STIMATORI E INTERVALLI DI CONFIDENZA

VARIANZA NOTA E IGNOTA

Dr. R. Lo Franco – Lezione 9-10

Riferimenti: [1] Sheldon M. Ross, *Introduzione alla statistica*, Apogeo Editore; [2] Maria Garetto, *Statistica*, Università di Torino

7. Stima dei parametri

7.1 Introduzione

Abbiamo visto come la teoria dei campioni possa essere usata per ottenere informazioni riguardanti campioni estratti casualmente da una popolazione.
Da un punto di vista applicativo è però spesso più importante trarre conclusioni sull'intera popolazione utilizzando i risultati ottenuti su campioni estratti da essa. Questi sono i problemi di cui si occupa l'**inferenza statistica**.
I metodi della statistica inferenziale riguardano essenzialmente due aree: la **stima dei parametri** e i **test di ipotesi**.
Il primo importante problema dell'inferenza statistica, di cui ci occupiamo in questo capitolo, è la stima dei parametri di una popolazione, media, varianza, scarto quadratico medio, per mezzo dei corrispondenti parametri campionari o statistiche del campione.
Il valore del parametro da stimare per la popolazione è incognito, e possiamo solo chiederci se, dopo ripetuti campionamenti, la distribuzione della statistica ha certe proprietà che possono garantirci che la statistica è vicina al valore incognito del parametro.
Ad esempio, sappiamo dal teorema 1, Cap. 6, che la distribuzione della media campionaria ha la stessa media della popolazione da cui è stato ottenuto il campione: ci aspettiamo perciò che, dopo più campionamenti, la media campionaria sia vicina alla media della popolazione.

7.2 Stime puntuali e stime per intervallo

Per i parametri di una popolazione è possibile calcolare due tipi di stima: una stima puntuale e una stima per intervallo.

Definizioni 1
Se la stima di un parametro della popolazione è data da un singolo numero, tale valore è detto **stima puntuale** del parametro.
Se invece la stima di un parametro della popolazione fornisce gli estremi di un intervallo fra i quali si può supporre, con un certo grado di fiducia, che il parametro sia compreso, tale stima è detta **stima per intervallo** del parametro.

I parametri che più frequentemente accade di dover stimare sono:
1 – la media μ di una popolazione;
2 – la varianza σ^2 di una popolazione;
3 – la proporzione p di individui di una popolazione che appartengono a una certa classe di interesse;
4 – la differenza fra le medie di due popolazioni $\mu_1 - \mu_2$;
5 – la differenza fra le proporzioni di due popolazioni $p_1 - p_2$.
Ragionevoli stime puntuali di questi parametri sono:
1 – per μ, la media campionaria $\bar{x}$;
2 – per σ^2, la varianza campionaria s^2;
3 – per p, la proporzione campionaria $\hat{p} = \frac{x}{n}$, dove x è il numero di individui in un campione di ampiezza n appartenenti alla classe di interesse;
4 – per $\mu_1 - \mu_2$, la differenza $\bar{x}_1 - \bar{x}_2$ fra le medie di due campioni indipendenti;
5 – per $p_1 - p_2$, la differenza $\hat{p}_1 - \hat{p}_2$ fra le proporzioni di due campioni indipendenti.

Si possono avere più stime puntuali per lo stesso parametro; per esempio se si vuole stimare la media di una popolazione, si potrebbe usare anche la mediana campionaria, o magari la media fra il più piccolo e il più grande fra i valori del campione[1].
Per decidere quale fra le possibili stime puntuali è preferibile usare, ci basiamo sulla verifica di alcune proprietà che gli stimatori devono possedere per essere giudicati i più adatti.
Una di queste è la proprietà della **correttezza** o **non distorsione**.

Definizione 2
Se la media di una distribuzione campionaria di una statistica è uguale al corrispondente parametro della popolazione, la statistica è detta **stimatore corretto** o **non distorto** del parametro.

I valori corrispondenti a tali statistiche sono detti **stime corrette**. In altre parole, una statistica è uno stimatore corretto se "in media" i suoi valori uguagliano il parametro che valuta.
Ad esempio la media della distribuzione campionaria della media è

$$\mu_{\overline{X}} = \mu$$

quindi la media campionaria $\overline{x}$ è una stima corretta della media μ di una popolazione.
Lo stimatore corretto di un parametro non è unico. Ad esempio anche la mediana campionaria è una stima corretta della media della popolazione.
Occorre quindi un'ulteriore proprietà, detta **efficienza**, per decidere quale tra più stime corrette sia la migliore per stimare un parametro.

