Università degli Studi di Napoli "Parthenope"

Corso di Laurea in Statistica e Informatica per l'Azienda, la Finanza e le Assicurazioni (SIAFA)

STATISTICA
II MODULO

Sergio LONGOBARDI
longobardi@uniparthenope.it

Stima puntuale

Stima puntuale

L'inferenza permette di stimare i parametri di una popolazione utilizzando le informazioni derivanti da un campione (dati campionari).

La *stima puntuale* dei parametri consiste nell'individuare una funzione dei dati campionari che fornisca una buona approssimazione del parametro ignoto

Stima puntuale

Sia $X_1, X_2, \ldots, X_n$ un campione di dimensione n e

$x_1, x_2, \ldots, x_n$ il corrispondente campione osservato

Obiettivo:

Ottenere attraverso un’opportuna funzione delle osservazioni:

$$t(x_1, x_2, .. x_n)$$

una stima del parametro incognito θ

Stima puntuale

La stima $t=t(x_1, x_2, \ldots, x_n)$ può essere considerata come una realizzazione della variabile casuale $T=t(X_1, X_2, \ldots, X_n)$ chiamata **stimatore** di θ

Variabili casuali $X_1, \ldots, X_n$ *campione casuale*	Variabile casuale $T = t(X_1, \ldots, X_n)$ *stimatore*
↓	↓
Realizzazioni $x_1, \ldots, x_n$ *campione osservato*	Realizzazione $t = t(x_1, \ldots, x_n)$ *stima*
(Insieme di n valori: i *dati del campione*)	(Singolo valore: la stima di θ)

Esempio: campione osservato (2,5,3,6,4,4,1,2,2,5)

Parametro: θ =μ media della popolazione.

