STATISTICA INFERENZIALE - SCHEDA N. 1
CAMPIONAMENTO E STIMA

Molto spesso non si hanno o non si possono avere informazioni su una popolazione o su un fenomeno naturale considerati nella loro globalità, ma si conosce solo un sottoinsieme della popolazione o una parte del fenomeno, cioè si conosce un *campione*.
Nella scheda sulla "legge empirica del caso" abbiamo visto che, per valutare la probabilità di uscita di testa nel lancio di una moneta, possiamo utilizzare i risultati ottenuti ripetendo molte volte il lancio della moneta stessa. Questa procedura, pur fornendo informazioni importanti, non ci assicura che il risultato sperimentale sia il "vero" valore della probabilità cercata.
In questa scheda e nella prossima affronteremo il problema di come passare dalle informazioni relative ad un campione a considerazioni su una popolazione o su un fenomeno, valutando in termini probabilistici gli errori che si commettono. Il nostro obiettivo sarà quello di stimare i parametri di variabili casuali definite sull'intera popolazione.
Questo tipo di metodologie rientra in quella parte della statistica che viene detta *statistica inferenziale*.

ESEMPIO: Si vuole sapere se un lotto di 10 milioni di pezzi prodotti da una fabbrica soddisfa le condizioni imposte dall'acquirente al momento del contratto. Il solo modo di sapere con certezza la risposta è quello di testare ogni singolo pezzo. Questa strategia non è conveniente in termini di prezzo; inoltre, il test potrebbe essere tale da rovinare il prodotto e quindi renderlo non più vendibile. Sarebbe meglio ottenere una risposta più conveniente in termini di tempo e costi, eseguendo il test solo su *alcuni* pezzi del lotto e usando i risultati ottenuti per fare una previsione su tutti i pezzi prodotti dalla fabbrica. Non possiamo essere sicuri dell'esattezza della nostra previsione ma possiamo giustificarla in senso probabilistico se scegliamo i pezzi *secondo certe modalità*. Ma come scegliere il campione? Come stimare l'errore?

ESEMPIO: Se vogliamo conoscere il prezzo medio di un prodotto in una determinata regione geografica, ragionevolmente non possiamo risalire al prezzo in tutti i punti vendita e poi farne la media. Scegliendo *opportunamente* un numero ridotto di punti vendita, dalle risultanze di questo sottoinsieme, possiamo dedurre quanto vale all'incirca il prezzo medio. Meglio ancora possiamo definire un *intervallo* entro il quale si trova, con una certa probabilità, il prezzo medio del prodotto dell'intera area geografica.. Ma come sapere quanto il valore medio ottenuto dal campione è effettivamente vicino al valore medio reale?

1. Popolazioni e campioni

Esempi di popolazioni sono l'insieme di tutti gli abitanti di una città o di una regione, l'insieme degli studenti iscritti a un corso di laurea, un prodotto alimentare venduto in una determinata regione geografica. Talvolta però non è possibile avere i dati relativi a tutte le unità sperimentali di una popolazione (troppo costoso in termini di tempo o denaro, materialmente impossibile...)
È molto importante allora selezionare un **campione** in modo corretto, cioè in modo che sia

- **Rappresentativo** della popolazione (se, ad esempio, si vuole studiare il prezzo medio di un prodotto non si può avere un campione formato solo da supermercati, senza piccoli negozi);
- Formato da elementi fra di loro **indipendenti** (se, ad esempio, si estrae un campione da una popolazione umana per effettuare misurazioni sull'altezza non è opportuno avere un campione formato solo da elementi della stessa famiglia, in quanto l'altezza dei figli dipende, in parte, da quella dei genitori).

Esistono vari metodi per reclutare un campione considerando determinate caratteristiche della popolazione, ma queste tecniche non saranno oggetto del nostro studio, mentre ci occuperemo solo di campioni scelti casualmente con probabilità uniforme sull'intera popolazione. Una volta formato un campione si incorre comunque in un errore intrinseco al campionamento stesso, detto **errore campionario**, dato dalla differenza fra i valori ottenuti nel campione e il corrispondente parametro della variabile definita sull'intera popolazione. Vogliamo valutare questo errore sulla base di considerazioni di tipo probabilistico, utilizzando le conoscenze sulla distribuzione degli elementi del campione.