Definizione 3
Se due statistiche sono entrambe stimatori corretti di un parametro, la statistica per cui la varianza della sua distribuzione campionaria è minore è detta **stimatore più efficiente**.

Si può dimostrare che, fra tutte le statistiche che stimano la media di una popolazione, la media campionaria è la più efficiente. Anche la varianza campionaria è una stima corretta ed efficiente della varianza di una popolazione.

Esempio 1

Dato un campione di 5 misurazioni del diametro di una sferetta in cm

6.33 6.37 6.36 6.32 6.37

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

La stima corretta ed efficiente per la media della popolazione è la media campionaria

$$\bar{x} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35\,\text{cm}$$

Anche per la varianza la stima corretta ed efficiente è la varianza campionaria

$$s^2 = \frac{1}{4}\sum_{i=1}^{5}\left(x_i - \bar{x}\right)^2 = \frac{(0.02)^2 + (0.02)^2 + (0.01)^2 + (0.03)^2 + (0.02)^2}{4} = 0.00055\,\text{cm}^2$$

Poiché non ci si può aspettare che una stima puntuale coincida esattamente con la quantità che essa deve stimare, è spesso preferibile usare una stima per intervallo, ossia un intervallo per il quale si può affermare con un certo grado di fiducia che conterrà il parametro della popolazione che si vuole stimare.

Tali stime per intervallo vengono comunemente chiamate **intervalli di confidenza**.

7.3 Intervalli di confidenza per la media (varianza nota)

Come già detto, la media campionaria $\bar{x}$ è una buona stima, corretta ed efficiente, della media μ di una popolazione. Tuttavia, non c'è alcuna probabilità che la stima sia esattamente uguale a μ; ha quindi più significato stimare μ con un intervallo, che in qualche modo ci dà informazioni sulla probabile grandezza di μ.
Per ottenere una stima per intervallo, si utilizzano le proprietà delle distribuzioni campionarie. In questo caso, poiché si vuole stimare la media di una popolazione per mezzo della media di un campione, facciamo ricorso alla distribuzione della media campionaria.

In base al teorema del limite centrale possiamo affermare che, per grandi valori di n, la statistica

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (7.1)$$

ha approssimativamente la distribuzione normale standardizzata.

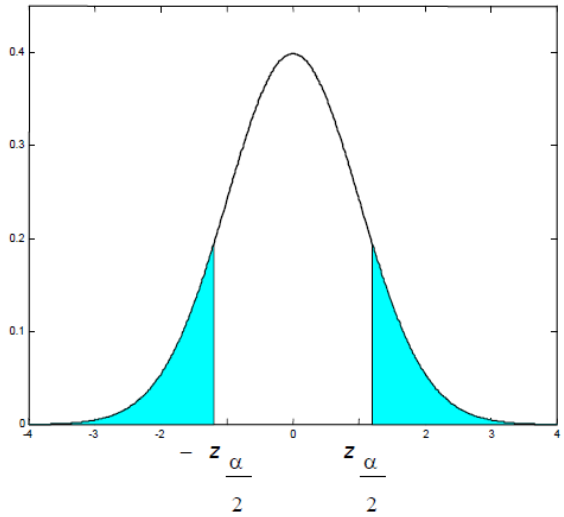


Figura 1

Se l'area sottesa dalla distribuzione normale a destra di $z_{\frac{\alpha}{2}}$ vale $\frac{\alpha}{2}$ (figura 1), allora l'area compresa fra $-z_{\frac{\alpha}{2}}$ e $z_{\frac{\alpha}{2}}$ vale $1-\alpha$, perciò

$$P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Di conseguenza, si può asserire, con probabilità uguale a $1-\alpha$, che è soddisfatta la disuguaglianza

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

Pertanto, una volta estratto il campione di ampiezza n, con n sufficientemente grande (**$n \geq 30$, grande campione**) e calcolato il valore $\bar{x}$ della media del campione, si ottiene la seguente stima per intervallo per la media μ, soddisfatta con probabilità $1 - \alpha$.

$$\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \qquad (7.3)$$

Si può quindi affermare con probabilità $1 - \alpha$ che l'intervallo $\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$ contiene la media μ della popolazione.

