Elementi di Statistica

Introduzione

Contrariamente a quanto si potrebbe pensare, la possibilità di attingere a una grandissima quantità di informazioni rischia di impedirci di fatto di utilizzare anche solo una parte di queste: non basta infatti avere solo l'accesso teorico a una informazione, ma occorre che essa sia effettivamente e praticamente fruibile.

Il compito principale della statistica è proprio quello di rendere utilizzabili grandi quantità di informazioni, difficilmente gestibili, relative agli oggetti della propria indagine. Infatti tutte le informazioni, per contribuire effettivamente ad accrescere la conoscenza di un fenomeno, hanno bisogno di essere trattate da vari punti di vista: occorrono tecniche accurate di rilevazione, occorre procedere ad accurate selezioni, occorre un lavoro di organizzazione e di sintesi.

La statistica raccoglie e restituisce in forma organizzata grandi quantità di informazioni. Nel fare ciò obbedisce alla duplice esigenza *predittiva* e *descrittiva*.

- Ogni comunità sente il bisogno, a fini di documentazione, di raccogliere una serie di dati sugli usi, sui costumi, sulle attività sociali ed economiche dei suoi componenti; i censimenti costituiscono uno strumento fondamentale attraverso cui la statistica esplica questa funzione.
- un'altra esigenza a cui risponde la statistica è quella predittiva: la raccolta e l'elaborazione dei dati, e quindi la "fotografia" del passato e del presente, serve per prevedere i comportamenti futuri, per operare scelte, per assumere decisioni.

Durante l'operazione di raccolta dei dati, è spesso impossibile raccogliere tutti i dati, per cui si preferisce riferirsi a una parte significativa di essi detta *campione*. L'insieme del quale il campione è rappresentativo viene definito *popolazione*.

Naturalmente il campione deve essere il più possibile rappresentativo della popolazione; la parte della statistica che stabilisce i criteri di rappresentatività si chiama *inferenza statistica* o *statistica induttiva*.

Per esempio in un sondaggio d'opinione non si intervisteranno tutti gli abitanti di una città, ma solo una parte di essi, scelti in base a determinati criteri.

Potenze decimali e cifre significative

È spesso utile l'utilizzo della notazione scientifica, che consente di scrivere un numero con molti zeri in maniera più compatta.

Si utilizzano potenze di 10. Per esempio $1000 = 10^3$, dove 10 è detto base e 3 esponente.

Spesso inoltre conviene arrotondare i dati; per esempio 11,22, a seconda dell'arrotondamento scelto, si può arrotondare a 11 o 11,2.

Quando l'ultima cifra è un 5, in statistica si usa approssimare alla cifra pari che precede il 5. Esempio: 13,425 diventa 13,42 ma 13.435 diventa 13,44.

Dal concetto di arrotondamento discende quello di cifre significative: le cifre significative sono le cifre di un numero escludendo gli zeri necessari a localizzare la virgola.

Esempio: 11,22 ha 4 cifre significative; 1,23456789 ne ha 10; 0,005 ne ha 1 (difatti, in notazione scientifica lo si può esprimere come $5 \cdot 10^{-3}$).

Un'importante regola riguardo le cifre significative asserisce che in un calcolo il numero di cifre significative del risultato dipende dal numero di cifre significative dei numeri presenti nel calcolo.

Nel caso di addizioni o sottrazioni, il risultato non può avere più cifre significative dopo la virgola del numero presente nel calcolo che ha minor numero di cifre significative dopo la virgola.

Nel caso di prodotti, divisioni o potenze, il risultato non può avere più cifre significative del numero presente nel calcolo che ha minor numero di cifre significative.

Esempi: $3{,}28 + 3{,}5 = 3{,}8$ $\qquad$ $7{,}2 \cdot 5{,}33 = 38$

Medie Ferme

media aritmetica	media ponderata	media geometrica	media armonica	media quadratica
$M = \frac{\sum_{i=1}^{n} x_i}{n}$	$M_p = \frac{\sum_{i=1}^{n} p_i x_i}{\sum_{i=1}^{n} p_i}$	$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$	$H = \frac{1}{\frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	$M_Q = \sqrt{\frac{\sum_{i=1}^{n} x^2{}_i}{n}}$

La *media armonica* è il reciproco della media aritmetica dei loro reciproci.

Le tre medie hanno il seguente ordine di grandezza: $H \le G \le M$

Osservazioni

Nel caso di una distribuzione per classi, il calcolo della media viene fatto sostituendo ciascuna classe con il suo termine centrale, ottenuto calcolando la semisomma dei valori estremi.

La media calcolata fra valori troppo distanti fra di loro non è un valore significativo.

Esempio

100	200	2000	150	550	12000

$$M = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{100 + 200 + 2000 + 150 + 550 + 12000}{6} = \frac{15000}{6} = 2500$$ non è un valore significativo.

Medie lasche

La *mediana* di un insieme di n numeri ordinati dispari è il valore centrale. *Cioè quello che occupa il posto* $int\left(\frac{n}{2}\right)+1$

1	2	3	**5**	7	11	20

$M_e = 5$

La *mediana* di un insieme di numeri ordinati pari è la media dei due valori centrali. *Cioè la media fra il termine* $\frac{n}{2}$ *e il suo successivo. In questo caso* M_e *non è un dato della serie.*

1	2	3	**5**	**7**	9	11	20

$M_e = 6$

Geometricamente la mediana è il valore che divide l'istogramma dei dati in due aree di uguale estensione.

La *mediana* divide l'insieme dei dati in due parti uguali.

I *quartili* dividono l'insieme dei dati in quattro parti uguali.

I *decili* dividono l'insieme dei dati in dieci parti uguali.

I *centili* dividono l'insieme dei dati in cento parti uguali.

La *moda* è il valore che si presenta con la più alta frequenza.

Nel caso in cui più dati hanno la stessa frequenza si parla di distribuzioni bimodali, trimodali, ecc…

Nel caso di una distribuzione in cui ogni valore ha la stessa frequenza, la moda non esiste.

Nel caso in cui la distribuzione sia per classi si parla di *classe modale*.

Se le classi hanno tutte la stessa ampiezza, la classe modale è quella che presenta frequenza maggiore.

Classi	*Frequenza*
$0 \le x < 10$	35
$10 \le x < 20$	**60**
$20 \le x < 30$	25
$30 \le x < 40$	20
$40 \le x < 50$	30
TOTALE	170

Se le classi hanno ampiezza diversa, si valuta il rapporto *Frequenza / Ampiezza*.

Classi	*Frequenza*	*Frequenza / Ampiezza*
$0 \le x < 10$	35	3,5
$10 \le x < 30$	60	3
$30 \le x < 60$	30	1
$60 \le x < 65$	20	**4**
$65 \le x < 70$	10	2
TOTALE	175	

Teoria delle distribuzioni di frequenze

La statistica studia la raccolta dei dati relativi a un insieme di entità, detto campione. In genere questi dati non sono ordinati. L'operazione di ordinamento dei dati prende il nome di distribuzione di frequenze.

Una *serie* è un ordinamento crescente o decrescente.

Il *campo di variazione* di una serie è la differenza tra il dato maggiore e il dato minore.

Nell'esempio a lato esso vale: $12 - (-1) = 13$

Temperature registrate			
4	3	2	2
1	-1	2	4
6	7	8	12

Dopo avere ordinato i dati occorre suddividerli in classi (almeno 5) di uguale ampiezza. Inoltre occorre calcolare sulla destra di ciascuna classe la relativa frequenza (assoluta) con cui il dato della corrispondente classe si presenta.

Una *distribuzione di frequenze* è un ordinamento tabulare in classi e frequenze del tipo precedente.

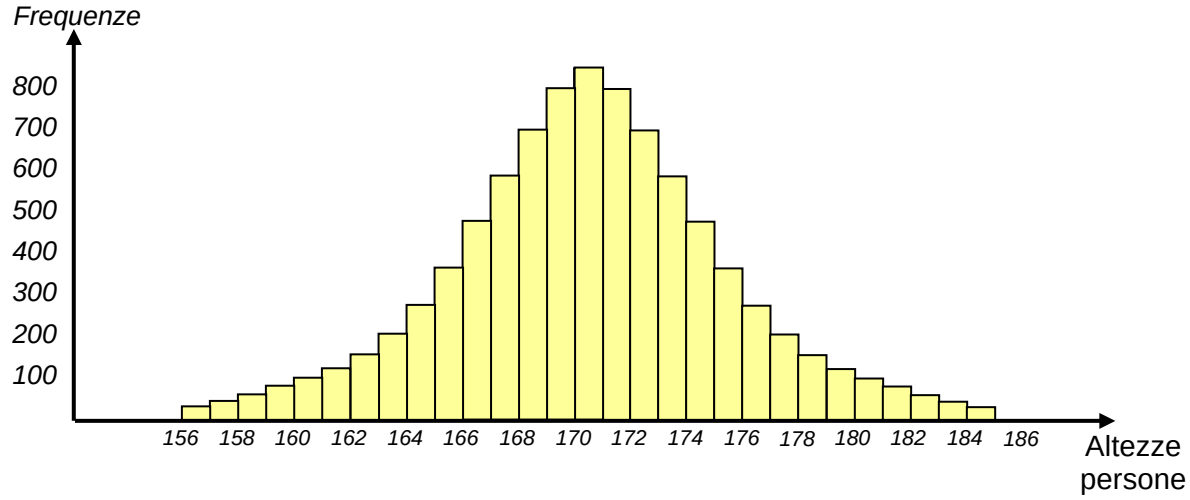
L'*intervallo* o *ampiezza* è la differenza fra il limite superiore e il limite inferiore di una classe.

Se una classe ha come limiti, inferiore e superiore, i numeri 147 e 163, i suoi *confini* reali sono i numeri 146,5 e 163,5.

Il *valore centrale* di una classe è la media aritmetica dei suoi confini reali.

Una distribuzione di frequenza può essere rappresentata mediante *istogrammi*.

L'*istogramma* consiste in una serie di rettangoli affiancati (la cui base inferiore poggia sull'asse orizzontale del grafico, è centrata sul valor centrale ed è larga quanto l'ampiezza della classe) la cui altezza è proporzionale al valore rappresentato.



La *frequenza relativa* di una classe (espressa in percentuale) è il rapporto fra la frequenza assoluta di una classe e il numero totale dei dati.

La *frequenza cumulativa* è la somma delle frequenze delle classi minori o uguali ad una data classe.

Voti	*Frequenza*	*Frequenza cumulata*	*Frequenza relativa*
4	*2*	*2*	*2/30 = 6,7%*
5	*8*	*10*	*8/30 = 26,7%*
6	*12*	*22*	*12/30 = 40%*
7	*6*	*28*	*6/30 = 20%*
8	*2*	*30*	*2/30 = 6,7%*
Totale	*30*		

Le frequenze cumulative possono essere rappresentate tramite istogrammi o tramite poligoni di frequenze cumulative, dette *ogive*.