Riassumendo, se abbiamo una popolazione di numerosità ***N*** e vogliamo costruire un campione di numerosità ***n***, possiamo immaginare la popolazione come un'urna contenente biglie, con etichette che le identificano, e il campionamento come *l'estrazione con reinserimento* di n biglie. L'ipotesi della reimmissione, che a livello intuitivo non è del tutto ragionevole, è fatta per garantire l'indipendenza e per semplificare i calcoli. Comunque, quando la dimensione della popolazione è molto superiore a quella del campione, questa ipotesi incide poco sui risultati (si pensi a quanto è bassa la probabilità che intervistando 1000 abitanti di una regione si finisca per intervistare due volte lo stesso individuo).

2. Stima puntuale

Consideriamo ora una popolazione su cui è definita una variabile aleatoria X che rappresenta la caratteristica della popolazione che si vuole analizzare. La variabile X avrà una sua densità di probabilità che indichiamo con f_X.
Un campione estratto dalla popolazione è un insieme di n elementi su ciascuno dei quali si osserva la caratteristica oggetto di studio.
Indichiamo con X_1 la variabile aleatoria che corrisponde ai risultati per la prima unità campionaria e con x_1 il valore che X_1 assume nel campione considerato. La distribuzione di probabilità di X_1 sarà uguale a quella di X (ossia della variabile aleatoria che rappresenta la caratteristica della popolazione da studiare). Ad esempio, se volessimo studiare l'altezza media della popolazione adulta italiana, x_1 sarà l'altezza del primo soggetto intervistato mentre X_1 è la variabile aleatoria relativa a tutti i possibili valori assumibili con le rispettive probabilità. In maniera analoga introduciamo X_2 (variabile aleatoria) e x_2 (valore che la variabile aleatoria assume nel campione), X_3 e x_3, ..., X_n e x_n. Ricordiamo che in generale si usano le lettere maiuscole per le variabili aleatorie e con le lettere minuscole le loro realizzazioni.
Le variabili aleatorie $X_1, X_2, \ldots, X_n$ sono dette *variabili aleatorie campionarie.* Avremo, quindi, che

$$f_X = f_{X_2} = f_{X_3} = \cdots = f_{X_n}$$

e inoltre le variabili $X_1, X_2, \ldots, X_n$ sono indipendenti.

ESEMPIO: Consideriamo il lancio di una moneta. Si vuole stabilire se la moneta è truccata. La variabile X che modella l'uscita della faccia T ha distribuzione $B(1, p)$, dove p è il parametro incognito che vorremmo conoscere. Per stimare tale probabilità possiamo lanciare la moneta 10 volte e considerare la percentuale di uscite della faccia T. Gli esiti dei singoli lanci sono rappresentati dalle variabili aleatorie $X_1, X_2, \ldots, X_{10}$ ciascuna con distribuzione $B(1, p)$. È evidente infatti che in ciascun lancio non debbano essere modificate le condizioni dell'esperimento.

Occupiamoci ora di stimare il valore atteso di una variabile aleatoria, che indicheremo con μ. L'idea è quella di *scegliere opportunamente* una funzione T (**stimatore**) che dipenda solo da $X_1, X_2, \ldots, X_n$ e non esplicitamente da μ. T, essendo una funzione di variabili aleatorie, è a sua volta una variabile aleatoria con una sua distribuzione di probabilità. Una volta note le misurazioni sul campione, ossia i valori numerici $x_1, x_2, \ldots, x_n$ si utilizzano questi per avere una stima del parametro incognito mediante la funzione $T(X_1, X_2, \ldots, X_n)$. Il valore numerico t che si ottiene sostituendo i valori osservati $x_1, x_2, \ldots, x_n$ alle variabili casuali $X_1, X_2, \ldots, X_n$ nell'espressione esplicita di T, è una **stima** del parametro incognito. Si noti che la stima dipende dal campione, mentre lo stimatore è una variabile casuale indipendente dal campione prescelto.

ESEMPIO: Consideriamo un campione di 8 individui, le cui altezze sono, in cm

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
166	168	173	176	166	169	171	175

Supponiamo di voler stimare la media delle altezze X sulla base di questo campione.