L'intervallo (7.3) è detto anche **intervallo di confidenza per la media μ**, per **grandi campioni**, con **grado di fiducia** $(1-\alpha)\cdot 100\%$.

grado di fiducia del 90%	$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$
grado di fiducia del 95%	$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$
grado di fiducia del 99%	$z_{\frac{\alpha}{2}} = z_{0.005} = 2.576$

La lunghezza di un intervallo di confidenza con grado di fiducia $(1-\alpha)\cdot 100\%$ è

$$2z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

e dipende quindi da tre fattori

– n: al crescere dell'ampiezza del campione, la lunghezza dell'intervallo diminuisce, quindi la stima è più precisa;

– α: al crescere del grado di fiducia richiesto, la lunghezza dell'intervallo aumenta, quindi la stima è meno precisa;

– σ: al crescere della deviazione standard, che riflette la variabilità del campione, la lunghezza dell'intervallo aumenta.

Indicando il massimo dell'errore con

$$E = \max \left| \overline{X} - \mu \right|$$

la stima di E con probabilità $1 - \alpha$ è

$$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \qquad (7.4)$$

In altre parole, se si vuole stimare la media μ della popolazione con la media campionaria di un campione di ampiezza n ($n \geq 30$), si può affermare, con probabilità $1 - \alpha$, che l'errore $\left| \overline{X} - \mu \right|$ sarà al più uguale a $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

Dalla formula (7.4), risolvendo rispetto a n, si ricava l'ampiezza del campione necessaria per stimare la media con un errore prefissato E e con un dato grado di fiducia (si ricordi che n deve essere un intero)

$$n \geq \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E} \right)^2 \qquad (7.5)$$

Se si è interessati alla stima della media della popolazione che sia più grande o più piccola di un certo estremo inferiore o superiore, si trovano gli **estremi di confidenza.**

Esempio 2

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

Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Applicando la formula (7.3) si ottiene l'intervallo di confidenza

$$21.6 - 1.96 \cdot \frac{5.1}{\sqrt{100}} < \mu < 21.6 + 1.96 \cdot \frac{5.1}{\sqrt{100}}$$

$$20.6 < \mu < 22.6$$

Questo intervallo può anche non contenere μ, ma abbiamo un grado di fiducia del 95% che lo contenga. In altre parole, se applichiamo ripetutamente su tutti i campioni di ampiezza $n = 100$ estraibili dalla popolazione la formula (7.3) per calcolare l'intervallo di confidenza, il 95% degli intervalli di confidenza conterrà la media μ della popolazione.

Esempio 4

Le misure dei diametri di un campione casuale di 200 sferette da cuscinetto prodotte da una macchina in una settimana hanno una media campionaria $\bar{x} = 0.824$ cm e una deviazione standard campionaria $s = 0.042$ cm.

Determinare gli intervalli di confidenza per la media della popolazione con grado di fiducia del 95% e del 99%.

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$0.824 - 1.96 \cdot \frac{0.042}{\sqrt{200}} < \mu < 0.824 + 1.96 \cdot \frac{0.042}{\sqrt{200}}$$

$$0.818 < \mu < 0.830$$

b –Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$0.824 - 2.576 \cdot \frac{0.042}{\sqrt{200}} < \mu < 0.824 + 2.576 \cdot \frac{0.042}{\sqrt{200}}$$

$$0.816 < \mu < 0.832$$

Si osservi che aumentando il grado di fiducia l'ampiezza dell'intervallo aumenta, ossia a parità di numero di elementi del campione la stima è meno precisa.

Esempio 5

Si vuole stimare il numero medio di battiti cardiaci al minuto per una certa popolazione. Il numero medio di battiti al minuto per un campione di 49 soggetti è risultato uguale a 90. La popolazione è distribuita in modo normale con uno scarto quadratico medio $\sigma = 10$.
Trovare gli intervalli di confidenza per la media della popolazione con i gradi di fiducia del 90%, 95% e 99%.

a – Per il grado di fiducia del 90% il valore critico è $z_{\frac{\alpha}{2}} = 1.645$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$90 - 1.645 \cdot \frac{10}{\sqrt{49}} < \mu < 90 + 1.645 \cdot \frac{10}{\sqrt{49}}$$

$$87.65 < \mu < 92.35$$

b – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$90 - 1.96 \cdot \frac{10}{\sqrt{49}} < \mu < 90 + 1.96 \cdot \frac{10}{\sqrt{49}}$$

$$87.20 < \mu < 92.80$$

c – Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$

Esempio 6

Sia dato un campione di 100 studenti tratto da una popolazione di studenti di sesso maschile iscritti ad un'università; la tabella 1 rappresenta la distribuzione di frequenza dei pesi in kg degli studenti. Trovare gli intervalli di confidenza al 95% e al 99% per il peso medio di tutti gli studenti.