La variabilità e la concentrazione

Le misure di posizione, come le medie, la moda e la mediana, sono utili come valori di sintesi di una distribuzione o per confrontare distribuzioni diverse. Tali valori però non descrivono in modo esauriente il fenomeno, perché non ci dicono quanto ciascun dato si discosta dal valore di sintesi considerato.

Per esempio, se una serie di dati relativi alla temperatura rilevata giornalmente e alla stessa ora nell'arco di un mese dà un valore medio di 25°C per il mese di Luglio e di 25°C per il mese di Agosto, questo non ci dà informazioni sul fatto che ci sia stata una variazione di temperatura maggiore in Luglio o in Agosto.

Per avere un quadro più chiaro del fenomeno, ci servono informazioni più precise su come variano e su come si distribuiscono i dati attorno al valore medio calcolato. Lo studio della variabilità si pone proprio l'obiettivo di dare risposte su come si distribuiscono i dati attorno al valore di sintesi in modo da poter confrontare agevolmente diverse serie di dati.

Diventa così possibile dare risposte a domande del tipo:

- nella razza bovina, c'è più variabilità fra il peso o fra le dimensioni dell'animale?
- i prezzi degli appartamenti nelle grandi città sono più o meno variabili di quelli in provincia o di quelli nelle regioni turistiche?
- le somme destinate dalle famiglie al consumo dei beni di prima necessità è più o meno variabile di quello destinato ai beni voluttuari?

Il campo di variabilità

In relazione ad un fenomeno statistico una prima informazione sulla variabilità può essere data dalla differenza fra il valore più grande e quello più piccolo osservati.

Tale differenza si dice *campo di variabilità* e, per come è stato definito, è un numero positivo espresso nella stessa unità di misura dei dati.

Il campo di variabilità è però un indice piuttosto grossolano della variabilità ed ha il difetto di essere grandemente influenzato dai valori estremi delle rilevazioni.

Supponiamo, ad esempio, che i rilevamenti compiuti su un campione di individui sulla pressione minima sanguigna abbia dato i seguenti risultati: 80 80 85 90 85 60 90 95 95 80 85 115.

Il campo di variabilità di questi dati è dato da 115 – 60 = 55. Se basassimo le nostre considerazioni solo su questo valore, saremmo portati a dire che in quel gruppo di persone vi è un'alta variabilità fra i dati, mentre in realtà, osservando meglio, si nota che la maggior parte di essi (tutti tranne due) si distribuiscono in un ambito più ristretto compreso fra 80 e 95. Dobbiamo allora costruire degli strumenti capaci di misurare la variabilità in modo significativo.

Scostamento, scarto quadratico medio e varianza

Supponiamo che quattro studenti, che indicheremo con A, B, C, D, abbiano conseguito i seguenti punteggi in una serie di 4 test di ammissione ad un corso di specializzazione.

A	*26*	*16*	*24*	*30*
B	*10*	*26*	*30*	*30*
C	*25*	*26*	*23*	*22*
D	*26*	*24*	*24*	*22*

Se solo due di essi potranno essere ammessi al corso, come stendere una graduatoria di ammissione?

La prima cosa che viene in mente di fare è calcolare la media aritmetica dei punteggi conseguiti da ognuno di essi: tale media è però 24 in tutti e quattro i casi; quindi non ci possiamo basare su di essa per il confronto fra gli studenti.

Se però confrontiamo le distribuzioni dei punteggi nei quattro casi, ci accorgiamo che essi si distribuiscono in modo molto diverso uno dall'altro rispetto alla media. Questo fatto ci suggerisce di studiare la variabilità

come studio della dispersione intorno ad un valore fissato, detto polo, che di solito coincide con una delle misure di posizione, nel nostro caso la media aritmetica.

Cominciamo allora a calcolare la distanza di ciascuno dei dati dalla media. Si ha che:

per lo studente A gli scarti sono: $26 - 24 = 2$ $16 - 24 = -8$ $24 - 24 = 0$ $30 - 24 = 6$

per lo studente B gli scarti sono: $10 - 24 = -14$ $26 - 24 = 2$ $30 - 24 = 6$ $30 - 24 = 6$

per lo studente C gli scarti sono: $25 - 24 = 1$ $26 - 24 = 2$ $23 - 24 = -1$ $22 - 24 = -2$

per lo studente D gli scarti sono: $26 - 24 = 2$ $24 - 24 = 0$ $24 - 24 = 0$ $22 - 24 = -2$

Per sintetizzare questi scarti potremmo calcolare la loro media; tuttavia, poiché sappiamo che la somma degli scarti dalla media aritmetica è nulla, questo calcolo non ci darebbe informazioni aggiuntive sulla dispersione.

Allora, riflettendo sul fatto che la somma degli scarti è nulla perché gli scarti negativi compensano quelli positivi, possiamo pensare di eliminare l'influenza del segno considerando i quadrati degli scarti e facendone poi la media che chiameremo media quadratica.

Nel caso dei nostri studenti avremo dunque che la media quadratica degli scarti è:

Studente A	*Studente B*	*Studente C*	*Studente D*
$\sqrt{\frac{2^2+(-8)^2+0^2+6^2}{4}}=5{,}1$	$\sqrt{\frac{14^2+2^2+6^2+6^2}{4}}=8{,}2$	$\sqrt{\frac{1^2+2^2+1^2+2^2}{4}}=1{,}6$	$\sqrt{\frac{2^2+0^2+0^2+2^2}{4}}=1{,}4$

Si può allora concludere che lo studente D presenta una minor variabilità, seguito nell'ordine dagli studenti C, A, B. I due studenti ammessi al corso saranno quindi D e C, in quanto il loro rendimento è più costante.

Abbiamo detto che è conveniente studiare la variabilità mediante uno studio della dispersione intorno ad un polo prefissato; tale polo coincide generalmente con una delle misure di posizione. Gli indici per misurare la dispersione sono di solito delle sintesi delle distanze fra il polo considerato e le osservazioni fatte. Nell'esempio che abbiamo preso in considerazione, la sintesi è stata fatta considerando come polo la media aritmetica e calcolando poi la media quadratica degli scarti, ma in altre circostanze il polo potrebbe benissimo essere la mediana o un altro valore ed il calcolo riferirsi ad altre quantità. Occorre poi tenere presente che, affinché la funzione che sintetizza le distanze dal polo consenta di fare confronti fra distribuzioni diverse, è opportuno considerare come polo il valore che minimizza la funzione stessa.

Poiché la media aritmetica rende minima la somma dei quadrati degli scarti, ecco perché è stata usata proprio questa funzione per sintetizzare la dispersione.

In altri casi, si sceglie come polo la mediana, poiché essa rende minima la somma dei valori assoluti degli scarti.

Lo *scarto semplice medio assoluto* è un indice che misura il grado di dispersione dei dati rispetto a quello del valore medio:

$$\overline{|x_i - M|} = \frac{\sum_{i=1}^{n} |x_i - M|}{n}$$

Se si conoscono anche le frequenze si applica:

$$\overline{|x_i - M|} = \frac{\sum_{i=1}^{n} f_i \cdot |x_i - M|}{n}$$

(dove M rappresenta la media)

Se si considera la mediana al posto della media si ottiene lo *scostamento medio*. Lo scostamento medio ha la proprietà di avere valore minimo.

$$\overline{|x_i - M_e|} = \frac{\sum_{i=1}^{n} |x_i - M_e|}{n}$$

(dove Me rappresenta la mediana)

Lo *scarto quadratico medio* è quello più utilizzato: $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2}{n}}$$

Se si conoscono anche le frequenze si applica: $$s = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (x_i - M)^2}{n}}$$

Si usa s per indicare lo *s.q.m.* riferito ad un *campione*. Si usa σ per indicare lo *s.q.m.* riferito alla *popolazione*.

La *varianza* è il quadrato dello scarto quadratico medio $$s^2 = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n}$$

Esempio

La seguente tabella riporta la distribuzione dei pesi di un campione di 100 persone.

Fascia di peso (kg)	*N° persone*
60 – 62	*5*
63 – 65	*18*
66 – 68	*42*
69 – 71	*27*
72 - 74	*11*

Calcolare :

A. il campo di variazione dei pesi
B. lo scostamento semplice medio assoluto
C. lo scarto quadratico medio dei pesi
D. la varianza
E. la varianza corretta secondo Sheppard

Soluzione

A. il campo di variazione dei pesi può essere calcolato in due modi:

1. come differenza tra il confine superiore della classe più pesante e quello inferiore della classe meno pesante
   $C = 74{,}5 - 59{,}5 = 15$
2. come differenza dei valori centrali tra la classe più pesante e quella meno pesante
   $C = 73 - 61 = 12$

B. Occorre calcolare prima la media provvisoria dei valori centrali delle classi : $A = \frac{61 + 64 + 67 + 70 + 73}{5} = 67$

x (valore centrale)	$D = x_i - A$	f	$f \cdot D$
61	*- 6*	*5*	*- 30*
64	*-3*	*18*	*- 54*
67	*0*	*42*	*0*
70	*3*	*27*	*81*
73	*6*	*8*	*48*
		$\sum f \cdot D$	*45*

e calcolare poi la media mediante la formula : $M = A + \frac{\sum f \cdot D}{n} = 67 + \frac{45}{100} = 67{,}45$

Lo *scostamento semplice medio assoluto* è dato dalla formula : $\overline{|x_i - M|} = \dfrac{\sum_{i=1}^{n} f \cdot |x_i - M|}{n}$

x (valore centrale)	$\|x_i - M\|$	*f*	$f \cdot \|x_i - M\|$
61	*6,45*	*5*	*32,25*
64	*3,45*	*18*	*62,10*
67	*0,45*	*42*	*18,90*
70	*2,55*	*27*	*68,85*
73	*5,55*	*8*	*44,40*

Pertanto lo scostamento semplice medio è :

$$\overline{|x_i - M|} = \frac{\sum_{i=1}^{n} f \cdot |x_i - M|}{n} = \frac{32,25 + 62,10 + 18,90 + 68,85 + 44,40}{100} = 2,26$$

C. lo scarto quadratico medio dei pesi è dato dalla formula : $s = \sqrt{\dfrac{\sum_{i=1}^{n} f_i \cdot (x_i - M)^2}{n}}$

x (valore centrale)	$(x_i - M)^2$	*f*	$f \cdot (x_i - M)^2$
61	*41,6025*	*5*	*208,0125*
64	*11,9025*	*18*	*214,2450*
67	*0,2025*	*42*	*8,5050*
70	*6,05025*	*27*	*175,5675*
73	*30,8025*	*8*	*246,4200*
		$\sum_{i=1}^{n} f_i \cdot (x_i - M)^2$	*852,75*

$$s = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (x_i - M)^2}{n}} = \sqrt{\frac{852,75}{100}} = 2,92$$

D. la varianza vale $s^2 = 8,5275$

E. la correzione di Sheppard: $s^2_{Sheppard} = s^2 - \dfrac{c^2}{12} = 8,5275 - \dfrac{3^2}{12} = 8,5275 - 0,75 = 7,7775$.
dove c rappresenta l'ampiezza delle classi.