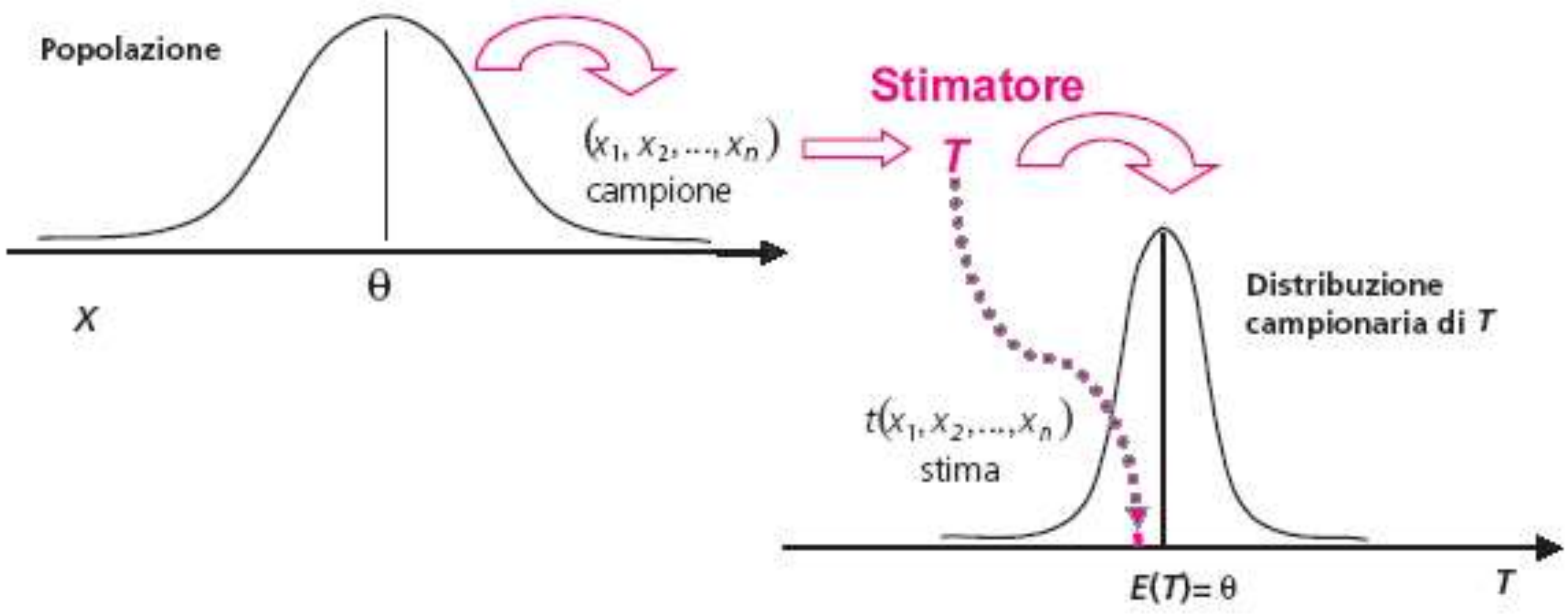
Stimatore: media campionaria

$$\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$$

Stima: $\bar{X} = 3{,}4$

Stima puntuale

Lo stimatore, dipendendo dal campione, è una variabile casuale e quindi possiede una **distribuzione campionaria** la cui conoscenza permette di capire se lo stimatore scelto produrra con elevata probabilità stime “vicine al valore vero del parametro.



Stima puntuale

Per valutare la "bontà" di uno stimatore si può guardare alle sue proprietà:

Proprietà per ***n*** finito:

- Correttezza
- Efficienza

Proprietà per $n \to +\infty$ (asintotiche):

- Consistenza
- Correttezza asintotica

Stima puntuale

Correttezza

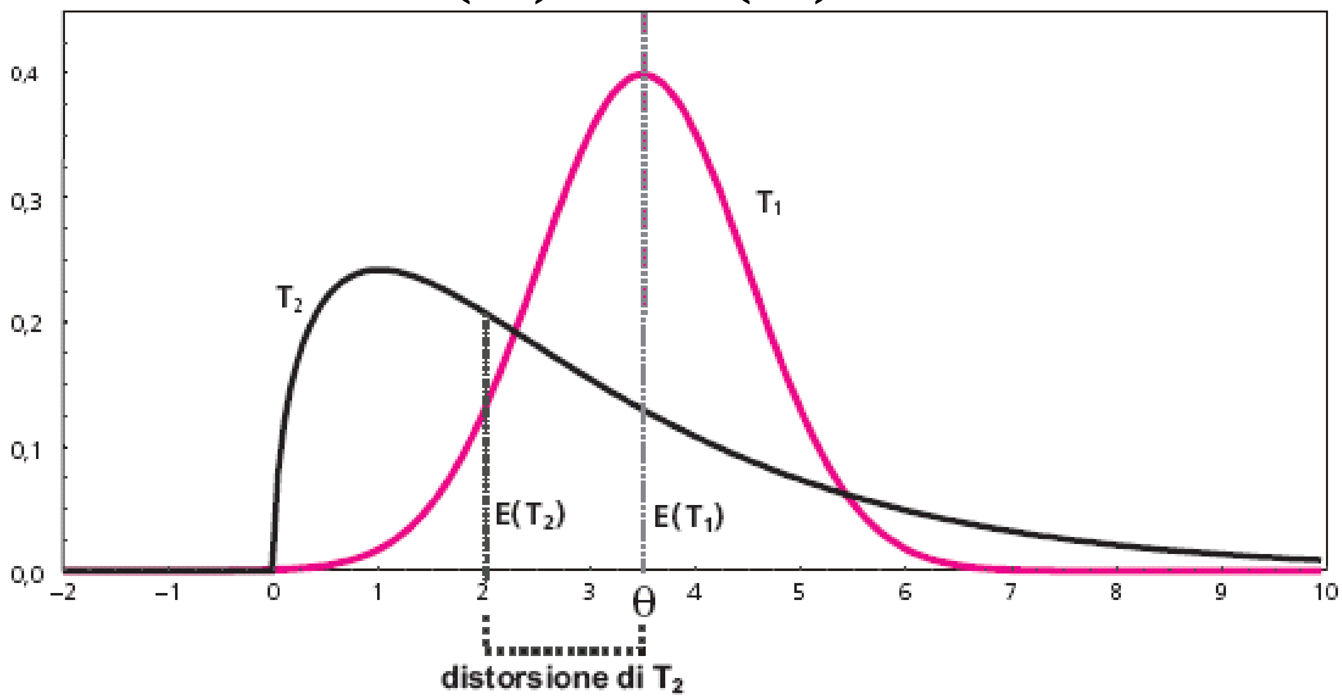
Lo stimatore ***T*** è uno stimatore **corretto** di θ se

$$E(T) = \theta$$

Per tutti i possibili valori di θ

La **distorsione** di uno stimatore è uguale a

$$B(T) = E(T) - \theta$$



Stima puntuale

Esempio

Si considerino due stimatori differenti per stimare la media della popolazione sulla base di un campione di n=10 osservazioni:

$T_1=(X_1, X_2, \ldots,X_{10})/10$ $\qquad$ $T_1=(X_1, X_2, \ldots,X_{10})/5$

Il valore atteso di T_1 è pari a:

$$E(T_1) = \frac{1}{10}E(X_1 + X_2 + \cdots + X_{10}) = \frac{E(X_1) + E(X_2) + \cdots + E(X_{10})}{10} = \frac{10\mu}{10} = \mu$$

Il valore atteso di T_2 è pari a:

$$E(T_2) = \frac{1}{5}E(X_1 + X_2 + \cdots + X_{10}) = \frac{E(X_1) + E(X_2) + \cdots + E(X_{10})}{5} = \frac{10\mu}{5} = 2\mu$$

La distorsione di T_2 sarà pari a:

$$B(T_2) = 2\mu - \mu = \mu$$

Stima puntuale

Per valutare la prossimità di ***T*** a ***θ*** possiamo usare l'**errore quadratico medio** (mean square error) dato dalla quantità:

$$\boxed{MSE(T) = E[(T - \theta)]^2}$$

Proprietà:

$$MSE(T) = E[(T - \theta)]^2 = Var(T) + B(T)^2$$

$$Var(T) = E[T - E(T)]^2$$

Diremo che ***T_1* è più efficiente di *T_2*** se

$$\boxed{MSE(T_1) < MSE(T_2)}$$

Per tutti i possibili valori di ***θ***

Stima puntuale

Se lo stimatore è corretto e quindi è nulla la distorsione si ha:

$$MSE(T) = Var(T)$$

per tutti i possibili valori di $\boldsymbol{\theta}$.

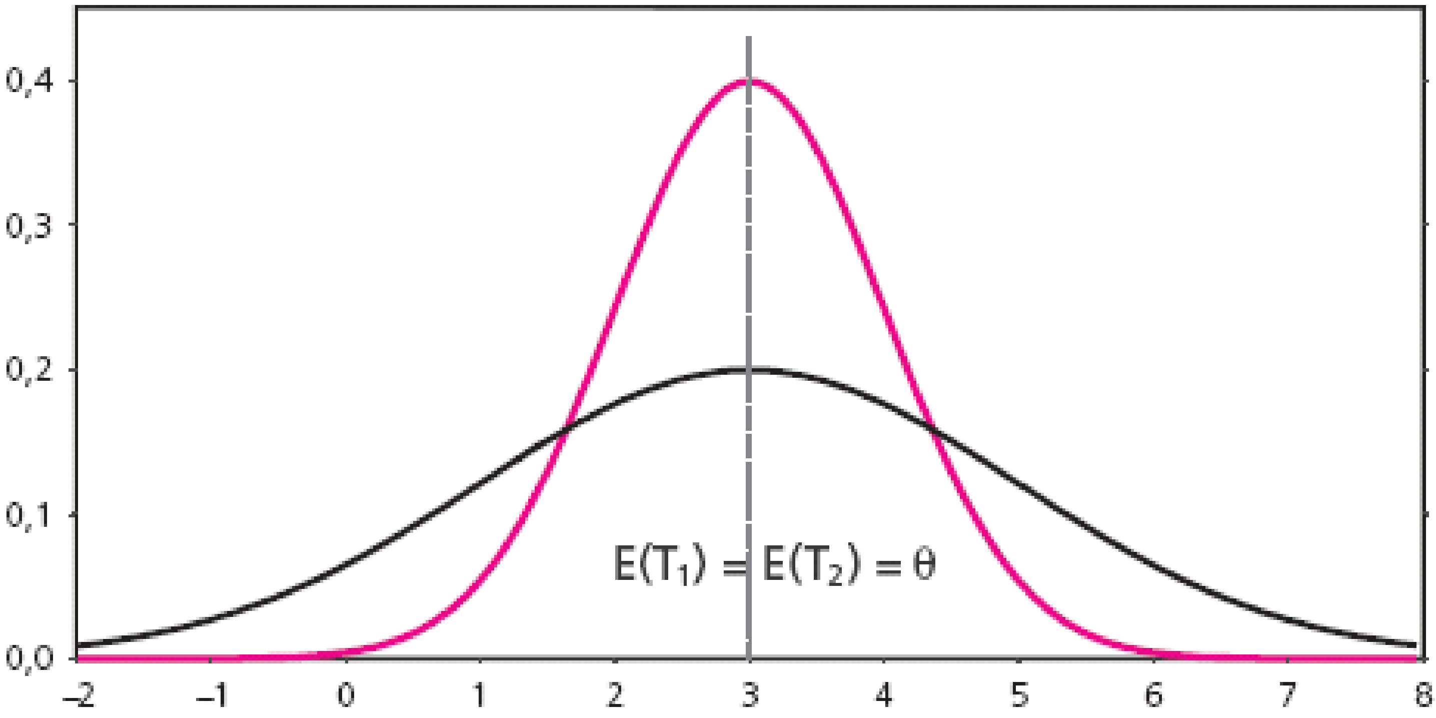
Dati due stimatori corretti $\boldsymbol{T_1}$ e $\boldsymbol{T_2}$, si dirà che $\boldsymbol{T_1}$ è più efficiente di $\boldsymbol{T_2}$ se:

