Azzolini Riccardo 2020-05-12

Stime puntuali e per intervallo

1 Statistiche e stimatori

Definizione: Dato un campione $X = (X_1, \ldots, X_n)$, si chiama **statistica** una variabile aleatoria $T = t(X_1, \ldots, X_n)$, cioè una variabile aleatoria che è funzione del campione.

Come al solito, le variabili X_i che costituiscono il campione sono indipendenti e identicamente distribuite, e la loro distribuzione coincide con quella della popolazione. Nei problemi di statistica, tale distribuzione è tipicamente sconosciuta.

Un caso comune è quello in cui si sa il tipo di legge (binomiale, normale, ecc.), ma non se ne conosce un determinato parametro, indicato solitamente con θ: esso può essere un singolo valore (ad esempio, per una legge $B(n, p)$ con n noto, $\theta = p$), oppure multidimensionale (come ad esempio la coppia $\theta = (\mu, \sigma^2)$ di media e varianza che determinano una distribuzione normale). Convenzionalmente, si indica poi con Θ l'insieme dei valori che θ può assumere (ad esempio, $\Theta = [0, 1]$ per $\theta = p$, e $\Theta = \mathbb{R} \times \mathbb{R}^+$ per $\theta = (\mu, \sigma^2)$).

Siccome il parametro θ determina la distribuzione delle X_i, esso influisce sui risultati dei calcoli di probabilità, valore medio, varianza, ecc. effettuati su tali variabili, e su altre variabili definite in funzione delle X_i, come ad esempio una statistica $T = t(X_1, \ldots, X_n)$. Per mettere in evidenza questo fatto, spesso si aggiunge θ alla notazione: $P^\theta\{X_i \in A\}$, $E^\theta(X_i)$, $\mathrm{Var}^\theta(X_i)$ o $\mathrm{Var}_\theta(X_i)$, ecc.

Il parametro sconosciuto θ può essere stimato da un'opportuna statistica T. Più in generale, una statistica può essere usata per stimare un valore calcolabile a partire da θ, cioè una qualche funzione $\psi(\theta)$ – con un certo abuso di linguaggio, anche $\psi(\theta)$ può essere chiamato un parametro. Formalmente:

Definizione: Data una funzione $\psi : \Theta \to \mathbb{R}^m$, una statistica T (a valori in $\mathbb{R}^m$) è uno **stimatore** del parametro $\psi(\theta)$ se viene usata per stimare il valore (sconosciuto) di $\psi(\theta)$.

- Uno stimatore $T = t(X_1, \ldots, X_n)$ è **non distorto** o **corretto** quando $E^\theta(T) = \psi(\theta)$.
- Uno stimatore $T = t(X_1, \ldots, X_n)$ è **consistente** quando $\mathrm{Var}^\theta(T) \to 0$ per $n \to \infty$.
- Uno stimatore T_1 è **più efficiente** di un altro stimatore T_2 quando $\mathrm{Var}^\theta(T_1) \leq \mathrm{Var}^\theta(T_2)$.

2 Stimatori di media e varianza

Sia $X_1, \ldots, X_n$ un campione estratto da una popolazione caratterizzata da un parametro θ. Degli stimatori per la media $E^\theta(X_i)$ e la varianza $\mathrm{Var}^\theta(X_i)$ della popolazione sono le seguenti statistiche, già presentate in precedenza.

2.1 Media campionaria

La *media campionaria*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

è uno stimatore non distorto e consistente della media della popolazione:

$$E^\theta(\overline{X}) = E^\theta(X_i)$$
$$\mathrm{Var}^\theta(\overline{X}) = \frac{\mathrm{Var}^\theta(X_i)}{n} \to 0 \quad \text{per } n \to \infty$$

2.2 Varianza campionaria a media nota

Supponendo nota la media della popolazione, uno stimatore non distorto e consistente della varianza è dato dalla *varianza campionaria a media nota*:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - E^\theta(X_i))^2$$

2.3 Varianza campionaria a media incognita

Se, invece, la media della popolazione non è nota, si può calcolare la varianza rispetto alla media campionaria,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

ma, come già accennato, questa fornisce una stima distorta di $\mathrm{Var}^\theta(X_i)$:

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\overline{X})^2 &= \sum_{i=1}^{n}(X_i^2-2\overline{X}X_i+\overline{X}^2)\\
&= \sum_{i=1}^{n}X_i^2-\sum_{i=1}^{n}2\overline{X}X_i+\sum_{i=1}^{n}\overline{X}^2\\
&= \sum_{i=1}^{n}X_i^2-2\overline{X}\sum_{i=1}^{n}X_i+\overline{X}^2\sum_{i=1}^{n}1\\
&= \sum_{i=1}^{n}X_i^2-2\overline{X}\cdot\frac{n}{n}\sum_{i=1}^{n}X_i+n\overline{X}^2\\
&= \sum_{i=1}^{n}X_i^2-2\overline{X}\cdot n\overline{X}+n\overline{X}^2\\
&= \sum_{i=1}^{n}X_i^2-2n\overline{X}^2+n\overline{X}^2\\
&= \sum_{i=1}^{n}X_i^2-n\overline{X}^2
\end{aligned}$$

e, dalla formula per la varianza $\mathrm{Var}(X)=E(X^2)-(E(X))^2$, si ricava

$$\begin{aligned}
\mathrm{Var}^\theta(X_i) &= E^\theta(X_i^2)-(E^\theta(X_i))^2 & \mathrm{Var}^\theta(\overline{X}) &= E^\theta(\overline{X}^2)-(E^\theta(\overline{X}))^2\\
E^\theta(X_i^2) &= \mathrm{Var}^\theta(X_i)+(E^\theta(X_i))^2 & E^\theta(\overline{X}^2) &= \mathrm{Var}^\theta(\overline{X})+(E^\theta(\overline{X}))^2\\
& & E^\theta(\overline{X}^2) &= \frac{\mathrm{Var}^\theta(X_i)}{n}+(E^\theta(X_i))^2
\end{aligned}$$

quindi

$$\begin{aligned}
E^\theta\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) &= \frac{1}{n}E^\theta\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}E^\theta(X_i^2)-nE^\theta(\overline{X}^2)\right)\\
&= \frac{1}{n}\left(nE^\theta(X_i^2)-nE^\theta(\overline{X}^2)\right)\\
&= E^\theta(X_i^2)-E^\theta(\overline{X}^2)\\
&= \mathrm{Var}^\theta(X_i)+(E^\theta(X_i))^2-\frac{\mathrm{Var}^\theta(X_i)}{n}-(E^\theta(X_i))^2\\
&= \frac{n\,\mathrm{Var}^\theta(X_i)-\mathrm{Var}^\theta(X_i)}{n}\\
&= \frac{n-1}{n}\,\mathrm{Var}^\theta(X_i)<\mathrm{Var}^\theta(X_i)
\end{aligned}$$

cioè la stima ottenuta è in media più piccola della reale varianza della popolazione.

Da questo, si può tuttavia ricavare uno stimatore non distorto, chiamato *varianza campionaria* (*a media incognita*), dividendo per $n-1$ invece che per n:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

3 Problema: stime di media e varianza

Problema: Dato un campione di 5 misurazioni di diametro di una sferetta, in centimetri,

$$6.33 \quad 6.37 \quad 6.36 \quad 6.32 \quad 6.37$$

trovare stime corrette per la media e la varianza della popolazione.

Soluzione: La media della popolazione può essere stimata in modo corretto (e, si può dimostrare, efficiente) usando semplicemente la media campionaria $\overline{X}$, che sul campione dato assume il valore

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35\,\mathrm{cm}$$

Non essendo nota la media della popolazione, una stima corretta (e anche questa efficiente) della varianza è data dalla varianza campionaria a media incognita, S^2, che per questo campione vale:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{0.02^2 + 0.02^2 + 0.01^2 + 0.03^2 + 0.02^2}{4} = 0.00055\,\mathrm{cm}^2$$

Osservazione: Per indicare i valori assunti dalle variabili aleatorie X_i, $\overline{X}$, S^2, ecc. su uno specifico campione, si usano le lettere minuscole (x_i, $\bar{x}$, s^2, ecc.), al fine di distinguere questi singoli valori dalle variabili aleatorie stesse.