Altri indici di dispersione

Altri indici di dispersione sono:

- la *variabile standardizzata* data dal rapporto fra la deviazione dalla media e lo *s.q.m.* : $z = \frac{x - M}{s}$
- i *coefficienti di variazione*
- la *semidifferenza interquartile*: $Q = \frac{Q_3 - Q_1}{2}$ dove Q_1 e Q_3 sono il I e il III quartile

Esempio - variabile standardizzata

Uno studente ha ottenuto 84/100 come voto finale dell'esame di Fisica, nel quale ha ottenuto voto medio 76 e scarto quadratico medio 10. Considerando che ha anche ottenuto 90/100 come voto finale all'esame di Chimica, nel quale ha ottenuto voto medio 82 e scarto quadratico medio 16, in quale delle due materie il voto è stato relativamente più alto ?

Soluzione

Considerando la variabile standardizzata z, la quale indica la deviazione (relativa a s) della variabile x, si ha:

$$z_{Fisica} = \frac{84 - 76}{10} = 0{,}8 \qquad z_{Chimica} = \frac{90 - 82}{16} = 0{,}5\,.$$

Pertanto il voto in Fisica è relativamente più alto di quello preso in Chimica.

I coefficienti di variazione

Per confrontare lo stesso carattere su due popolazioni diverse basta confrontare i corrispondenti *s.q.m.* (*in questo caso il confronto viene effettuato nella stessa unità di misura*).

Se invece si devono confrontare due caratteri diversi, espressi con unità di misura diverse oppure con la stessa unità di misura ma di ordini di grandezza diversi (si vuole ad esempio stabilire se c'è più variabilità fra i pesi o fra le altezze di una certa popolazione), occorre utilizzare indici adimensionali:

- la *dispersione relativa* data dal rapporto fra la dispersione assoluta e la media aritmetica.: $D_{Rel} = \frac{D_{Ass}}{M}$
- il *coefficiente di dispersione* o *coefficiente di variazione*, dato dal rapporto fra lo scarto quadratico medio e la media aritmetica: $V = \frac{s.q.m.}{M}$

Esempio 1 - coefficienti di variazione

Nell'analisi relativa alla statura di un gruppo di individui si è ottenuto una media: $M = 169{,}5\ cm$ e uno *s.q.m.* : $\sigma = 6{,}42\ cm$. Nell'analisi relativa al peso dello stesso gruppo di individui si è ottenuto una media: $M = 72{,}58\ kg$ e uno *s.q.m.* : $\sigma = 4{,}93\ kg$. Si vuole sapere se c'è maggior variabilità nei pesi o nelle altezze degli individui.

Soluzione

Il coefficiente di variazione delle altezze è: $V_{Altezze} = \frac{s.q.m.}{M} = \frac{6{,}42}{169{,}5} = 3{,}79\%$.

Il coefficiente di variazione dei pesi è: $V_{Pesi} = \frac{s.q.m.}{M} = \frac{4{,}93}{72{,}58} = 6{,}79\%$.

Pertanto possiamo dedurre che, in quel gruppo di individui, c'è più variabilità fra i pesi che fra le altezze.

Esempio 2 - coefficienti di variazione

Nell'analisi relativa alle dimensioni di un gruppo di semi si è ottenuto una media: $M = 6{,}5\ mm$ e uno *s.q.m.*: $\sigma = 1{,}56\ mm$. Nell'analisi relativa all'altezza di un gruppo di piante adulte si è ottenuto una media: $M = 3{,}8\ m$ e uno *s.q.m.*: $\sigma = 1{,}07\ m$. Si vuole sapere se c'è maggior variabilità nelle dimensioni dei semi o nelle altezze delle piante.

Soluzione

Il coefficiente di variazione delle dimensioni dei semi è: $V_{Semi} = \frac{s.q.m.}{M} = \frac{1{,}56}{6{,}5} = 24\%$.

Il coefficiente di variazione delle altezze delle piante è: $V_{Altezze} = \frac{s.q.m.}{M} = \frac{1{,}07}{3{,}8} = 28{,}16\%$.

Pertanto possiamo dedurre che c'è maggior variabilità fra le altezze delle piante generate da quei semi che non fra i semi stessi.

Osservazione

In entrambi gli esempi, i valori dello *s.q.m.* ci avrebbero dato delle indicazioni errate.

Osservazione

Se una grandezza x ha distribuzione normale com media M e varianza s^2	
il 68,27% dei casi è compreso tra	$M - s$ e $M + s$
il 95,45% dei casi è compreso tra	$M - 2s$ e $M + 2s$
il 99,73% dei casi è compreso tra	$M - 3s$ e $M + 3s$

I momenti

Il momento di ordine r di una serie di dati $x_1, x_2, \ldots, x_n$ si calcola con la seguente formula:

$$\overline{x^r} = \frac{x_1^{\,r} + x_2^{\,r} + \ldots + x_n^{\,r}}{n} = \frac{\sum x_i^{\,r}}{n}$$

Per $n = 1$ si ottiene la media aritmetica: $\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$

Il momento di ordine r dalla media aritmetica M è dato dalla seguente formula: $m_r = \frac{\sum (x_i - M)^r}{n}$.

Per $r = 1$ si ottiene la varianza: $m_1 = s^2 = \frac{\sum (x_i - M)}{n}$.

Esempio

Dati i numeri: $2,3,5,7$, si calcolino:

A. i momenti dei primi 3 ordini
B. i momenti dei primi 3 ordini rispetto al numero 4
C. i momenti dei primi 3 ordini rispetto alla media aritmetica dei dati

Soluzione A

Per $n = 1$ si ottiene: $\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{2 + 3 + 5 + 7}{4} = 4{,}25$

Per $n = 2$ si ottiene: $\overline{x^2} = \frac{x_1^{\,2} + x_2^{\,2} + \ldots + x_n^{\,2}}{n} = \frac{4 + 9 + 25 + 49}{4} = 22{,}25$

Per $n = 3$ si ottiene: $\overline{x^3} = \frac{x_1^{\,3} + x_2^{\,3} + \ldots + x_n^{\,3}}{n} = \frac{8 + 27 + 125 + 343}{4} = 125{,}75$

Soluzione B

Per $n = 1$ si ottiene: $\overline{x - 4} = \frac{(2-4) + (3-4) + (5-4) + (7-4)}{4} = 0{,}25$

Per $n = 2$ si ottiene: $\overline{(x-4)^2} = \frac{(2-4)^2 + (3-4)^2 + (5-4)^2 + (7-4)^2}{4} = 3{,}75$

Per $n = 3$ si ottiene: $\overline{(x-4)^3} = \frac{(2-4)^3 + (3-4)^3 + (5-4)^3 + (7-4)^3}{4} = 9{,}25$

Soluzione C

Per $n = 1$ si ottiene: $\overline{x - M} = \frac{(2-4{,}25) + (3-4{,}25) + (5-4{,}25) + (7-4{,}25)}{4} = 0$

Per $n = 2$ si ottiene: $\overline{(x-M)^2} = \frac{(2-4{,}25)^2 + (3-4{,}25)^2 + (5-4{,}25)^2 + (7-4{,}25)^2}{4} = 3{,}69$

Per $n = 3$ si ottiene: $\overline{(x-M)^3} = \frac{(2-4{,}25)^3 + (3-4{,}25)^3 + (5-4{,}25)^3 + (7-4{,}25)^3}{4} = 1{,}69$

Distribuzione binomiale o di Bernoulli

La probabilità che un evento E si presenti x volte su n prove è: $p(x) = c_{n,x} \cdot p^x \cdot q^{n-x}$

con p probabilità di successo e q probabilità di insuccesso.

Essa rappresenta una distribuzione di probabilità discreta.

Nel caso di distribuzione binomiale, la media, la varianza, lo scarto quadratico medio, l'asimmetria e curtosi sono dati dalle seguenti formule:

Media	*Varianza*	*s. q. m.*	*asimmetria*	*curtosi*
$M = n \cdot p$	$\sigma^2 = n \cdot p \cdot q$	$\sigma = \sqrt{n \cdot p \cdot q}$	$a_3 = \frac{q-p}{\sqrt{n \cdot p \cdot q}}$	$a_4 = 3 + \frac{1-6pq}{n \cdot p \cdot q}$

La *curtosi* rappresenta lo sviluppo in altezza della curva di distribuzione.

Quando una distribuzione di frequenze asimmetrica è più allungata sul lato destro, rispetto al punto di massimo, si definisce *positivamente asimmetrica*.

Nel caso in cui la curva presenti un allungamento dal lato sinistro, si definisce *negativamente asimmetrica*.

Esempio 1

Lanciando una moneta 6 volte, la probabilità di ottenere 2 croci è:

$$p(x) = c_{n,x} \cdot p^x \cdot q^{n-x} = c_{6,2} \cdot p^2 \cdot q^4 = \frac{6 \cdot 5}{2 \cdot 1} \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^4 = 15 \cdot \left(\frac{1}{2}\right)^6 = \frac{15}{64}.$$

Esempio 2

Lanciando una moneta 6 volte, la probabilità di ottenere almeno 5 teste è data dalla somma di :

$$p(\geq 5) = p(5) + p(6) = c_{6,5} \cdot p^5 \cdot q^{6-5} + c_{6,6} \cdot p^6 \cdot q^{6-6} =$$

$$= \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} \cdot \left(\frac{1}{2}\right)^5 \cdot \left(\frac{1}{2}\right)^1 + 1 \cdot \left(\frac{1}{2}\right)^6 \cdot \left(\frac{1}{2}\right)^0 = 6 \cdot \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^6 = 7 \cdot \left(\frac{1}{2}\right)^6 = \frac{7}{64}.$$

Esempio 3

Se il 10% dei pezzi prodotti da una macchina è difettoso, qual è la probabilità che su 5 pezzi scelti casualmente, accada che:

A. nessuno sia difettoso
B. uno sia difettoso
C. due siano difettosi
D. non più di due siano difettosi

Soluzioni

$$p(0) = c_{5,0} \cdot p^0 \cdot q^{5-0} = 1 \cdot \left(\frac{1}{10}\right)^0 \cdot \left(\frac{9}{10}\right)^5 = 1 \cdot 1 \cdot \frac{59049}{100000} = 0{,}59049.$$

$$p(1) = c_{5,1} \cdot p^1 \cdot q^{5-1} = 5 \cdot \left(\frac{1}{10}\right)^1 \cdot \left(\frac{9}{10}\right)^4 = 5 \cdot \frac{1}{10} \cdot \frac{6561}{10000} = 0{,}32805.$$

$$p(2) = c_{5,2} \cdot p^2 \cdot q^{5-2} = \frac{5 \cdot 4}{1 \cdot 2} \cdot \left(\frac{1}{10}\right)^2 \cdot \left(\frac{9}{10}\right)^3 = 10 \cdot \frac{1}{100} \cdot \frac{729}{1000} = 0{,}0729.$$

$$p(\leq 2) = p(0) + p(1) + p(2) = 0{,}59049 + 0{,}32805 + 0{,}0729 = 0{,}99144.$$

Esempio 4

Considerando l'esempio precedente, trovare media, varianza, scarto quadratico medio, coefficiente di asimmetria e curtosi della distribuzione dei pezzi difettosi su di un campione di *1000* pezzi.

