di confidenza della proporzione di individui ammalati

7) Un comitato vuole stimare la proporzione di persone che utilizzano un personal computer. Vengono intervistate 370 persone e si stabilisce che 214 di queste utilizzano un PC. Si determini la stima per intervalli con confidenza al 95% per la frequenza relativa di utilizzatori di PC.

8) Da una popolazione "grande" di studenti è stato estratto un campione casuale con i risultati di tabella:

Classi di peso (kg)	40-50	50-60	60-70	70-80	80-90
Frequenza assoluta	7	40	60	20	3

a) stimare la deviazione standard dei pesi degli alunni della scuola;
b) stimare, mediante intervalli di confidenza ad un livello di fiducia del 95%, il peso medio degli alunni della scuola.