4 Intervalli di confidenza

Quella considerata finora è la **stime puntuali** di un parametro, che ne indica il valore approssimato sotto forma di un singolo numero. Essa può essere generalizzata alla **stima per intervallo**, che fornisce invece (gli estremi di) un intervallo di valori tra i quali si può supporre, con una certa probabilità, che sia compreso il parametro. Formalmente:

Definizione: Date due statistiche $T_1 = t_1(X_1, \ldots, X_n)$ e $T_2 = t_2(X_1, \ldots, X_n)$, si dice che $I_X = [T_1, T_2]$ (o $I_X = (T_1, T_2)$) è un **intervallo di confidenza** (o **di fiducia**) per $\psi(\theta)$,

di **livello** (**grado di fiducia**) $1-\alpha$, con $0 < \alpha < 1$, se, per ogni valore $\theta \in \Theta$ del parametro sconosciuto, si ha

$$P^{\theta}\{I_X \ni \psi(\theta)\} \geq 1-\alpha$$

Osservazione: I_X è un intervallo aleatorio, che varia in funzione del campione $X = (X_1, \ldots, X_n)$, poiché i suoi estremi sono determinati dalle variabili aleatorie T_1 e T_2, e queste sono a loro volta definite in funzione di X. Dunque, $\{I_X \ni \psi(\theta)\}$ è un evento di appartenenza di un valore a un intervallo, ma, al contrario degli altri eventi del genere visti finora, qui è l'intervallo I_X a essere aleatorio, mentre il valore che vi appartiene, $\psi(\theta)$, è costante, in quanto proprietà intrinseca della popolazione. Proprio per mettere in evidenza questa particolarità, l'evento viene scritto come $\{I_X \ni \psi(\theta)\}$ (piuttosto che come $\{\psi(\theta) \in I_X\}$), in base alla convenzione di tenere a sinistra la variabile aleatoria.

5 Intervalli di confidenza per la media a varianza nota

Si consideri una popolazione di varianza nota σ^2 e media incognita μ, dalla quale viene estratto un campione di ampiezza n. Per il teorema del limite centrale, se è verificata la regola pratica $n \geq 30$, la statistica

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

ha approssimativamente la distribuzione normale standardizzata. Come caso particolare, se la popolazione ha già una distribuzione normale, Z è normale standardizzata anche per $n < 30$.

Si chiama **valore critico** $z_{\alpha/2}$ il valore tale che

$$P\{Z > z_{\alpha/2}\} = \frac{\alpha}{2}$$

cioè tale che l'area sottesa alla distribuzione normale standardizzata a destra di $z_{\alpha/2}$ valga $\frac{\alpha}{2}$. Per simmetria, si ha anche che

$$P\{Z < -z_{\alpha/2}\} = \frac{\alpha}{2}$$

e, allora, l'area compresa tra $-z_{\alpha/2}$ e $z_{\alpha/2}$ vale

$$P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - (P\{Z < -z_{\alpha/2}\} + P\{Z > z_{\alpha/2}\}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha$$

In altre parole, la disuguaglianza

$$-z_{\alpha/2} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}$$

è soddisfatta con probabilità $1-\alpha$. Risolvendo tale disuguaglianza rispetto a μ, si ottiene

$$-z_{\alpha/2} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}$$
$$-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$$
$$-\overline{X}-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X}+z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$$
$$\overline{X}-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$$

Quindi, estraendo dalla popolazione un campione di ampiezza n, e calcolando il valore $\overline{x}$ della media campionaria per tale campione, si ha probabilità $1-\alpha$ che l'intervallo

$$\left(\overline{x}-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}},\ \overline{x}+z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

contenga la media μ della popolazione. Questo è allora l'*intervallo di confidenza per la media* μ, con *grado di fiducia* $1-\alpha$ (spesso espresso in percentuale, $(1-\alpha)\cdot 100\ \%$), per *grandi campioni* ($n \geq 30$) o per *popolazioni normali* (anche con campioni piccoli).

6 Problema: effetti collaterali

Problema: In un campione di 400 persone, alle quali è stato somministrato un dato vaccino 136 di esse hanno avuto effetti collaterali di un certo rilievo. Determinare un intervallo di confidenza con grado di fiducia del 95 % per la proporzione della popolazione che soffre di tali effetti collaterali.

Soluzione: Sia p la proporzione (incognita) della popolazione che soffre di effetti collaterali. Allora, la popolazione corrisponde a una variabile aleatoria di Bernoulli che assume il valore 1 (cioè "l'individuo soffre di effetti collaterali") con probabilità p, che è anche la media della popolazione (in generale, la media di una Bernoulli $B(1,p)$ è p).