Soluzioni

La media è: $M = n \cdot p = 1000 \cdot \frac{1}{10} = 100$.

La varianza è: $\sigma^2 = n \cdot p \cdot q = 1000 \cdot \frac{1}{10} \cdot \frac{9}{10} = 90$.

Lo scarto quadratico medio è: $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{1000 \cdot \frac{1}{10} \cdot \frac{9}{10}} = \sqrt{90} = 9{,}49$

Il coefficiente di asimmetria è: $a_3 = \frac{q - p}{\sqrt{n \cdot p \cdot q}} = \frac{0{,}9 - 0{,}1}{\sqrt{1000 \cdot 0{,}1 \cdot 0{,}9}} = \frac{0{,}8}{\sqrt{,90}} = 0{,}084$

La curtosi è: $a_4 = 3 + \frac{1 - 6pq}{n \cdot p \cdot q} = 3 + \frac{1 - 6 \cdot 0{,}1 \cdot 0{,}9}{1000 \cdot 0{,}1 \cdot 0{,}9} = 3 + \frac{1 - 0{,}54}{90} = 3 + \frac{0{,}46}{90} = 3{,}005$.

Esempio 5

Una macchina A produce 800 lampadine al giorno, e di esse, in media, 38 sono difettose.
Una macchina B produce 750 lampadine al giorno, e di esse, in media, 27 sono difettose.
Qual è la produzione media di lampadine perfette ?
Quale dei due macchinari presenta un maggior grado precisione ?

Soluzione

La macchina A produce una lampadina difettosa con frequenza: $p_A = \frac{38}{800} = 0{,}0475$

La macchina A produce una lampadina perfetta con frequenza: $q_A = 1 - 0{,}0475 = 0{,}9525$

La macchina B produce una lampadina difettosa con frequenza: $p_B = \frac{27}{750} = 0{,}0360$

La macchina B produce una lampadina perfetta con frequenza: $q_B = 1 - 0{,}0360 = 0{,}9640$

La produzione media di lampadine perfette prodotte dalla macchina A è: $M_A(X) = n \cdot q_A = 800 \cdot 0{,}9525 = 762$.

La produzione media di lampadine perfette prodotte dalla macchina B è: $M_B(X) = n \cdot q_B = 750 \cdot 0{,}9640 = 723$.

La produzione media di lampadine perfette è: $M(X) = M_A(X) + M_B(X) = 762 + 723 = 1485$.

La precisione di un macchinario può essere misurata per mezzo dello s. q. m. relativo al valore medio delle lampadine perfette. Quindi:

$\sigma_A = \sqrt{n \cdot p \cdot q} = \sqrt{800 \cdot 0{,}0475 \cdot 0{,}9525} = 6{,}0162$.

$\sigma_B = \sqrt{n \cdot p \cdot q} = \sqrt{750 \cdot 0{,}0360 \cdot 0{,}9640} = 5{,}1018$.

Essendo quindi $\sigma_B < \sigma_A$, la macchina B è più precisa della macchina A.

Una conferma di tale risultato è dato dal calcolo dei coefficienti di variazione:

$V_A = \frac{\sigma_A}{M_A(X)} = \frac{s.q.m._A}{M_A(X)} = \frac{6{,}0162}{762} = 0{,}0079 = 0{,}79\%$.

$V_B = \frac{\sigma_B}{M_B(X)} = \frac{s.q.m._B}{M_B(X)} = \frac{5{,}1018}{723} = 0{,}0071 = 0{,}71\%$.

Distribuzione di Poisson

La distribuzione di probabilità discreta: $p(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$, (*con* $\lambda = np$ *costante*) è detta *distribuzione di Poisson*.

Essa si utilizza nella risoluzione di problemi sulle prove ripetute, con un numero di prove n molto grande e con probabilità p prossimo allo zero (e quindi q prossimo ad 1).

Al crescere di λ, la distribuzione Poisson si avvicina alla distribuzione normale con variabile standardizzata $z = \frac{x - \lambda}{\sqrt{\lambda}}$

Per essa valgono le seguenti formule:

Media	*Varianza*	*s. q. m.*	*asimmetria*	*curtosi*
$M = \lambda$	$\sigma^2 = \lambda$	$\sigma = \sqrt{\lambda}$	$a_3 = \frac{1}{\sqrt{\lambda}}$	$a_4 = 3 + \frac{1}{\sqrt{\lambda}}$

Esempio 1

Data la seguente distribuzione di Poisson $p(x) = \frac{0{,}72^x \cdot e^{-0{,}72}}{x!}$, calcolare: $p(0)$, $p(1)$, $p(2)$.

Soluzioni

$$p(0) = \frac{0{,}72^0 \cdot e^{-0{,}72}}{0!} = \frac{1 \cdot e^{-0{,}72}}{1} = 0{,}48675 .$$

$$p(1) = \frac{0{,}72^1 \cdot e^{-0{,}72}}{1!} = \frac{0{,}72 \cdot e^{-0{,}72}}{1} = 0{,}35046 .$$

$$p(2) = \frac{0{,}72^2 \cdot e^{-0{,}72}}{2!} = \frac{0{,}5184 \cdot e^{-0{,}72}}{2} = 0{,}12616 .$$

Esempio 2

La probabilità che un certo pezzo di un motore si guasti è 0,001. Determinare la probabilità che su 3000 motori:

A. 5 accusino il guasto di quel pezzo
B. più di 3 accusino il guasto di quel pezzo

Soluzione A

Essendo la probabilità che un certo pezzo di un motore si guasti molto piccola ($p = 0{,}001$), si può applicare distribuzione di Poisson.
Il valor medio è: $\lambda = np = 3000 \cdot 0{,}001 = 3$

La probabilità che su 3000 motori, 5 accusino il guasto di quel pezzo è: $p(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} = \frac{3^5 \cdot e^{-3}}{5!} = 0{,}1008 .$

Soluzione B

La probabilità che più di 3 accusino il guasto di quel pezzo può essere calcolata ricorrendo alla proprietà contraria.
$p(>3) = 1 - [\, p(0) + p(1) + p(2) + p(3) \,] = 1 - (0{,}0498 + 0{,}1494 + 0{,}2240 + 0{,}2240) = 0{,}3528 .$

Distribuzione normale o gaussiana

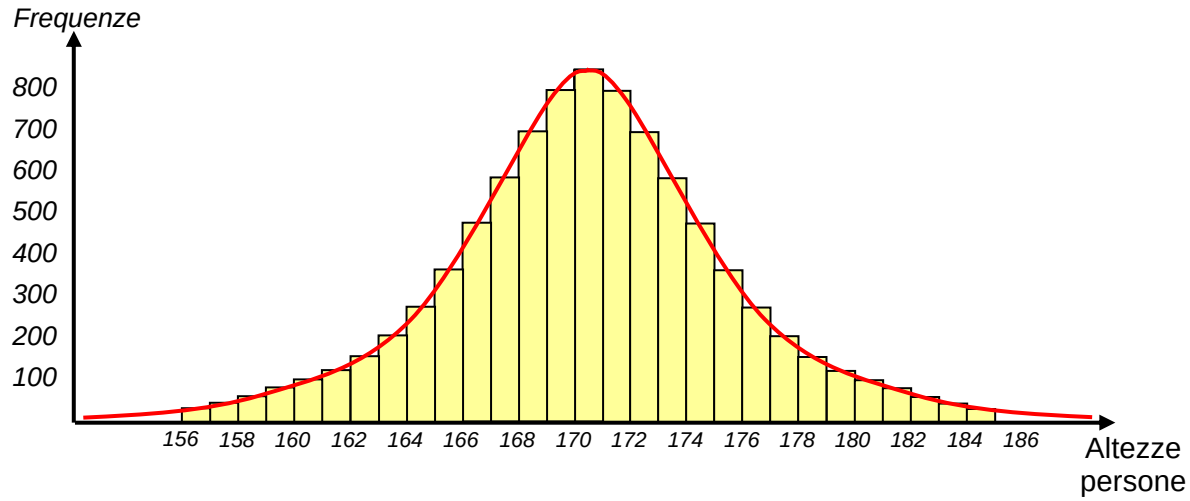
Come nel discreto, anche nel continuo vi sono alcune distribuzioni che regolano particolari fenomeni. Fra tutte, quella che riveste maggiore importanza è la distribuzione normale che approssima in modo soddisfacente molte situazioni.

In quasi tutte le distribuzioni di frequenza (peso di un gruppo di persone, altezze di un gruppo di persone, ecc…), si evidenziano le seguenti caratteristiche comuni:

- pochi dati appartengono alle classi più basse e alla classi più alte
- la maggior parte dei dati si concentra, in modo progressivo, attorno ad un valore medio che, occupa approssimativamente la classe centrale dell'istogramma.

L'istogramma che si ottiene da queste distribuzioni assume una forma di una campana rovesciata.

La curva a campana è la curva teorica che rappresenta la funzione densità di probabilità dei fenomeni osservati.

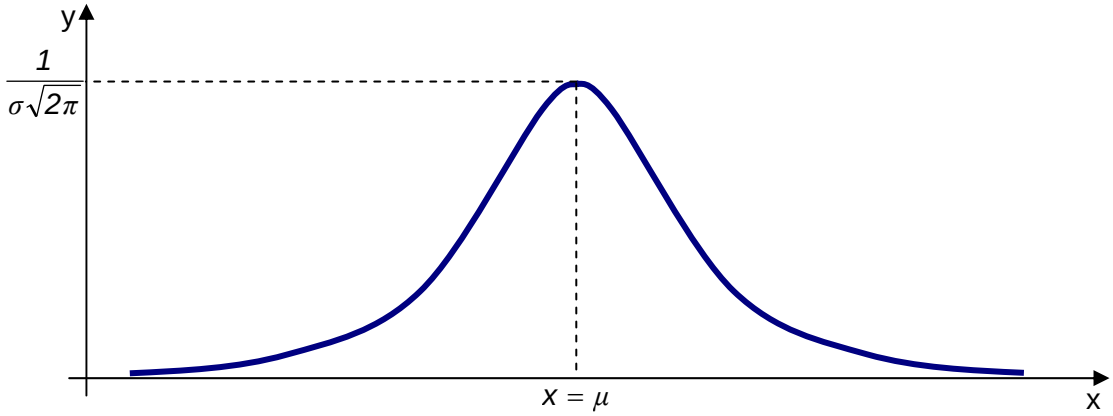


La sua equazione è data dalla formula : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}$

dove $\mu = media$ e $\sigma = scarto\ quadratico\ medio.$

La curva normale o gaussiana ha le seguenti caratteristiche:

- è simmetrica rispetto alla retta $x = \mu$
- assume valore massimo in $x = \mu$ e il suo valore è $\frac{1}{\sigma\sqrt{2\pi}}$
- ha come asintoto orizzontale l'asse delle x
- l'area sottesa tra la curva e l'asse x vale 1.
- l'area tra le due ordinate a e b rappresenta la probabilità che x sia compreso tra a e b $p(a < x \leq b)$.