Classi (peso)	*N° studenti (freq. ass.)*	*Valori centrali*
$60 \le x \le 62$	5	61
$63 \le x \le 65$	18	64
$66 \le x \le 68$	42	67
$69 \le x \le 71$	27	70
$72 \le x \le 74$	8	73

Tabella 1

Calcoliamo la media e la varianza campionarie usando i dati raggruppati

$$\bar{x} = \frac{5 \cdot 61 + 18 \cdot 64 + 42 \cdot 67 + 27 \cdot 70 + 8 \cdot 73}{100} = 67.45$$

$$s^2 = \frac{1}{99}\left[5 \cdot 61^2 + 18 \cdot 64^2 + 42 \cdot 67^2 + 27 \cdot 70^2 + 8 \cdot 73^2 - 100 \cdot 67.45^2\right] = 8.61$$

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$67.45 - 1.96 \cdot \frac{\sqrt{8.61}}{\sqrt{100}} < \mu < 67.45 + 1.96 \cdot \frac{\sqrt{8.61}}{\sqrt{100}}$$

$$66.87 < \mu < 68.02$$

b – Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$

Con la formula (7.3) si ottiene l'intervallo di confidenza

$$67.45 - 2.576 \cdot \frac{\sqrt{8.61}}{\sqrt{100}} < \mu < 67.45 + 2.576 \cdot \frac{\sqrt{8.61}}{\sqrt{100}}$$

$$66.69 < \mu < 68.21$$

Esempio 8

Un medico misura i tempi di reazione dei suoi pazienti a un determinato stimolo. La stima dello scarto quadratico medio è $s = 0.05$ sec.

Calcolare quanto deve essere grande il campione di misurazioni affinché si possa asserire con grado di fiducia del 95% e del 99%, che l'errore nello stimare il tempo medio di reazione nella popolazione non è superiore a 0.01 sec.

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$.

Con la formula (7.5) si ottiene

$$n \geq \left(\frac{1.96 \cdot 0.05}{0.01} \right)^2 = 96.04$$

Quindi possiamo avere un grado di fiducia del 95% che l'errore nella stima del tempo medio sarà al più 0.01 sec, se prendiamo un campione di ampiezza $n = 97$.

b – Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$. Con la formula (7.5) si ottiene

$$n \geq \left(\frac{2.576 \cdot 0.05}{0.01} \right)^2 = 165.9$$

Quindi il campione deve avere ampiezza $n = 166$.

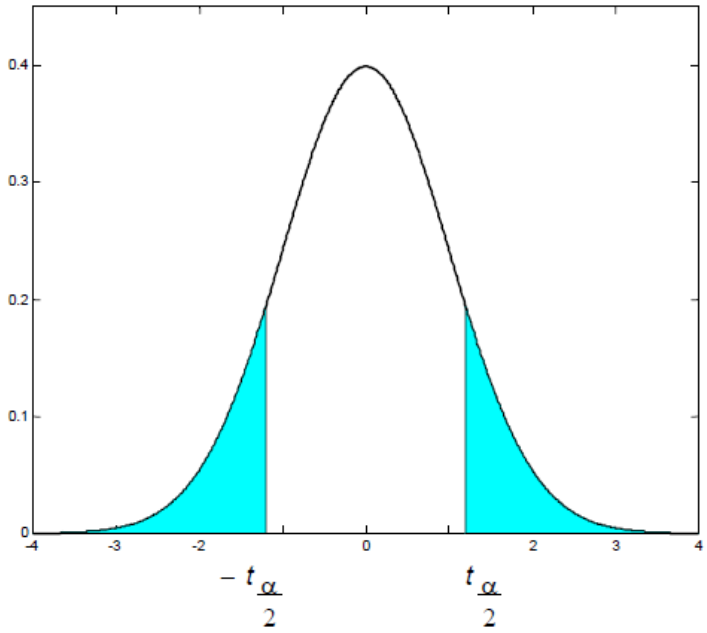
Si osservi (esempi 7 e 8) che per avere un maggior grado di fiducia, a parità di errore, bisogna usare un campione di ampiezza più grande.

7.4 Intervalli di confidenza per la media (varianza incognita)

L'applicazione della (7.3) richiede la conoscenza di σ; se σ non è noto, si è già osservato che per grandi campioni può essere sostituito con lo scarto quadratico medio campionario s.
Per **piccoli campioni** ($n < 30$), nell'ipotesi che la popolazione da cui si estrae il campione abbia distribuzione normale, ci si può servire del teorema 3, Cap. 6, in base al quale la statistica

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{7.6}$$

è una variabile aleatoria che ha la distribuzione t di Student con grado di libertà $\nu = n - 1$.