Dobbiamo scegliere uno stimatore; due possibili sono:

- La variabile aleatoria media campionaria $\overline{X}_n$

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

- La variabile aleatoria che indica il valore centrale dell'intervallo dei valori assunti nel campione:

$$T = \frac{\min(X_1, \ldots, X_4) + \max(X_1, \ldots, X_4)}{2}$$

Le stime sono rispettivamente

$$\overline{x_n} = 170 \qquad t = \frac{\min(x_1, \ldots, x_4) + \max(x_1, \ldots, x_4)}{2} = 171$$

Quale stimatore scegliere? La media è lo stimatore migliore perché ha buone proprietà, come vedremo nel seguente esempio e nelle pagine seguenti.

ESEMPIO. Per capire quali siano i possibili campioni estraibili da una popolazione, quali siano i valori e le corrispondenti probabilità dello stimatore $\overline{X}_n$ media campionaria, prendiamo in esame, per semplicità, una popolazione di 4 individui A, B, C, D e consideriamo i campioni di numerosità 2. Osserviamo che, nella pratica, solo uno di essi sarà estratto.
Supponiamo che le altezze in cm di questi 4 individui siano

A	B	C	D
165	169	171	173

Possono essere considerati le realizzazioni di una variabile X. La probabilità che assegnamo a ciascun valore è $\frac{1}{4}$ perché ciascuno compare una sola volta su 4. Poiché, in questo caso, abbiamo i dati sull'intera popolazione sappiamo che il valore medio delle altezze è

$$\mu = E(X) = \frac{165 + 169 + 171 + 173}{4} = 169.5\,\text{cm}$$

Mettiamoci ora nell'ottica di chi vuole stimare questo parametro senza avere le informazioni su tutta la popolazione ma solo quelle di un campione di numerosità 2.
Una scelta istintiva è quella di stimare la media della popolazione con la media empirica calcolata sul campione. Poiché consideriamo un campionamento con ripetizione abbiamo 16 campioni di numerosità 2 tutti ugualmente probabili. Lo stimatore che utilizziamo è

$$\overline{X}_2 = \frac{X_1 + X_2}{2}.$$

Nella tabella a fianco sono riportati tutti i campioni e le corrispondenti stime della media.

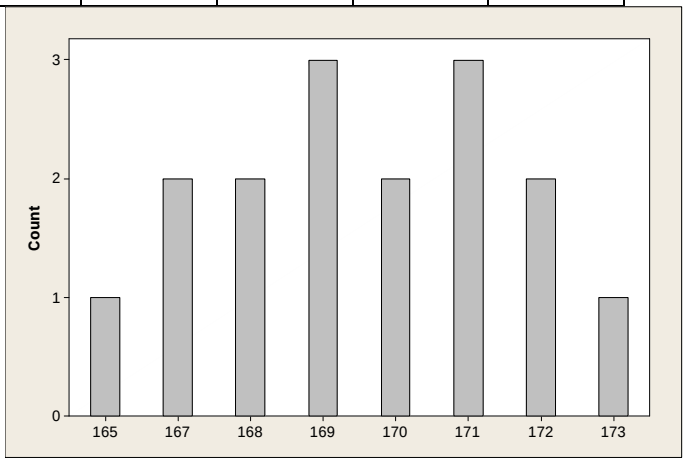
Ribadiamo il fatto che, nella situazione reale, si ha a disposizione un solo campione. Quello che stiamo facendo adesso serve per capire quali sono i possibili campioni, i possibili valori per lo stimatore $\overline{X}_n$ e le corrispondenti probabilità.

campione	x_1	x_2	$\overline{x}_n$ (in cm)
AA	165	165	165
AB	165	169	167
AC	165	171	168
AD	165	173	169
BA	169	165	167
BB	169	169	169
BC	169	171	170
BD	169	173	171
CA	171	165	168
CB	171	169	170
CC	171	171	171
CD	171	173	172
DA	173	165	169
DB	173	169	171
DC	173	171	172
DD	173	173	173

Qui sotto sono riportati i possibili valori dello stimatore $\bar{X}_2$ e le corrispondenti probabilità; dove sta la casualità? perché diciamo che $\bar{X}_2$ è una variabile aleatoria? la casualità sta nell'estrarre a caso un campione e nell'ottenere uno dei possibili valori con una determinata probabilità.