Per valutare la probabilità che un elemento a caso della distribuzione abbia un valore compreso in una determinata classe $p(a < x \le b)$, occorre calcolare l'area della parte di piano racchiusa dalla curva, dall'asse delle x e dalle rette di equazione $x = a$ e $x = b$, (*con* a *e* b *estremi della classe*).

Occorre cioè calcolare l'espressione della funzione di ripartizione determinando il valore del seguente integrale:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}} dt$$

Poiché tale funzione non è precisabile analiticamente (*non si riesce infatti a calcolare una primitiva di* $f(x)$), occorre far ricorso a tecniche di approssimazione.

Tuttavia non si può pensare di compilare tavole con i valori approssimati delle aree che rappresentano le varie probabilità per ogni valore dei parametri μ e σ.

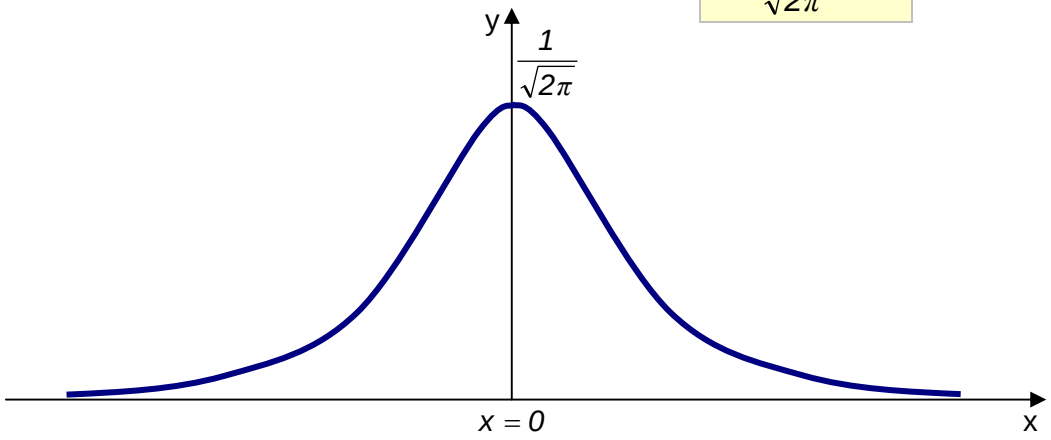
Si ricorre allora ad una particolare trasformazione che consente di ricondurre qualsiasi distribuzione normale di media μ e deviazione standard σ ad una distribuzione normale di media 0 e deviazione 1.

Per fare in modo che la media sia uguale a 0, basta operare la traslazione: $\begin{cases} x' = x - \mu \\ y' = y \end{cases}$.

Per fare in modo che la deviazione standard σ sia uguale ad 1, si deve operare una omotetia di rapporto $k = \frac{1}{\sigma}$.

In definitiva, combinando le due trasformazioni si ottiene la seguente sostituzione: $z = \frac{x - \mu}{\sigma}$ (*variabile standardizzata*).

L'equazione della curva gaussiana standardizzata che si ottiene è la seguente: $y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$.



I valori della funzione di ripartizione relativi alla gaussiana standardizzata sono stati calcolati una volta per tutte e sono riportati in qualunque manuale di statistica.

Osservazione

Se n è molto grande e sia p e sia q sono lontani dallo zero, la distribuzione binomiale può essere rappresentata da una distribuzione gaussiana, effettuando la sostituzione: $z = \frac{x - np}{\sqrt{n \cdot p \cdot q}}$.

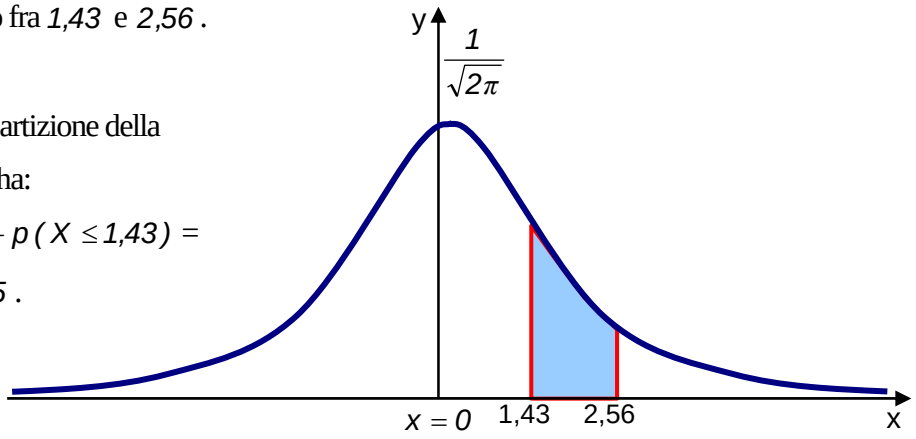
Al crescere di n l'approssimazione diminuisce e per n prossimo a infinito, le due distribuzioni coincidono.

Nel caso della distribuzione gaussiana si hanno le seguenti formule:

Media	*Varianza*	*s. q. m.*	*asimmetria*	*curtosi*	*Scostamento semplice medio*
$M = n \cdot p$	$\sigma^2 = n \cdot p \cdot q$	$\sigma = \sqrt{n \cdot p \cdot q}$	$a_3 = 0$	$a_4 = 3$	$0{,}9797 \cdot \sigma$

Esempio 1

Una variabile aleatoria X segue una distribuzione normale con media $\mu = 0$ e deviazione standard $\sigma = 1$. Calcolare la probabilità che un valore x sia compreso fra $1{,}43$ e $2{,}56$.



Soluzione

Utilizzando le tavole della funzione di ripartizione della distribuzione gaussiana standardizzata si ha:

$p(1{,}43 < X \le 2{,}56) = p(X \le 2{,}56) - p(X \le 1{,}43) =$

$= 0{,}994766 - 0{,}923641 = 0{,}0711125$.

Esempio 2

Le altezze di una popolazione di uomini seguono strettamente una distribuzione normale con valore atteso $\mu = 168{,}75\ cm$ e deviazione standard $\sigma = 6{,}25\ cm$. Calcolare le probabilità di avere individui:

A. che superano $176{,}25\ cm$ di altezza
B. che siano al di sotto di $167{,}5\ cm$
C. che siano al di sotto di $180\ cm$
D. che abbiano un'altezza compresa tra $162{,}5\ cm$ e $170\ cm$

Stimare inoltre, in una popolazione di 2000 individui il numero di quelli che appartengono alle classi dei punti A, B e C.

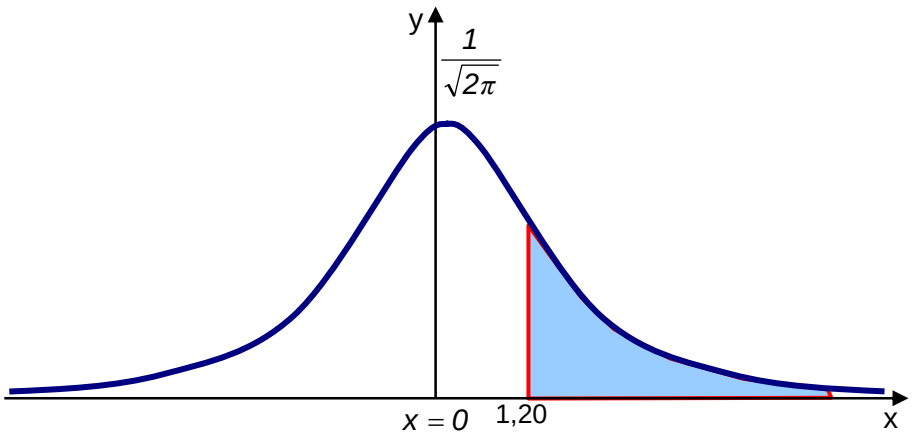
Soluzione A

Occorre trasformare la variabile X, che indica le altezze, nella variabile standardizzata, operando la trasformazione:

$z = \dfrac{x - \mu}{\sigma} = \dfrac{176{,}5 - 168{,}75}{6{,}25} = 1{,}20$ da cui:

$p(z > 1{,}20) = 1 - p(z \le 1{,}20) =$

$= 1 - 0{,}884930 = 0{,}11507$.



Soluzione B

Occorre trasformare la variabile X, che indica le altezze, nella variabile standardizzata, operando la trasformazione:

$z = \dfrac{x - \mu}{\sigma} = \dfrac{167{,}5 - 168{,}75}{6{,}25} = -0{,}20$

Per ragioni di simmetria, la probabilità che z sia minore di $-0{,}20$ è uguale alla probabilità che z sia maggiore di $0{,}20$

$p(z < -0{,}20) = p(z > 0{,}20) = 1 - p(z \le 0{,}20) = 1 - 0{,}579260 = 0{,}42074$.

Soluzione C

Occorre trasformare la variabile X, che indica le altezze, nella variabile standardizzata, operando la trasformazione:

$z = \dfrac{x - \mu}{\sigma} = \dfrac{180 - 168{,}75}{6{,}25} = 1{,}80$

$p(z < 1{,}80) = 0{,}964070$.

Soluzione D

Occorre trasformare la variabile X, che indica le altezze, nella variabile standardizzata, operando la trasformazione:

Per $x = 162{,}5$ si ha: $z = \frac{x-\mu}{\sigma} = \frac{162{,}5 - 168{,}75}{6{,}25} = -1$

Per $x = 170$ si ha: $z = \frac{x-\mu}{\sigma} = \frac{170 - 168{,}75}{6{,}25} = -2$

Pertanto: $p(-1 < z \le 0{,}2) = p(z \le 0{,}2) - p(z \le -1) = p(z \le 0{,}2) - p(z \ge 1) = p(z \le 0{,}2) - [1 - p(z < 1)] =$
$= p(z \le 0{,}2) - 1 + p(z < 1) = 0{,}579260 - 1 + 0{,}841345 = 0{,}420605$.

Per stimare infine, il numero di individui che appartengono alle varie classi, basta moltiplicare la numerosità della popolazione per il rispettivo valore di probabilità.
Pertanto su una popolazione di 2000 individui:

A. il numero stimato di individui che hanno un'altezza superiore a $176{,}25\ cm$ è: $0{,}11507 \cdot 2000 = 230$.
B. il numero stimato di individui che hanno un'altezza inferiore a $167{,}5\ cm$ è: $0{,}42074 \cdot 2000 = 841$.
C. il numero stimato di individui che hanno un'altezza inferiore a $180\ cm$ è: $0{,}964070 \cdot 2000 = 1928$.
D. il numero stimato di individui che hanno un'altezza compresa tra $162{,}5\ cm$ e $170\ cm$ è: $0{,}420605 \cdot 2000 = 841$.