$$Var(T_1) < Var(T_2)$$

Per tutti i possibili valori di $\boldsymbol{\theta}$

Stima puntuale

Nella figura sono riportate le distribuzioni campionarie di due stimatori corretti. lo stimatore T_1 (linea rossa) possiede un errore quadratico medio (ossia una varianza) più piccolo di T_2 (linea nera).



Esempio

Si consideri il campione $X_1, X_2, .. X_4$ estratto da una popolazione con media μ e varianza σ^2, e i due stimatori della media:

$$T_1 = \bar{X} = \frac{1}{4}\sum X_i \qquad T_2 = \frac{X_1 + 2X_2 + 3X_3 + X_4}{4}$$

Allora si ha:

$$E(T_1) = E(\bar{X}) = \mu$$

$$Var(T_1) = Var(\bar{X}) = \sigma^2/4$$

$$E(T_2) = (7/4)\mu$$

$$B(T_2) = (7/4)\mu - \mu = (3/4)\mu$$

$$Var(T_2) = (15/16)\sigma^2$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$MSE(T_1) = \sigma^2/4 \quad < \quad MSE(T_2) = (15/16)\sigma^2 + (9/16)\mu^2$$

Stima puntuale

Consistenza

Lo stimatore T_n di un parametro θ , dove l'indice ***n*** indica la dipendenza dello stimatore dalla numerosità campionaria, è uno stimatore **consistente in media quadratica** se

$$\lim_{n\to\infty} MSE(T_n) = \lim_{n\to\infty} E(T_n - \theta)^2 = 0$$

Quindi

$$\lim_{n\to\infty} MSE(T_n) = 0 \text{ se e solo se } \lim_{n\to\infty} Var(T_n) = 0 \quad \lim_{n\to\infty} B(T_n) = 0$$

Allora uno stimatore **corretto** è consistente se:

$$\lim_{n\to\infty} MSE(T_n) = \lim_{n\to\infty} Var(T_n) = 0$$

Stima puntuale

Correttezza asintotica

Uno stimatore T_n di un parametro θ è uno stimatore **asintoticamente corretto** se:

$$\lim_{n\to\infty} E(T_n) = \theta$$

per ogni possibile valore di θ

Esempio

$$T_n = \frac{X_1 + \cdots + X_{n-1}}{n-1} + \frac{X_n}{n}$$

$$E(T_n) = \mu + \mu/n \quad B(T_n) = \mu/n$$

$$Var(T_n) = \left(\sigma^2/(n-1)\right) + \left(\sigma^2/n^2\right)$$

$$\Rightarrow \begin{cases} \lim\limits_{n\to\infty} B(T_n) = 0 \\ \lim\limits_{n\to\infty} Var(T_n) = 0 \end{cases}$$

asint. corretto

consistente

$$\lim_{n\to\infty} MSE(T_n) = 0$$

Stima puntuale

Si consideri una popolazione X con media μ e varianza σ^2

La media campionaria $\overline{X}$ è uno stimatore corretto per la media della popolazione, ossia

$$E(\overline{X}) = \mu$$

La varianza della media campionaria è $Var(\overline{X}) = \sigma^2 / n$ pertanto è uno stimatore consistente, poiché

$$\lim_{n \to \infty} MSE(\overline{X}_n) = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0$$

- Se la popolazione è distribuita come una Normale, $N(\mu, \sigma^2)$ allora anche la media campionaria si distribuisce come una Normale

$$\overline{X} \to N\left(\mu, \frac{\sigma^2}{n}\right)$$

Media campionaria (esercizi)

- Se la popolazione è distribuita come una Normale, $N(\mu, \sigma^2)$ allora anche la media campionaria si distribuisce come una Normale .