$\bar{x}_2$	165	167	168	169	170	171	172	173
$P(\bar{X}_2 = \bar{x}_2)$	1/16	2/16	2/16	3/16	2/16	3/16	2/16	1/16

Osserviamo che con nessun campione otteniamo una stima della media della popolazione uguale alla media effettiva. Stime "lontane" da 169.5 sono però in numero minore delle stime "vicine".



Il grafico della **distribuzione campionaria** di $\bar{X}_2$ è riportato qui a fianco.

Il valore atteso della variabile aleatoria $\bar{X}_2$ è quindi:

$$E(\bar{X}_2) = \frac{165 + 2\times167 + 2\times168 + 3\times169 + 2\times170 + 3\times171 + 2\times172 + 173}{16} = 169.5$$

Quindi è una variabile centrata proprio nel valore del parametro che vuole stimare.

Il fatto che $E(\bar{X}_n) = \mu$ è una proprietà generale e non dipende dai particolari valori del nostro esempio.

3. Proprietà degli stimatori

Analizziamo ora le principali proprietà che dovrebbe soddisfare un *buon* stimatore.

- *Uno stimatore T di un parametro θ è detto **non distorto** se E(T)= θ.*

Se uno stimatore è non distorto, le stime che si ottengono saranno *centrate* attorno al valore vero del parametro; in caso contrario, le stime saranno mediamente superiori o inferiori al valore vero del parametro. È, quindi, del tutto ragionevole richiedere che uno stimatore sia non distorto, anche se non sempre esistono stimatori non distorti e non sempre risultano i migliori.
La distorsione, da sola, spesso non basta. Affinché uno stimatore sia buono è importante anche che la sua varianza sia piccola.

- *Uno stimatore T è detto **consistente** se è non distorto e la sua varianza tende a zero quando la numerosità del campione tende all'infinito.*

Per confrontare due stimatori di uno stesso parametro θ si utilizza l'**errore quadratico medio**, definito da

$$MSE(T)=E((T-\theta)^2).$$

Questo indice coincide con la varianza nel caso di stimatori non distorti.

4. Alcuni stimatori

In questa sezione esamineremo nei dettagli lo stimatore della media e della varianza di una variabile aleatoria quantitativa e lo stimatore della frequenza di una variabile aleatoria qualitativa con due possibili valori.

4.1 Stimatore del valore atteso μ di una variabile aleatoria X

Consideriamo una variabile casuale X con valore atteso μ e varianza σ^2, ed un campione di numerosità n con le variabili aleatorie campionarie $X_1, X_2, \ldots, X_n$. Come stimatore del valore atteso di X si utilizza la variabile aleatoria media campionaria

$$X_n = \frac{X_1 + \cdots + X_n}{n}$$

Come abbiamo già visto nella scheda precedente

$$E(\bar{X}_n)=\mu \quad \text{e} \quad Var(\bar{X}_n)=\sigma^2/n$$

Quindi $\overline{X}$ è uno stimatore per la media μ non distorto e consistente.

ESEMPIO: Consideriamo un campione di numerosità 5 estratto da una popolazione su cui è definita una variabile normale X e le variabili aleatorie campionarie X_1, X_2, X_3, X_4, X_5. Si vuole stimare il valore atteso di X. Si utilizza allora lo stimatore media campionaria

$$\bar{X}_5 = \frac{X_1 + \cdots + X_5}{5}$$

Se i valori osservati nel campione sono

22 26 30 21 27

La stima che si ottiene è 25.2. Se dalla stessa popolazione si estrae un altro campione di numerosità 5 i cui valori osservati sono

32 27 25 29 20

Lo stimatore è lo stesso di prima, ma la stima in questo caso è 26.6. Nonostante lo stimatore sia identico nei due casi, la stima varia a seconda del particolare campione che si estrae dalla popolazione.

4.2 Stimatore della frequenza p di una variabile di Bernoulli

Se si vuole stimare la frequenza p di una caratteristica presente in una popolazione, si può effettuare un campionamento della popolazione e utilizzare come stimatore di p la funzione $\hat{P}$ ottenuta calcolando la frequenza (relativa) di successi nel campione. La variabile aleatoria $\hat{P}$ sarà

$$\hat{P} = \frac{X_1 + \cdots + X_n}{n}$$

dove $X_1, X_2, \ldots, X_n$ sono variabili aleatorie campionarie, ciascuna X_i è una variabile aleatoria con distribuzione $B(1, p)$. Avremo quindi

$$E(\hat{P}) = E\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n}E\left(X_1 + \ldots + X_n\right) = \frac{1}{n}np = p$$

$$Var(\hat{P}) = \frac{1}{n^2}Var(\mathrm{X}_1 + \ldots + \mathrm{X}_n) = \frac{p(1-p)}{n}$$

Anche in questo caso $\hat{P}$ è uno stimatore non distorto e consistente per p. Questi risultati derivano dal caso precedente; infatti $\hat{P}$ è una particolare media campionaria.
Osserviamo che, a parità di n, la varianza è massima quando p=1/2.