Distribuzione multinomiale

Se n eventi $E_1, E_2, \ldots E_k$ hanno probabilità $p_1, p_2, \ldots p_k$ di presentarsi (dove $x_1, x_2, \ldots x_k = n$), è possibile calcolare la probabilità che detti eventi si presentino rispettivamente $x_1, x_2, \ldots x_n$ volte:	$p_{tot} = \frac{n!}{x_1! \cdot x_2! \cdots x_k!} \cdot {p_1}^{x_1} \cdot {p_2}^{x_2} \cdots {p_k}^{x_k}$

Esempio 1

Lanciando un dado 12 volte, la probabilità di ottenere 2 volte il numero 1, 2 volte il numero 2, 2 volte il numero 3, 2 volte il numero 4, 2 volte il numero 5, 2 volte il numero 6 è:

$$p_{tot} = \frac{n!}{x_1! \cdot x_2 \cdots x_k} \cdot {p_1}^{x_1} \cdot {p_2}^{x_2} \cdots {p_k}^{x_k} = \frac{12!}{2! \cdot 2! \cdot 2! \cdot 2! \cdot 2! \cdot 2!} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 = 0{,}00344$$

Esempio 2

I partecipanti a un esame hanno ottenuto un voto medio di $6{,}7\ /\ 10$ con scarto quadratico medio di $1{,}2$. Ipotizzando una distribuzione normale, calcolare:

A. la percentuale di studenti che ha ottenuto la sufficienza
B. il voto massimo del peggior 10% della classe
C. il voto minimo del miglior 10% della classe

Soluzione A

Nonostante che i voti siano discreti (numeri interi), occorre trattarli come dati continui per poter applicare la distribuzione normale.
Di conseguenza la sufficienza è rappresentata da un voto compreso tra $5{,}5$ e $6{,}5$ (*e non 6*).
Trasformiamo questi dati in unità standard:

$$x_{5,5} = \frac{x - M}{s} = \frac{5{,}5 - 6{,}7}{1{,}2} = -1 \qquad x_{6,5} = \frac{x - M}{s} = \frac{6{,}5 - 6{,}7}{1{,}2} = -0{,}17$$

Per calcolare l'area compresa tra -1 e $-0{,}17$, si considera che a causa della simmetria della curva questa area è pari a quella tra $0{,}17$ e 1, quindi consultando la tavola delle aree della curva normale standardizzata, si trova che l'area tra 0 e 1 è $0{,}3413$, mentre l'area compresa tra 0 e $0{,}17$ è $0{,}0675$.

Di conseguenza l'area cercata è pari a $0{,}3413 - 0{,}0675 = 0{,}2738$, pari quindi al 27% dell'area totale sotto la curva (pari a 1). Quindi il 27% degli studenti ha ottenuto la sufficienza.

Soluzione B

Si deve cercare sulle tavole il valore x_i in corrispondenza del quale l'area (tra 0 e x) vale $0{,}4$. Il valore più prossimo è $1{,}28$.

Considerando le coordinate standardizzate si ha: $0{,}1 = \dfrac{x_1 - \bar{x}}{s} = \dfrac{6{,}5 - 6{,}7}{1{,}2} = -0{,}128$

da cui il voto cercato risulta 5.

Soluzione C

Per la simmetria della curva, si trova: $0{,}1 = \dfrac{x_1 - \bar{x}}{s} = \dfrac{6{,}5 - 6{,}7}{1{,}2} = 0{,}128$

da cui il voto cercato risulta $8{,}2$ approssimato al valore 8, dato che i voti sono numeri interi.

Campioni e campionamento

La teoria dei campioni studia la significatività dei dati campionari ed entro quali limiti essi siano applicabili all'intera popolazione.

Essa si occupa di stimare determinate grandezze (per es., la media, la varianza ecc.) riferite a una popolazione quando siano note quelle campionarie, oppure di determinare entro quali limiti siano significative le differenze riscontrate tra campioni diversi della stessa popolazione (ovvero in quale parte queste differenze siano da attribuire al caso).

Alla teoria dei campioni è correlata l'*inferenza statistica*, ovvero lo studio delle inferenze di una popolazione ottenute mediante suoi campioni. L'inferenza statistica si occupa ovviamente anche dell'accuratezza di queste inferenze.

Per rendere accettabili i risultati ottenuti dai campioni, questi devono essere rappresentativi della popolazione. La maniera migliore per far ciò è estrarli a caso (purché ogni membro della popolazione abbia la stessa probabilità di essere incluso nel campione). Questo processo prende il nome di *campionamento casuale* e corrisponde alla classica estrazione di un bigliettino o di un numero da un'urna.

Quando noi estraiamo un elemento del campione, possiamo scegliere di escluderlo o meno dalla popolazione. Nel primo caso questo elemento potrà entrare a far parte del campione più di una volta (***campionamento con ripetizione***), mentre nel secondo caso no (***campionamento senza ripetizione***).

Le popolazioni dalle quali estraiamo il campione possono essere finite (per es., l'estrazione di un numero della tombola, dato che i numeri sono 90) o infinite (per es., il lancio di una moneta: possono solo verificarsi i due eventi testa/croce, ma il numero di lanci è teoricamente illimitato).

Una popolazione finita può essere considerata infinita se vi si compie un campionamento con ripetizione. Per esempio, se si estraggono i numeri della tombola ma si rimette il numero estratto nel sacchetto, si possono effettuare infinite estrazioni anziché solo 90.

Tipi di distribuzioni campionarie

Per ogni campione che consideriamo, noi possiamo calcolare determinati parametri (media, varianza ecc.); questi sono dunque detti parametri campionari e la loro distribuzione prende il nome di *distribuzione del parametro campionario*.

Nel caso di una popolazione di n_p elementi. Indicando con μ e σ rispettivamente la media e lo scarto quadratico medio della popolazione e con $\mu_{\bar{x}}$ e $\sigma_{\bar{x}}$ rispettivamente la media e lo scarto quadratico medio della distribuzione della media campionaria, si ha:

Se il campionamento è effettuato ***senza ripetizioni*** (*in blocco*), la media e lo *s.q.m.* sono:	$\begin{cases} \mu_{\bar{x}} = \mu \\ \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{n_p - n}{n_p - 1}} \end{cases}$
Se il campionamento è effettuato ***con ripetizione***, i limiti di confidenza sono:	$\begin{cases} \mu_{\bar{x}} = \mu \\ \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} \end{cases}$

Per n molto grande ($n > 30$), la distribuzione della media campionaria tende a una distribuzione normale di media $\mu_{\bar{x}}$ e scarto quadratico $\sigma_{\bar{x}}$.

Nel caso di una popolazione infinita e un evento di probabilità di successo p e di insuccesso q, la proporzione dei successi relativa a ogni campione di ampiezza n genera come risultato la distribuzione delle proporzioni campionarie, che ha come media $\mu_{\bar{x}}$ e come scarto quadratico $\sigma_{\bar{x}}$:

Se il campionamento è effettuato ***senza ripetizioni*** (*in blocco*), la media e lo *s.q.m.* sono:	$\begin{cases} \mu_{\bar{x}} = p \\ \sigma_{\bar{x}} = \dfrac{n_P - n}{n_P - 1} \cdot \sqrt{\dfrac{p \cdot q}{n}} \end{cases}$
Se il campionamento è effettuato ***con ripetizione***, i limiti di confidenza sono:	$\begin{cases} \mu_{\bar{x}} = p \\ \sigma_{\bar{x}} = \sqrt{\dfrac{p \cdot q}{n}} \end{cases}$

Distribuzione delle differenze tra campioni

Consideriamo due popolazioni: della prima calcoliamo una certa statistica S_1 per ciascun campione di ampiezza n_1 da essa estratto. Calcoliamo inoltre la distribuzione campionaria della statistica, che avrà media e scarto quadratico medio μ_{S_1} e σ_{S_1}.

Procediamo nella stessa maniera con la seconda popolazione, ottenendo i relativi dati.

Definiamo distribuzione delle differenze delle statistiche campionarie la distribuzione delle differenze $S_1 - S_2$, ottenuta da tutte le possibili combinazioni dei campioni (supposti tra loro indipendenti) estratti dalle due popolazioni.

La media e lo scarto quadratico sono: $\begin{cases} \mu_{S_1 - S_2} = \mu_{S_1} - \mu_{S_2} \\ \sigma_{S_1 - S_2} = \sqrt{\sigma_{S_1}^{\ 2} + \sigma_{S_2}^{\ 2}} \end{cases}$

Se le statistiche S_1 e S_2 sono le medie (ovvero $\bar{x}_1$ e $\bar{x}_2$), questa è la distribuzione delle differenze delle medie campionarie: $\begin{cases} \mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 \\ \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^{\ 2} + \sigma_{\bar{x}_2}^{\ 2}} = \sqrt{\dfrac{\sigma_1^{\ 2}}{n_1} + \dfrac{\sigma_2^{\ 2}}{n_2}} \end{cases}$

Se n_1 e n_2 sono grandi ci si avvicina, come nei casi precedenti, a distribuzioni normali.

Esempio

La ditta A produce diodi con durata media di 10000 ore e scarto quadratico medio 100 ore, mentre i diodi della ditta B hanno durata media 12000 ore di scarto quadratico medio 125 ore. Considerando due campioni di 120 e 140 pezzi rispettivamente, si trovino i limiti di confidenza al 95% e al 99% per la somma delle durate medie delle popolazioni:

Soluzione

I limiti di confidenza per la somma delle durate medie delle intere popolazioni sono dati dalla formula:

$$\overline{x_A} + \overline{x_B} \pm z_C \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

Applicandola nel caso dei limiti al 95% si ha:

$$\overline{x_A} + \overline{x_B} \pm z_C \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = 10000 + 12000 \pm 1{,}96 \sqrt{\frac{100^2}{120} + \frac{125^2}{140}} = 22000 \pm 27{,}36$$

Applicandola nel caso dei limiti al 99 si ha:

$$\overline{x_A} + \overline{x_B} \pm z_C \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = 10000 + 12000 \pm 2{,}58 \sqrt{\frac{100^2}{120} + \frac{125^2}{140}} = 22000 \pm 36{,}02$$

L’errore standard

Viene spesso definito errore standard lo scarto quadratico medio della distribuzione di una statistica campionaria.

Distribuzione campionaria	*Errore standard*		*Distribuzione campionaria*	*Errore standard*
media	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$		mediana	$\sigma_{mdn} = \sigma\sqrt{\frac{\pi}{2n}}$
proporzioni	$\sigma_p = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{p \cdot (1-p)}{n}}$		varianza	$\sigma_s^{\,2} = \sigma^2\sqrt{\frac{2}{n}} = \sqrt{\frac{\mu_1 - \mu_2^{\,2}}{n}}$
s. q. m.	$\sigma_s = \frac{\sigma}{\sqrt{2n}}\sqrt{\frac{\mu_1 - \mu_2^{\,2}}{4n\mu_2}}$		Coefficiente di variazione	$\sigma_v = v\sqrt{\frac{1+2v^2}{2n}}$

Esempio 1

Considerando i numeri $2, 3, 6, 8, 10$, si estraggano tutti i possibili campioni di ampiezza 2 (con ripetizione) e si calcoli:

A. media e scarto quadratico medio della popolazione
B. media e scarto quadratico medio della distribuzione della media campionaria

Soluzione A

La media della popolazione è: $\mu = \frac{2+3+6+8+10}{5} = 5,8$.