Procedendo come nel caso dei grandi campioni, se l'area sottesa dalla distribuzione t a destra di $t_{\frac{\alpha}{2}}$ vale $\frac{\alpha}{2}$ (figura 2), allora l'area compresa fra $-t_{\frac{\alpha}{2}}$ e $t_{\frac{\alpha}{2}}$ vale $1-\alpha$, perciò

$$P\left(-t_{\frac{\alpha}{2}} < T < t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

In altre parole si può asserire, con probabilità uguale a $1-\alpha$, che è soddisfatta la disuguaglianza

$$-t_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{\frac{\alpha}{2}} \qquad (7.7)$$

Pertanto, una volta estratto il campione di ampiezza n, con $n < 30$, e calcolati i valori della media $\bar{x}$ e dello scarto quadratico medio s del campione, si ottiene la stima per intervallo per la media μ, con probabilità $1-\alpha$, o con grado di fiducia $(1-\alpha)\cdot 100\%$

$$\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \qquad (7.8)$$

L'intervallo (7.8) è detto **intervallo di confidenza per la media μ**, per **piccoli campioni**, con grado di fiducia $(1-\alpha)\cdot 100\%$.

Si ricordi che il grado di libertà della distribuzione t è $\nu = n - 1$.

I valori più comunemente usati per $1 - \alpha$ sono 0.90, 0.95 e 0.99 ; i relativi gradi di fiducia sono il 90%, il 95% e il 99%; i corrispondenti valori di $t_{\frac{\alpha}{2}}$ sono

grado di fiducia del 90% $\quad t_{\frac{\alpha}{2}} = t_{0.05}$

grado di fiducia del 95% $\quad t_{\frac{\alpha}{2}} = t_{0.025}$

grado di fiducia del 99% $\quad t_{\frac{\alpha}{2}} = t_{0.005}$

Questi valori possono essere letti sulla tabella della distribuzione t in corrispondenza al grado di libertà $\nu = n - 1$.

$$P(t > t_{\alpha,n-1}) = \alpha$$

N.B. Nella tabella n ≡ v, indica il numero gradi di libertà della t di Student, da non confondere con la numerosità del campione.

Campione di numerosità n
ha v = n-1
e dalla tabella bisogna prendere il valore corrispondente a *n* = n-1

Tabella A.3 Valori assunti da $t_{\alpha,n}$

	α				
n	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
70	1.294	1.667	1.994	2.381	2.648
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.326	2.576

Esempio 9

Sia dato un campione di 16 oggetti di cui si misura il peso, trovando un peso medio $\bar{x} = 3.42$ g e uno scarto quadratico medio $s = 0.68$ g.
Determinare un intervallo di confidenza con grado di fiducia del 99% per il peso medio della popolazione.

Poiché si tratta di misure, si può ragionevolmente ipotizzare che la popolazione da cui proviene il campione abbia distribuzione normale.
Il campione ha ampiezza $n = 16$, perciò il grado di libertà è

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

Dalle tavole della distribuzione t si ottiene

$$t_{0.005} = 2.947.$$

Con la formula (7.8) si ottiene l'intervallo di confidenza

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

$$2.91 < \mu < 3.93$$

Esempio 10

Un campione di 10 misurazioni del diametro di una sferetta ha una media campionaria $\bar{x} = 4.38$ cm e una deviazione standard campionaria $s = 0.06$ cm. Determinare gli intervalli di confidenza con grado di fiducia del 90%, 95% e 99% per il diametro medio della popolazione.

Poiché si tratta di misure, si può ragionevolmente ipotizzare che la popolazione da cui proviene il campione abbia distribuzione normale.
Il campione ha ampiezza $n = 10$, perciò il grado di libertà è $\nu = n - 1 = 9$.

a – Per il grado di fiducia del 90% e il grado di libertà $\nu = 9$, si ha $t_{\frac{\alpha}{2}} = t_{0.05} = 1.833$

Con la formula (7.8) si ottiene l'intervallo di confidenza

$$4.38 - 1.833 \cdot \frac{0.06}{\sqrt{10}} < \mu < 4.38 + 1.833 \cdot \frac{0.06}{\sqrt{10}}$$

$$4.34 < \mu < 4.42$$

b – Per il grado di fiducia del 95% e il grado di libertà $\nu = 9$, si ha $t_{\frac{\alpha}{2}} = t_{0.025} = 2.262$

$$4.38 - 2.262 \cdot \frac{0.06}{\sqrt{10}} < \mu < 4.38 + 2.262 \cdot \frac{0.06}{\sqrt{10}}$$

$$4.33 < \mu < 4.43$$

c – Per il grado di fiducia del 99% e il grado di libertà $\nu = 9$, si ha $t_{\frac{\alpha}{2}} = t_{0.005} = 3.250$

$$4.38 - 3.250 \cdot \frac{0.06}{\sqrt{10}} < \mu < 4.38 + 3.250 \cdot \frac{0.06}{\sqrt{10}}$$

$$4.31 < \mu < 4.45$$

Si osservi come, restando invariata l'ampiezza del campione, all'aumentare del grado di fiducia cresce l'ampiezza dell'intervallo di confidenza, ossia la stima è meno precisa.