ESEMPIO: Vogliamo stimare la proporzione di maschi in una popolazione sulla base di un campione di numerosità 200. Lo stimatore è $\hat{P}$. Se sul campione di 200 persone si osservano 96 maschi e 104 femmine, la stima della frequenza relativa è

$$\hat{p} = 96/200 = 0.48$$

Nella prossima scheda vedremo come costruire un intervallo (sempre a partire dai dati campionari) a cui il parametro incognito appartenga con un certo grado di fiducia.

4.3 Stimatore della varianza σ^2 di una variabile aleatoria

Consideriamo una variabile casuale X con valore atteso μ e varianza σ^2, ed un campione di numerosità n con le variabili aleatorie campionarie $X_1, X_2, \ldots, X_n$. In analogia a quanto fatto con il valore atteso si può pensare di utilizzare come stimatore di σ^2 la variabile aleatoria $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, ma si dimostra che questo stimatore è distorto, cioè il suo valore atteso non è σ^2.

Uno stimatore non distorto di σ^2 è la variabile aleatoria che viene indicata con S^2:

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

Come abbiamo già visto nelle schede di statistica descrittiva, nei calcoli può essere più comodo sviluppare il quadrato, ottenendo:

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

ESEMPIO. In un'indagine campionaria sul prezzo di un litro di latte intero rilevato 57 negozi si è ottenuto che prezzo medio campionario è $\bar{x}$ =1.34 euro; la somma dei quadrati dei 57 valori di un litro di latte è 105.857. Quindi la stima della varianza del prezzo di un litro di latte è:

$$s^2 = \frac{1}{56}105.857 - \frac{57}{56}1.34^2 = 0.0626$$

ESERCIZI

1) Sia X una variabile aleatoria la cui funzione di densità dipende da un parametro θ.
Siano D_1 e D_2 due stimatori indipendenti del parametro entrambi non distorti e di varianza rispettivamente σ_1^2 e σ_2^2. Si consideri lo stimatore D:

$$D = \lambda D_1 + (1-\lambda) D_2 \qquad \text{con } \lambda \in (0,1)$$

a) Dire se D è distorto.
b) Calcolare la varianza di D.
c) Calcolare il valore di λ per cui la varianza di D è minima.

2) Sia X una variabile aleatoria con densità dipendente da un parametro reale θ. Siano S_1 e S_2 due stimatori del parametro basati sulle variabili aleatorie campionarie $X_1,\ldots,X_n$ tali che: $E(S_1) = \theta$, $E(S_2) = \theta - \frac{1}{n}$, $\text{Var}(S_1) = \frac{\theta}{n}$, $\text{Var}(S_2) = \frac{\theta}{n^2}$.
Calcolare l'errore quadratico medio (MSE).dei due stimatori e dire quale è preferibile.

3) Sia X una variabile aleatoria e siano $X_1,\ldots,X_n$ variabili aleatorie campionarie e $x_1,\ldots,x_n$ le loro realizzazioni; scrivere la differenza fra stimatore della media e la sua stima.

4) Si considerino le variabili casuali campionarie $X_1,\ldots,X_n$ di una variabile casuale con la seguente distribuzione discreta:

x	0	1	2	3	4
P(x)	θ	2θ	3θ	4θ	1-10θ

a) Determinare i valori assumibili dal parametro θ affinché P(x) sia una distribuzione di probabilità.
b) Scrivere uno stimatore T_n per θ.
c) Calcolare il valore atteso di T_n e stabilire se è distorto.
d) Calcolare la varianza di T_n e il suo errore quadratico medio (MSE).

5) Spiega, eventualmente anche con un esempio, che cosa sono lo stimatore della media di una variabile aleatoria, il valor medio dello stimatore e la stima.