Lo s. q. m. della popolazione è: $\sigma = \sqrt{\frac{(2-5,8)^2 + (3-5,8)^2 + (6-5,8)^2 + (8-5,8)^2 + (10-5,8)^2}{5}} =$

$= \sqrt{8,96} = 2,99$

Soluzione B

I 25 $(D'^{5,2} = 5^2)$ campioni possibili sono:	$2, 2$	$2, 3$	$2, 6$	$2, 8$	$2, 10$
	$3, 2$	$3, 3$	$3, 6$	$3, 8$	$3, 10$
	$6, 2$	$6, 3$	$6, 6$	$6, 8$	$6, 10$
	$8, 2$	$8, 3$	$8, 6$	$8, 8$	$8, 10$
	$10, 2$	$10, 3$	$10, 6$	$10, 8$	$10, 10$

Le medie sono:	2	$2,5$	4	5	6
	$2,5$	3	$4,5$	$5,5$	$6,5$
	4	$4,5$	6	7	8
	5	$5,5$	7	8	9
	6	$6,5$	8	9	10

La media delle medie dei 25 campioni rappresenta la media della distribuzione della media. Essa vale $5,8$.

Osservazione

La media della distribuzione della media è uguale alla media della popolazione.

Dati					Somme
2	$2,5$	4	5	6	19,5
$2,5$	3	$4,5$	$5,5$	$6,5$	22
4	$4,5$	6	7	8	29,5
5	$5,5$	7	8	9	34,5
6	$6,5$	8	9	10	39,5
			Somma totale		145
				Media	*5,8*

Lo scarto quadratico medio delle medie dei *25* campioni rappresenta invece lo scarto quadratico medio della distribuzione della media, cioè l'errore standard. Esso vale *2,12*.

Media	**5,8**	
xi	**xi - M**	**(xi - M)^2**
2	-3,8	14,44
2,5	-3,3	10,89
4	-1,8	3,24
5	-0,8	0,64
6	0,2	0,04
2,5	-3,3	10,89
3	-2,8	7,84
4,5	-1,3	1,69
5,5	-0,3	0,09
6,5	0,7	0,49
4	-1,8	3,24
4,5	-1,3	1,69
6	0,2	0,04
7	1,2	1,44
8	2,2	4,84
5	-0,8	0,64
5,5	-0,3	0,09
7	1,2	1,44
8	2,2	4,84
9	3,2	10,24
6	0,2	0,04
6,5	0,7	0,49
8	2,2	4,84
9	3,2	10,24
10	4,2	17,64
Somma (xi - M)^2		112
Somma (xi - M)^2 / n		4,48
s. q. m.		2,12

Osservazione

La varianza della media campionaria $\sigma_{\bar{x}}$ è uguale alla varianza della popolazione diviso l'ampiezza (*2*) del campione.

Infatti: $\sigma = 2{,}99 \qquad \sigma^2 = 2{,}99^2 = 8{,}9401$

$\sigma_{\bar{x}} = 2{,}12 \qquad \sigma_{\bar{x}}^2 = 2{,}12^2 = 4{,}4944$

In generale vale la regola: $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ in accordo con la formula vista precedentemente: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Esempio 2

Considerando i dati dell'esercizio precedente, si estraggano tutti i possibili campioni di ampiezza 2 senza ripetizione e si calcoli media e scarto quadratico medio della distribuzione della media campionaria.

Soluzione

Occorre non considerare le coppie con ripetizione *(2, 2)* e considerare una sola delle coppie uguali *(2, 3) e (3, 2):*

I *10* campioni possibili sono: $C_{5,2} = \frac{5 \cdot 4}{2 \cdot 1}$					
	2, 2	2, 3	2, 6	2, 8	2, 10
	3, 2	3, 3	3, 6	3, 8	3, 10
	6, 2	6, 3	6, 6	6, 8	6, 10
	8, 2	8, 3	8, 6	8, 8	8, 10
	10, 2	10, 3	10, 6	10, 8	10, 10

Le medie dei *10* campioni sono:					
	2	2,5	4	5	6
	2,5	3	4,5	5,5	6,5
	4	4,5	6	7	8
	5	5,5	7	8	9
	6	6,5	8	9	10

La media delle medie dei *10* campioni rappresenta la media della distribuzione della media.

Essa vale $M = \frac{2{,}5 + 4 + 5 + 6 + 4{,}5 + 5{,}5 + 6{,}5 + 7 + 8 + 9}{10} = 5{,}8$.

Anche questo valore coincide con la media della popolazione.

Lo scarto quadratico medio delle medie dei *10* campioni rappresenta invece lo scarto quadratico medio della distribuzione della media, cioè l'errore standard. Esso vale *1,83* .

Media	**5,8**	
xi	**xi - M**	**(xi - M)^2**
2,5	-3,3	10,89
4	-1,8	3,24
5	-0,8	0,64
6	0,2	0,04
4,5	-1,3	1,69
5,5	-0,3	0,09
6,5	0,7	0,49
7	1,2	1,44
8	2,2	4,84
9	3,2	10,24
Somma (xi - M)^2		33,6
Somma (xi - M)^2 / 10		3,36
s. q. m.		1,83

Esempio 3

Considerando i dati dell'esempio 1, si calcolino media e scarto quadratico medio della distribuzione della varianza campionaria.

Soluzione

Le varianze campionarie relative alle *25* coppie:

2, 2	*2, 3*	*2, 6*	*2, 8*	*2, 10*
3, 2	*3, 3*	*3, 6*	*3, 8*	*3, 10*
6, 2	*6, 3*	*6, 6*	*6, 8*	*6, 10*
8, 2	*8, 3*	*8, 6*	*8, 8*	*8, 10*
10, 2	*10, 3*	*10, 6*	*10, 8*	*10, 10*

sono:

0	*0,25*	*4*	*9*	*16*
0,25	*0*	*2,25*	*6,25*	*12,25*
4	*2,25*	*0*	*1*	*4*
9	*6,25*	*1*	*0*	*1*
16	*12,25*	*4*	*1*	*0*

Coppia		Media	x1-M	x2-M	(x1-M)^2	(x2-M)^2	(x1-M)^2 + (x2-M)^2	((x1-M)^2 + (x2-M)^2) / 2
2	2	2	0	0	0	0	0	0
2	3	2,5	-0,5	0,5	0,25	0,25	0,5	0,25
2	6	4	-2	2	4	4	8	4
2	8	5	-3	3	9	9	18	9
2	10	6	-4	4	16	16	32	16
3	2	2,5	0,5	-0,5	0,25	0,25	0,5	0,25
3	3	3	0	0	0	0	0	0
3	6	4,5	-1,5	1,5	2,25	2,25	4,5	2,25
3	8	5,5	-2,5	2,5	6,25	6,25	12,5	6,25
3	10	6,5	-3,5	3,5	12,25	12,25	24,5	12,25
6	2	4	2	-2	4	4	8	4
6	3	4,5	1,5	-1,5	2,25	2,25	4,5	2,25
6	6	6	0	0	0	0	0	0
6	8	7	-1	1	1	1	2	1
6	10	8	-2	2	4	4	8	4
8	2	5	3	-3	9	9	18	9
8	3	5,5	2,5	-2,5	6,25	6,25	12,5	6,25
8	6	7	1	-1	1	1	2	1
8	8	8	0	0	0	0	0	0
8	10	9	-1	1	1	1	2	1
10	2	6	4	-4	16	16	32	16
10	3	6,5	3,5	-3,5	12,25	12,25	24,5	12,25
10	6	8	2	-2	4	4	8	4
10	8	9	1	-1	1	1	2	1
10	10	10	0	0	0	0	0	0

La media delle *25* varianze campionarie è: $M = 4{,}48$.

Dati					Somme
0	*0,25*	*4*	*9*	*16*	29,25
0,25	*0*	*2,25*	*6,25*	*12,25*	21
4	*2,25*	*0*	*1*	*4*	11,25
9	*6,25*	*1*	*0*	*1*	17,25
16	*12,25*	*4*	*1*	*0*	33,25
			Somma totale		112
				Media	*4,48*

Osservazione

La media della distribuzione della varianza campionaria poteva essere calcolata con una formula più facile:

$$\mu_{S^2} = \frac{n-1}{n}\sigma^2 = \frac{2-1}{2}2{,}99^2 = \frac{1}{2}8{,}96 = 4{,}48 .$$

In generale

Se il campionamento è effettuato ***senza ripetizioni*** (*in blocco*), la media e lo *s.q.m.* sono:	$\mu_{S^2} = \frac{n-1}{n} \cdot \frac{n_P}{n_P - 1} \sigma^2$
Se il campionamento è effettuato ***con ripetizione***, i limiti di confidenza sono:	$\mu_{S^2} = \frac{n-1}{n} \sigma^2$

Lo s. q. m. delle *25* varianze campionarie è: *s.q.m. = 5,01* .

Media	***4,48***	
xi	***xi - M***	***(xi - M)^2***
0	*-4,48*	*20,0704*
0,25	*-4,23*	*17,8929*
4	*-0,48*	*0,2304*
9	*4,52*	*20,4304*
16	*11,52*	*132,71*
0,25	*-4,23*	*17,8929*
0	*-4,48*	*20,0704*
2,25	*-2,23*	*4,9729*
6,25	*1,77*	*3,1329*
12,25	*7,77*	*60,3729*
4	*-0,48*	*0,2304*
2,25	*-2,23*	*4,9729*
0	*-4,48*	*20,0704*
1	*-3,48*	*12,1104*
4	*-0,48*	*0,2304*
9	*4,52*	*20,4304*
6,25	*1,77*	*3,1329*
1	*-3,48*	*12,1104*
0	*-4,48*	*20,0704*
1	*-3,48*	*12,1104*
16	*11,52*	*132,71*
12,25	*7,77*	*60,3729*
4	*-0,48*	*0,2304*
1	*-3,48*	*12,1104*
0	*-4,48*	*20,0704*
Somma (xi - M)^2		*628,74*
Somma (xi - M)^2 / n		*25,1496*
s. q. m.		*5,01*

Stime ed errori

Un importante compito dell'inferenza statistica è la determinazione dei parametri della popolazione (media, varianza ecc.) attraverso lo studio dei campioni da essa estratti.

Uno stimatore T, è ***corretto***, o anche ***non distorto***, se il suo valore atteso è uguale al valore del parametro θ che deve stimare. In simboli: $E(T_n) = \theta \quad \forall n \geq 1$.
La differenza $E(T_n) - \theta$ viene detta ***errore sistematico*** o ***distorsione*** dello stimatore.

Uno stimatore non distorto non garantisce stime precise, ma ha il seguente significato: se estraessimo tutti i possibili campioni o comunque un grande numero di essi calcolando le corrispondenti stime, la media di questi valori coinciderebbe o sarebbe molto vicina al vero valore del parametro.

La media della distribuzione della media campionaria $\mu_{\bar{x}}$ ***è uno stimatore corretto della media della popolazione*** μ. *Infatti la media campionaria e la media della popolazione coincidono* ($\mu_{\bar{x}} = \mu$)

La varianza campionaria μ_{S^2} *(media della distribuzione della varianza campionaria)* ***è uno stimatore distorto per la varianza della popolazione*** σ^2. *Infatti* $\mu_{S^2} = \frac{n-1}{n}\sigma^2$

La sua distorsione vale *(nel caso di campionamento bernoulliano)*	$E(S_n^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$

quindi S_n^2 fornisce in media delle sottostime della varianza della popolazione.