Nel caso di popolazioni infinite:

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right)$$

Nel caso di popolazioni finite

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\frac{N-n}{N-1}\right)$$

Media campionaria (esercizi)

Esercizio 3.5.2.1

Data la popolazione finita indicata in tabella, si verifichi empiricamente il teorema del limite centrale costruendo la distribuzione delle medie campionarie nei casi in cui: a) *n* = 2; b) *n* = 3; c) *n* = 4;

Famiglie	Reddito mensile (migliaia di euro)
A	1.500
B	3.150
C	1.200
D	1.350
E	1.050

$$\mu = \frac{1.500 + 3.150 + 1.200 + 1.350 + 1.050}{5} = 1.650$$

$$\sigma = \sqrt{\frac{(1.500 - 1.650)^2 + (3.150 - 1.650)^2 + \ldots + (1.050 - 1.650)^2}{5}} = 764{,}85$$

Media campionaria (esercizi)

Famiglie	Reddito totale	Reddito medio
AB	1.500 + 3.150 = 4.650	2.325
AC	1.500 + 1.200 = 2.700	1.350
AD	1.500 + 1.350 = 2.850	1.425
AE	1.500 + 1.050 = 2.550	1.275
BC	3.150 + 1.200 = 4.350	2.175
BD	3.150 + 1.350 = 4.500	2.250
BE	3.150 + 1.050 = 4.200	2.100
CD	1.200 + 1.350 = 2.550	1.275
CE	1.200 + 1.050 = 2.250	1.125
DE	1.350 + 1.050 = 2.400	1.200
	Totale	16.500

$$E(\overline{x}) = \frac{\sum \overline{x}}{n_{camp}} = \frac{16.500}{10} = 1.650 = \mu$$

Lo scarto quadratico medio delle medie campionarie è:

$$\sigma_{\overline{x}} = \sqrt{\frac{(2.325-1.650)^2 + (1.350-1.650)^2 + \ldots + (1.200-1.650)^2}{10}} = 468{,}37$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{2}}\sqrt{\frac{5-2}{5-1}} = \frac{764{,}8529}{\sqrt{2}}\sqrt{\frac{3}{4}} = 468{,}3748$$

Media campionaria (esercizi)

Famiglie	Reddito totale	Reddito medio
ABC	1.500 + 3.150 + 1.200 = 5.850	1.950
ABD	1.500 + 3.150 + 1.350 = 6.000	2.000
ABE	1.500 + 3.150 + 1.050 = 5.700	1.900
ACD	1.500 + 1.200 + 1.350 = 4.050	1.350
ACE	1.500 + 1.200 + 1.050 = 3.750	1.250
ADE	1.500 + 1.350 + 1.050 = 3.900	1.300
BCD	3.150 + 1.200 + 1.350 = 5.700	1.900
BCE	3.150 + 1.200 + 1.050 = 5.400	1.800
BDE	3.150 + 1.350 + 1.050 = 5.550	1.850
CDE	1.200 + 1.350 + 1.050 = 3.600	1.200
	Totale	16.500

$$\mathrm{E}(\overline{\mathrm{x}}) = \frac{\sum \overline{x}}{n_{camp}} = \frac{16.500}{10} = 1.650 = \mu$$

Lo scarto quadratico medio delle medie campionarie è:

$$\sigma_{\overline{\mathrm{x}}} = \sqrt{\frac{(1.950-1.650)^2 + (2.000-1.650)^2 + \ldots + (1.200-1.650)^2}{10}} = 312{,}24$$

$$\sigma_{\overline{\mathrm{x}}} = \frac{\sigma}{\sqrt{3}}\sqrt{\frac{5-3}{5-1}} = \frac{764{,}8529}{\sqrt{3}}\sqrt{\frac{2}{4}} = 312{,}2499$$

Media campionaria (esercizi)

Esercizio 3.5.2.5 (Quintano-Castellano, pag.106)

Sia data X ~ ?(220, 400) e si abbia un campione di numerosità n = 100 estratto da una popolazione infinita ($N = \infty$). Si risolvano le seguenti relazioni:

$$P(\bar{x} \geq 225) = ?$$

$$P(216 \leq \bar{x} \leq 223) = ?$$

$$P(\bar{x} \geq x^*) = 0{,}3156$$

Media campionaria (esercizi)