Esempio 11

Le misure in kg del peso di un campione di 10 studenti maschi del primo anno di un'università sono

60	63	60	68	70	72	65	61	69	67

Trovare un intervallo di confidenza con grado di fiducia del 99% per il peso medio della popolazione universitaria maschile del primo anno di quella università.

Calcoliamo la media e la varianza campionaria

$$\bar{x} = \frac{60+63+60+68+70+72+65+61+69+67}{10} = 65.5$$

$$s^2 = \frac{1}{9}\cdot\left[60^2+63^2+60^2+68^2+70^2+72^2+\right.$$

$$\left.+65^2+61^2+69^2+67^2-10\cdot 65.5^2\right] = 18.94$$

Il campione ha ampiezza $n = 10$, perciò il grado di libertà è $\nu = n - 1 = 9$.

Per il grado di fiducia del 99% e il grado di libertà $\nu = 9$, si ha $t_{\frac{\alpha}{2}} = t_{0.005} = 3.250$

$$65.5 - 3.250\cdot\frac{\sqrt{18.94}}{\sqrt{10}} < \mu < 65.5 + 3.250\cdot\frac{\sqrt{18.94}}{\sqrt{10}}$$

$$61.02 < \mu < 69.98$$

7.5 Intervalli di confidenza per la proporzione

Un caso particolarmente importante di stima della media per una popolazione non normale e per grandi campioni è quello di una popolazione bernoulliana. Si vuole stimare il valore del parametro p (probabilità di successo), che rappresenta la frequenza relativa o proporzione con cui una certa caratteristica si presenta negli individui di una data popolazione.
Esempi tipici di questa situazione sono i seguenti.
1 – Il sondaggio di opinione: si vuole stimare la proporzione p della popolazione complessiva che è d'accordo con una certa opinione, osservando il valore che questa proporzione ha su un campione di n individui.
2 – La produzione di un dato tipo di oggetto: il produttore vuole poter garantire che la proporzione di pezzi difettosi in una data produzione non superi un certo valore prefissato; occorre quindi determinare, esaminando un campione, un intervallo di confidenza per la proporzione p di pezzi difettosi in una produzione, ed eventualmente intervenire sulla produzione affinché la proporzione di pezzi difettosi non superi una certa soglia fissata.

3 – Lo studio della diffusione di una data malattia: si vuole stimare qual è la proporzione di pazienti di una certa popolazione che ha una data malattia, studiando il valore di questa proporzione su un campione di n persone appartenenti a quella popolazione.
Per stimare la proporzione di una popolazione procediamo nello stesso modo in cui abbiamo stimato la media di una popolazione.

Si estraggono campioni di ampiezza n dalla popolazione e si considera la proporzione campionaria $\hat{P} = \frac{X}{n}$, dove X è il numero di volte in cui la caratteristica osservata si presenta nel campione. Questa proporzione campionaria è uno stimatore corretto della proporzione p della popolazione e viene usato come stima puntuale.

Nel § 5. 5 abbiamo visto che, quando si ha sia $np \geq 5$ che $n(1-p) \geq 5$, la distribuzione binomiale di parametri n e p può essere approssimata da una distribuzione normale avente media $\mu = np$ e varianza $\sigma^2 = np(1-p)$. In altri termini la statistica

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{7.9}$$

ha approssimativamente la distribuzione normale standardizzata per grandi valori di n.

Quindi quando n è grande si può costruire un intervallo di confidenza per il parametro p, usando l'approssimazione normale per la distribuzione binomiale. Possiamo affermare che

$$P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

ossia, con probabilità $1 - \alpha$, vale la disuguaglianza

$$-z_{\frac{\alpha}{2}} < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}} \tag{7.10}$$

In questo modo, estraendo un campione di ampiezza n da una popolazione bernoulliana e indicando con $\hat{p}$ la proporzione del campione, si ottiene il seguente **intervallo di confidenza per la proporzione *p*** della popolazione bernoulliana, con grado di fiducia $(1-\alpha)\cdot 100\%$, valido per grandi campioni.

$$\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (7.11)$$

Il valore critico $z_{\frac{\alpha}{2}}$ viene scelto con la stessa regola già indicata per l'intervallo di confidenza per la media, nel caso dei grandi campioni.