Quindi, considerando un campione di $n = 400$ persone,

$$X_1, \ldots, X_{400}, \quad X_i \sim B(1,p)$$

la *proporzione campionaria* $\hat{p}$, cioè la proporzione di individui che hanno avuto effetti collaterali nel campione,

$$\hat{p} = \frac{136}{400} = 0.34$$

non è altro che la media campionaria.

Siccome il campione è piuttosto grande, la varianza campionaria avrà un valore molto vicino alla varianza della popolazione, quindi quest'ultima può di fatto essere considerata nota. Inoltre, il fattore correttivo per la varianza campionaria a media incognita vale

$$\frac{n}{n-1} = \frac{400}{399} \approx 1.0025$$

cioè è molto vicino a 1, e dunque può essere trascurato: in pratica, si può calcolare la varianza campionaria come se la media della popolazione fosse nota. Così, il calcolo si riduce a quello della varianza di una Bernoulli $B(1, \hat{p})$, che vale $\hat{p}(1-\hat{p})$. Allora, riassumendo, si può affermare che

$$\sigma^2 \approx \hat{p}(1-\hat{p}) = 0.34(1-0.34) = 0.2244$$

Sempre per la dimensione del campione, $n = 400 \geq 30$, la distribuzione della media campionaria sarà approssimativamente normale, quindi un intervallo di confidenza può essere calcolato con il metodo illustrato sopra. Volendo un grado di fiducia del 95 %, si ha

$$1 - \alpha = 0.95 \implies \alpha = 0.05 \implies \frac{\alpha}{2} = 0.025$$

e, dalle tavole (usando eventualmente delle simmetrie, a seconda delle particolari tavole di cui si dispone), si ricava che $z_{\alpha/2} = z_{0.025} = 1.96$. Da qui, per ricavare l'intervallo di confidenza I è sufficiente applicare la formula:

$$\begin{aligned}
I &= \left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right) \\
&= \left(\hat{p} - z_{0.025} \cdot \sqrt{\frac{\sigma^2}{n}},\ \hat{p} + z_{0.025} \cdot \sqrt{\frac{\sigma^2}{n}} \right) \\
&= \left(0.34 - 1.96 \cdot \sqrt{\frac{0.2244}{400}},\ 0.34 + 1.96 \cdot \sqrt{\frac{0.2244}{400}} \right) \\
&\approx (0.34 - 0.05,\ 0.34 + 0.05) = (0.29, 0.39)
\end{aligned}$$

In conclusione, si ha il 95 % di probabilità che la proporzione p della popolazione che soffre di effetti collaterali sia $0.29 < p < 0.39$.

Osservazione: La variabile aleatoria che conta le occorrenze degli effetti collaterali nel campione è una binomiale $X \sim B(n = 400, p)$, che, se si suppone nota la media della popolazione, ponendo $p \approx \hat{p}$, diventa $X \sim B(n = 400, \hat{p} = 0.34)$. Allora, sono verificate le condizioni per poter approssimare tale binomiale con una distribuzione normale,

$$n\hat{p} = 400 \cdot 0.34 = 136 \geq 5 \qquad n(1-\hat{p}) = 400 \cdot 0.66 = 264 \geq 5$$

e perciò anche la proporzione campionaria $\hat{P} = \frac{X}{n}$ (della quale $\hat{p}$ è il valore per il campione dato) è approssimativamente normale. Questo è un modo alternativo (al teorema del limite centrale, con la regola $n \geq 30$) per giustificare il metodo di calcolo dell'intervallo di confidenza appena usato.

7 Problema: intervallo di confidenza per la media

Problema: Sia dato un campione di ampiezza $n = 100$, con media campionaria $\bar{x} = 21.6$, estratto da una popolazione avente deviazione standard (scarto quadratico medio) $\sigma = 5.1$. Costruire l'intervallo di confidenza al 95 % per la media μ della popolazione.