Soluzione

Poiché la popolazione è infinita, la media campionaria ha la seguente distribuzione:

$$\bar{\mathrm{x}} \sim \mathrm{N}\left(\mu_x ,\ \frac{\sigma_x^2}{n} \right)$$

Il valore standardizzato è:

$$\mathrm{z} = \frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}$$

$$P(\bar{x} \geq 225) = \mathrm{P}\left(z > \frac{225 - 220}{20/10} \right)$$

$$= \mathrm{P}(z > 2{,}5) = 0{,}0062$$

Media campionaria (esercizi)

$$P(216 \le \bar{x} \le 223) = \mathrm{P}\left(\frac{216 - 220}{20/10} < z < \frac{223 - 220}{20/10} \right)$$

$$= \mathrm{P}(-2 < z < 1{,}5) = \mathrm{P}(z > -2) - \mathrm{P}(z > 1{,}5)$$

$$= 0{,}9772 - (1 - 0{,}9332) = 0{,}9104$$

$$\mathrm{P}(\bar{x} \ge x^*) = 0{,}3156 = \mathrm{P}\left(z > \frac{x^* - 220}{20/10} \right) = 0{,}3156$$

$$\mathrm{P}(z > z^*) = 0{,}3156 = \mathrm{P}(z > 0{,}48) = 0{,}3156$$

$$z^* = \frac{x^* - 220}{2} = 0{,}48 \qquad x^* = 220{,}96$$

Stima puntuale (parametro proporzione)

Si consideri una popolazione X distribuita come una Bernoulli con parametro π

La media campionaria $\bar{X}$ è uno stimatore corretto della proporzione π della popolazione, ossia:

$$E(\bar{X}) = \pi$$

La varianza della media campionaria è

$$Var(\bar{X}) = \pi(1-\pi)/n$$

pertanto è uno stimatore consistente, poiché

$$\lim_{n\to\infty} MSE(\bar{X}_n) = \lim_{n\to\infty} \frac{\pi(1-\pi)}{n} = 0$$

Stima puntuale (parametro proporzione)

Sia *p* la proporzione del campione che possiede il carattere dato, e sia *π* la relativa proporzione nella popolazione (infinita). La proporzione campionaria (per *n* grande) si distribuisce normalmente

$$\mathrm{p} \sim \mathrm{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Più in generale, se la popolazione è finita, la proporzione campionaria si distribuisce normalmente :

$$\mathrm{p} \sim \mathrm{N}\left(\pi, \ \frac{\pi(1-\pi)}{n} \cdot \frac{N-n}{N-1}\right)$$

Proporzione campionaria

Esercizio 3.5.3.2

Sia π = 0,20 la proporzione di successi in una popolazione infinita. Si estragga un campione di *n* = 100 elementi. Calcolare la varianza della distribuzione campionaria (e lo scostamento quadratico medio) e la probabilità di avere valori della proporzione campionaria inferiori di *p*=0,14.

Soluzione

La varianza campionaria è data dalla formula

$$\sigma_p^2 = \frac{\pi(1-\pi)}{n} = \frac{0{,}2 \cdot 0{,}8}{100} = \frac{0{,}16}{100}$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0{,}2 \cdot 0{,}8}{100}} = \sqrt{\frac{0{,}16}{100}} = \frac{0{,}4}{10} = 0{,}04$$

Proporzione campionaria

Poiché *p* si distribuisce normalmente, si può impiegare la distribuzione di probabilità normale per calcolare la probabilità cercata

$$P(p < 0{,}14) = \mathrm{P}\left(z < \frac{0{,}14 - 0{,}20}{\sqrt{\dfrac{(0{,}2)\,(0{,}8)}{100}}} \right) = P\left(z < -\frac{0{,}06}{0{,}04} \right)$$

= P(*z* < –1,5) = P(*z* > 1,5) = 0,0668

Proporzione campionaria

Esercizio 3.5.3.3

Dalle statistiche di produzione risulta che il 30% di batterie è guasto. Qual è la probabilità di selezionare un campione di n = 64 batterie con proporzione di difetti del 20% o più, campionando da un lotto di N= 1.000 pezzi?

Soluzione: La proporzione di batterie guaste nella popolazione è π = 0,30.

Lo scostamento quadratico medio della proporzione campionaria è:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}\,\frac{N-n}{N-1}} = \sqrt{\frac{0{,}3 \cdot 0{,}7}{64} \cdot \frac{936}{999}} = 0{,}0554$$

La probabilità di avere una proporzione di difetti superiore al 20% viene calcolata dalla normale:

$$P(p > 0{,}20) \;= P\left(z > \frac{0{,}20 - 0{,}30}{0{,}0554} \right) = P(z > -1{,}81) = 0{,}9649$$