Osserviamo che per ottenere l'intervallo di confidenza (7.11) sono state fatte tre approssimazioni:

1 – l'approssimazione normale della binomiale;

2 – l'approssimazione di p con $\hat{p} = \frac{x}{n}$, nell'espressione $\sqrt{\frac{p(1-p)}{n}}$;

3 – non è stata fatta la correzione di continuità per l'approssimazione normale[2].

Questo implica che l'intervallo di confidenza trovato è un intervallo approssimato.

Per verificare le condizioni di applicabilità dell'approssimazione della binomiale con la normale, ossia $np \geq 5$ e $n(1-p) \geq 5$, possiamo solo verificare che sia $n\hat{p} \geq 5$ e $n(1-\hat{p}) \geq 5$; questa verifica si può fare solo dopo aver effettuato il campionamento: se le condizioni precedenti non sono soddisfatte, il risultato è privo di valore, e occorre ripetere il campionamento aumentando l'ampiezza n del campione.

Esempio 13

In un campione di 400 persone a cui 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.

Nel campione di $n = 400$ persone la proporzione campionaria è

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

Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$ e con la formula (7.11) si trova l'intervallo di confidenza

$$0.34 - 1.96 \cdot \sqrt{\frac{0.34 \cdot (1 - 0.34)}{400}} < p < 0.34 + 1.96 \cdot \sqrt{\frac{0.34 \cdot (1 - 0.34)}{400}}$$

$$0.29 < p < 0.39$$

Osserviamo che le condizioni per poter usare l'approssimazione della binomiale con la normale sono verificate, essendo

$$n\hat{p} = 400 \cdot 0.34 = 135 \qquad \text{e} \qquad n(1 - \hat{p}) = 400 \cdot 0.66 = 264 .$$

Esempio 14

Un campione di 100 votanti scelto a caso fra tutti i votanti di una regione ha indicato che il 55% di essi è favorevole ad un certo candidato.

a – Determinare gli intervalli di confidenza con grado di fiducia del 95% e del 99% per la proporzione di tutti i votanti a favore del candidato.

b – Confrontare queste stime con la stima che si trova se si usa un campione di 2000 votanti, con la stessa percentuale campionaria di favorevoli.

a – Per il grado di fiducia del 95% il valore critico è $z_{\frac{\alpha}{2}} = 1.96$; il risultato campionario indica che $\hat{p} = 0.55$ si e con la formula (7.11) si trova l'intervallo di confidenza

$$0.55 - 1.96 \cdot \sqrt{\frac{0.55 \cdot (1 - 0.55)}{100}} < p < 0.55 + 1.96 \cdot \sqrt{\frac{0.55 \cdot (1 - 0.55)}{100}}$$

$$0.45 < p < 0.65$$

Possiamo asserire con grado di fiducia del 95% che il candidato avrà a suo favore una percentuale di votanti compresa fra il 45% e il 65%.

Per il grado di fiducia del 99% il valore critico è $z_{\frac{\alpha}{2}} = 2.576$ e con la formula (7.11) si trova l'intervallo di confidenza

$$0.55 - 2.576 \cdot \sqrt{\frac{0.55 \cdot (1 - 0.55)}{100}} < p < 0.55 + 2.576 \cdot \sqrt{\frac{0.55 \cdot (1 - 0.55)}{100}}$$

$$0.42 < p < 0.68$$

Possiamo in questo caso asserire con grado di fiducia del 99% che il candidato avrà a suo favore una percentuale di votanti compresa fra il 42% e il 69%.

L'ampiezza degli intervalli di confidenza trovati è troppo grande, ossia la precisione delle stime è troppo bassa.

Esempio 14

Un campione di 100 votanti scelto a caso fra tutti i votanti di una regione ha indicato che il 55% di essi è favorevole ad un certo candidato.

a – Determinare gli intervalli di confidenza con grado di fiducia del 95% e del 99% per la proporzione di tutti i votanti a favore del candidato.

b – Confrontare queste stime con la stima che si trova se si usa un campione di 2000 votanti, con la stessa percentuale campionaria di favorevoli.

b – Se il campione è di 2000 votanti, con il grado di fiducia del 95% si trova il seguente intervallo di confidenza

$$0.55 - 1.96 \cdot \sqrt{\frac{0.55 \cdot (1 - 0.55)}{2000}} < p < 0.55 + 1.96 \cdot \sqrt{\frac{0.55 \cdot (1 - 0.55)}{2000}}$$

$$0.52 < p < 0.58$$

In questo caso, con un grado di fiducia del 95%, il candidato avrà a suo favore una percentuale di votanti compresa fra il 52% e il 58%, con una stima decisamente più precisa. La maggior precisione dipende dalla maggiore ampiezza del campione.