Soluzione: Al grado di fiducia del 95 %

$$1 - \alpha = 0.95 \implies \alpha = 0.05 \implies \frac{\alpha}{2} = 0.025$$

corrisponde (come sempre) il valore critico $z_{\alpha/2} = z_{0.025} = 1.96$. A questo punto, si hanno già tutti i dati necessari per compilare la formula dell'intervallo di confidenza:

$$\begin{aligned}
\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \\
21.6 - 1.96 \cdot \frac{5.1}{\sqrt{100}} < \mu < 21.6 + 1.96 \cdot \frac{5.1}{\sqrt{100}} \\
21.6 - 1.96 \cdot \frac{5.1}{10} < \mu < 21.6 + 1.96 \cdot \frac{5.1}{10} \\
21.6 - 1.96 \cdot 0.51 < \mu < 21.6 + 1.96 \cdot 0.51 \\
21.6 - 0.9996 < \mu < 21.6 + 0.9996 \\
20.6 < \mu < 22.6
\end{aligned}$$

8 Cenni al caso con varianza incognita

Il calcolo di un intervallo di confidenza per la media con il metodo presentato prima richiede la conoscenza della varianza σ^2 della popolazione. Se essa non è nota, per grandi campioni la si può sostituire con la varianza campionaria s^2. Rimane invece il problema nel caso di piccoli campioni ($n < 30$).

Se la popolazione da cui si estrae il campione ha una distribuzione normale, la statistica

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

(dove $S = \sqrt{S^2}$ è la deviazione standard campionaria) è una variabile aleatoria con una particolare distribuzione, la ***t* di Student** con $\nu = n - 1$ gradi di libertà, la cui densità ha una forma a campana, molto simile a quella della normale, ed è data dalla seguente formula (che, comunque, non è necessario conoscere per calcolare intervalli di confidenza):

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Come per la distribuzione normale, si definisce il valore critico $t_{\alpha/2}$, tale che

$$P\{T > t_{\alpha/2}\} = \frac{\alpha}{2}$$

da cui segue che

$$P\{-t_{\alpha/2} < T < t_{\alpha/2}\} = 1 - \alpha$$

e quindi, per un campione avente media campionaria $\bar{x}$ e deviazione standard campionaria s, l'intervallo di confidenza con grado di fiducia $1 - \alpha$ per la media della popolazione μ è

$$\left(\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)$$

ovvero la disuguaglianza

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

è verificata con probabilità $1 - \alpha$.

La principale differenza nel calcolo, rispetto al caso della varianza nota, è che il valore critico $t_{\alpha/2}$ dipende non solo dal grado di fiducia (come $z_{\alpha/2}$), ma anche dal grado di libertà $\nu = n - 1$, e dunque dall'ampiezza n del campione. Perciò, le tavole della distribuzione t di Student riportano il valore di $t_{\alpha/2}$ per gli $\frac{\alpha}{2}$ e ν comuni.

9 Problema: peso medio

Problema: Sia dato un campione di $n = 16$ oggetto di cui si misura il peso, trovando un peso medio (campionario) $\bar{x} = 3.42\,\mathrm{g}$ e una deviazione standard (campionaria) $s = 0.68\,\mathrm{g}$. Determinare un intervallo di confidenza con grado di fiducia del 99 % per il peso medio μ della popolazione.

Soluzione: Siccome la varianza della popolazione è incognita, e il campione non è abbastanza grande da poterla approssimare con la varianza campionaria, il calcolo dell'intervallo di confidenza non può essere effettuato in base al teorema del limite centrale.

Invece, poiché si tratta di misure, è ragionevole supporre che la popolazione da cui è stato estratto il campione sia normale, e allora si può ricorrere al calcolo con la distribuzione t di Stutent.

Volendo un grado di fiducia del 99 %, si ha

$$1 - \alpha = 0.99 \implies \alpha = 0.01 \implies \frac{\alpha}{2} = 0.005$$

e il grado di libertà è

$$\nu = n - 1 = 16 - 1 = 15$$

per cui, consultando le tavole, si legge che $t_{\alpha/2} = t_{0.005} = 2.947$. Infine, si applica la formula dell'intervallo di confidenza:

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$$3.42 - 2.947 \cdot \frac{0.68}{\sqrt{16}} < \mu < 3.42 + 2.947 \cdot \frac{0.68}{\sqrt{16}}$$

$$3.42 - 0.50 < \mu < 3.42 + 0.50$$

$$2.92\,\mathrm{g} < \mu < 3.92\,\mathrm{g}$$