Indicando con

$$E = \max \left| \hat{P} - p \right|$$

il massimo dell'errore che si commette approssimando la proporzione della popolazione p con la proporzione campionaria $\hat{P} = \frac{X}{n}$, la stima di E con probabilità $1 - \alpha$ è data da

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \qquad (7.12)$$

In altre parole, se si vuole stimare la proporzione p della popolazione con la proporzione campionaria $\hat{p} = \frac{x}{n}$ di un campione di ampiezza n ($n \geq 30$), si può affermare, con probabilità $1-\alpha$, che l'errore $\left| \frac{X}{n} - p \right|$ sarà al più uguale a $z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$.

Dalla formula (7.12), risolvendo rispetto a n, si ricava l'ampiezza del campione necessaria per stimare la proporzione p con un errore prefissato E e con un dato grado di fiducia (si ricordi che n deve essere un intero)

$$n \geq p(1-p)\left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \qquad (7.13)$$

Questa formula non può essere usata se non si ha qualche informazione sul valore di p; se tali informazioni non sono disponibili, si può far uso del fatto che il valore massimo[3] che può assumere la quantità $p(1-p)$ è $\frac{1}{4}$, corrispondente a $p = \frac{1}{2}$.

In questo caso l'ampiezza necessaria per il campione è (si ricordi che n deve essere un intero)

$$n \geq \frac{1}{4}\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \qquad (7.14)$$

Esempio 15

Problema del sondaggio di opinione. Supponiamo che si voglia stimare la proporzione di elettori che approva l'operato del capo del governo; su un campione di 150 persone intervistate, 90 si sono dichiarate favorevoli.

Determinare un intervallo di confidenza con grado di fiducia del 95% per la proporzione degli elettori favorevoli al capo del governo e valutare la precisione della stima.

La proporzione campionaria dei favorevoli è

$$\hat{p} = \frac{x}{n} = \frac{90}{150} = 0.6$$

L'intervallo di confidenza con grado di fiducia del 95% è il seguente

$$0.6 - 1.96 \cdot \sqrt{\frac{0.6 \cdot (1-0.6)}{150}} < p < 0.6 + 1.96 \cdot \sqrt{\frac{0.6 \cdot (1-0.6)}{150}}$$

$$0.52 < p < 0.68$$

La percentuale dei favorevoli, con un grado di fiducia del 95%, è compresa fra il 52% e il 68%: la stima è troppo imprecisa, l'ampiezza dell'intervallo è di 16 punti percentuali.

Può quindi essere utile determinare l'ampiezza del campione necessaria per ottenere una stima con precisione fissata. Stabiliamo ad esempio che si vuole una stima con una precisione dell'1% (corrispondente a un'ampiezza dell'intervallo non superiore a 2 punti percentuali), ossia fissiamo

$$E = 0.01.$$

Dato che non abbiamo informazioni circa la percentuale dei favorevoli nel nuovo campione, dobbiamo usare la formula (7.14) e in tal caso, per il grado di fiducia del 95% , si ottiene

$$n \geq \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604$$

Esempio 16

Supponiamo di voler stimare la proporzione di pezzi difettosi in un lotto di oggetti di un dato tipo con un errore $E = 0.04$ e un grado di fiducia del 95%; calcolare l'ampiezza necessaria per il campione, nel caso che

a – non si abbia alcuna informazione su quale possa essere la proporzione effettiva della popolazione;

b – si sappia che la proporzione della popolazione non supera il 12%.

a – Se non si ha alcuna informazione su p, si usa la formula (7.14), e con grado di fiducia del 95% si ricava

$$n \geq \frac{1}{4}\left(\frac{1.96}{0.04}\right)^2 = 600.3$$

Occorre quindi un campione di ampiezza $n = 601$.

b – Se sappiamo che $p \leq 0.12$, con la formula (7.13) e con grado di fiducia del 95% si ottiene

$$n \geq 0.12(1-0.12)\left(\frac{1.96}{0.04}\right)^2 = 253.5\,.$$

Occorre in questo caso un campione di ampiezza $n = 254$.

Questo esempio illustra come il fatto di avere qualche informazione sul possibile valore della proporzione può sensibilmente ridurre la dimensione del campione.