La ***varianza campionaria corretta*** $\hat{S}_n^2 = \frac{n}{n-1}S_n^2$ è invece uno stimatore corretto della varianza della popolazione.

In sintesi

Stima corretta ed efficiente	*Stima corretta e inefficiente*	*Stima distorta e inefficiente*
Media campionaria	Mediana	s. q. .m. campionario
Varianza campionaria corretta		s. q. .m. campionario corretto
		Scostamento semplice medio assoluto

Considerando la distribuzione di due statistiche con ugual media, quella con varianza minore è detta *stimatore efficiente* della media e i suoi valori sono detti *stime efficienti*.

Siccome le distribuzioni della media e della mediana campionaria hanno ugual media, e dato che la varianza della distribuzione della media è minore, questa è uno stimatore efficiente della media della popolazione.

La stima di un parametro della popolazione può essere effettuata in due modi:

- stima puntuale (*la stima è costituita da un solo numero*)
- stima per intervallo (*la stima è costituita da due numeri, estremo inferiore e superiore dell'intervallo, oppure da un numero e dall'ampiezza dell'intervallo su di esso centrato*)

Le stime puntuali

Esistono diversi metodi di deduzione degli stimatori dei parametri di una popolazione; quello più naturale è quello di considerare come stimatore del parametro l'analogo parametro campionario: la media del campione per la media della popolazione, la varianza del campione per la varianza della popolazione, la mediana del campione per la mediana della popolazione e così via.

Le stime per intervallo

Molto spesso è più interessante conoscere un intervallo in cui si sa che cade il valore vero del parametro con una certa probabilità. Ad esempio potrebbe essere più significativo stabilire che il valor medio della popolazione appartiene all'intervallo *(2,3; 2,5)* con una probabilità del *95%*, piuttosto che sapere che una sua stima puntuale è *2,397* con un errore del *6%*. Si vuole pertanto determinare l'intervallo cui appartiene il valore di un parametro della popolazione con una probabilità stabilita a priori.

Consideriamo una statistica S, con media μ_S, e scarto quadratico medio σ_S.

Se il campione è grande ($n \geq 30$), la distribuzione di S può essere considerata normale.

Si può dunque prevedere di trovare un valore di S che cada nell'intervallo $\mu_S \pm \sigma_S$, nel 68,27% dei casi, nell'intervallo $\mu_S \pm 2\sigma_S$ nel 95,45% dei casi e nell'intervallo $\mu_S \pm 3\sigma_S$ nel 99,73% dei casi.

Questi intervalli sono detti intervalli di confidenza e i loro estremi sono detti *limiti di confidenza* o *limiti fiduciari*.

Nella tabella sono riassunti i valori delle ampiezze dell'intervallo corrispondenti a livelli di confidenza usati frequentemente:

Intervalli di confidenza										
Livello %	*99,73*	*99*	*98*	*96*	*95,45*	*95*	*90*	*80*	*68,27*	*50,00*
Ampiezza z_C	*3,00*	*2,58*	*2,33*	*2,05*	*2,00*	*1,96*	*1,645*	*1,28*	*1*	*0,6745*

Stima per intervallo della media campionaria

Note la media μ e lo *s.q.m.* σ di una popolazione (supposta distribuita normalmente), l'intervallo di confidenza in cui si può trovare la media di un campione di ampiezza n (*media campionaria*), è dato da:

Se il campionamento è effettuato ***senza ripetizione*** (*in blocco*), i limiti di confidenza sono:	$\mu \pm z_C \frac{\sigma}{\sqrt{n}} \sqrt{\frac{n_p - n}{n_p - 1}}$
Se il campionamento è effettuato ***con ripetizione***, i limiti di confidenza sono:	$\mu \pm z_C \frac{\sigma}{\sqrt{n}}$

dove z_C è l'ampiezza ricavabile dalla tabella precedente (*o dalla tabella dell'area della curva normale, se non comparissero in tabella*).

Problemi di questo tipo si incontrano quando si effettuano controlli di qualità sulla produzione aziendale: in questi casi sono di solito note le caratteristiche degli impianti di produzione ed è quindi possibile risalire alla media e alla varianza (o allo scarto quadratico medio) della popolazione dei pezzi prodotti e si vuole determinare in quale intervallo deve essere compresa la media di un campione per considerare "buona" la produzione ad un certo livello di confidenza.

Esempio 1

La media dei voti di 50 studenti è 75/100, con scarto quadratico medio di 15. Calcolare:

A. I limiti di confidenza al 95% della popolazione di 200 studenti
B. Sotto quale grado di confidenza la media è compresa tra 74 e 76

Soluzione A

I limiti di confidenza al 95% valgono: $\mu \pm z_C \frac{\sigma}{\sqrt{n}} \sqrt{\frac{n_p - n}{n_p - 1}} = 75 \pm 1{,}96 \frac{15}{\sqrt{50}} \sqrt{\frac{200 - 50}{200 - 1}} = 75 \pm 3{,}6$.

Soluzione B

Imponendo che l'intervallo di confidenza sia l'intervallo $(74,76)$, cioè che i limiti distano dalla media 75 del valore ± 1, si ottiene:

$\mu \pm z_C \frac{\sigma}{\sqrt{n}} \sqrt{\frac{n_p - n}{n_p - 1}} = 75 \pm 1$ cioè: $75 \pm z_C \frac{15}{\sqrt{50}} \sqrt{\frac{200 - 50}{200 - 1}} = 75 \pm 1$; $75 \pm 1{,}84\, z_C = 75 \pm 1$.

Dalla quale si ricava: $1{,}84\, z_C = 1$; $z_C = 0{,}54$.

Esempio 2

Una lunghezza viene misurata con scarto quadratico medio di $0{,}10\ m$. Se vogliamo essere confidenti al 95% e al 99% che l'errore non superi il centimetro, quanto deve essere ampio il campione considerato?

Soluzione

Trasformando $1\ cm = 0{,}01\ m$ ed applicando la formula : $\mu \pm z_C \frac{\sigma}{\sqrt{n}} = \mu \pm 0{,}01$ (*la misura viene ripetuta*) si ha:

un errore nella stima, al 95%, se: $z_C \frac{\sigma}{\sqrt{n}} = 0{,}01$; $1{,}96 \frac{0{,}1}{\sqrt{n}} = 0{,}01$; $\frac{0{,}196}{\sqrt{n}} = 0{,}01$; $\sqrt{n} = \frac{0{,}196}{0{,}01}$;
$\sqrt{n} = 19{,}6$; $n = 19{,}6^2$; $n = 384{,}16$. Ciò significa che il campione dovrà contenere almeno 385 misurazioni.

un errore nella stima, al 99%, se: $z_C \frac{\sigma}{\sqrt{n}} = 0{,}01$; $2{,}58 \frac{0{,}1}{\sqrt{n}} = 0{,}01$; $\frac{0{,}258}{\sqrt{n}} = 0{,}01$; $\sqrt{n} = \frac{0{,}258}{0{,}01}$;
$\sqrt{n} = 25{,}8$; $n = 25{,}8^2$; $n = 665{,}64$. Ciò significa che il campione dovrà contenere almeno 667 misurazioni.

Stima per intervallo della media di una popolazione

Note la media μ e lo *s.q.m.* s di un campione (supposto distribuito normalmente), l'intervallo di confidenza in cui si può trovare la media della popolazione, è dato da:

Se il campionamento è effettuato ***senza ripetizione*** (*in blocco*), i limiti di confidenza sono:	$\mu \pm z_C \frac{s}{\sqrt{n-1}} \sqrt{1-\frac{n}{n_p}}$
Se il campionamento è effettuato ***con ripetizione***, i limiti di confidenza sono:	$\mu \pm z_C \frac{s}{\sqrt{n-1}}$

dove z_C è l'ampiezza ricavabile dalla tabella precedente (*o dalla tabella dell'area della curva normale, se non comparissero in tabella*).

Problemi di questo tipo si incontrano di frequente nelle indagini statistiche, perchè spesso si conoscono solo i parametri del campione e non quelli della popolazione.

In questo caso, supponendo un campionamento bernoulliano, non si può calcolare il valore di $\frac{\sigma}{\sqrt{n}}$ ma, se il campione è sufficientemente grande, si può stimare il valore di σ con la varianza corretta del campione.

Esempio 1

I pesi di un campione di 300 pezzi prodotti da un macchinario hanno media di 0,649 grammi e uno scarto quadratico medio di 0,052 grammi. Trovare i limiti di confidenza al 99% e al 95% del peso medio della popolazione.

Soluzione

Dalla tabella:

Livello %	*99,73*	*99*	*98*	*96*	*95,45*	*95*	*90*	*80*	*68,27*	*50,00*
Ampiezza	*3,00*	*2,58*	*2,33*	*2,05*	*2,00*	*1,96*	*1,645*	*1,28*	*1*	*0,6745*

I limiti di confidenza al 99 % valgono:

$$\mu \pm z_C \frac{s}{\sqrt{n-1}} = 0{,}649 \pm 2{,}58 \frac{0{,}052}{\sqrt{300-1}} = 0{,}649 \pm 0{,}0078 .$$

I limiti di confidenza al 95 % valgono:

$$\mu \pm z_C \frac{s}{\sqrt{n-1}} = 0{,}649 \pm 1{,}96 \frac{0{,}052}{\sqrt{300-1}} = 0{,}649 \pm 0{,}0059 .$$

Stima delle proporzioni

Nel caso in cui la statistica S sia la proporzione di successi in un campione di ampiezza n estratto da una popolazione di tipo binomiale, la quale ha proporzione di successi pari a p, i limiti di confidenza di p valgono $P \pm z_C \sigma_{\bar{x}}$, dove P è la proporzione di successi del campione di ampiezza n.

se il campionamento è effettuato senza ripetizione i limiti di confidenza valgono:	$P \pm z_C \sqrt{\frac{p \cdot q}{n}} \sqrt{\frac{n_p - n}{n_p - 1}}$
se il campionamento è effettuato con ripetizione i limiti di confidenza valgono:	$P \pm z_C \sqrt{\frac{p \cdot q}{n}}$

Esempio 1

Lanciando una monetina, si è ottenuto il punteggio "croce" per 22 volte su 50 lanci. Trovare i livelli di confidenza al 95% e al 98% nel caso in cui si potesse compiere un numero illimitato di lanci.

Soluzione

Dato il numero infinito di lanci, si può applicare la formula relativa al campionamento effettuato con ripetizione.

Dalla tabella:

Livello %	*99,73*	*99*	*98*	*96*	*95,45*	*95*	*90*	*80*	*68,27*	*50,00*
Ampiezza	*3,00*	*2,58*	*2,33*	*2,05*	*2,00*	*1,96*	*1,645*	*1,28*	*1*	*0,6745*

I limiti al 95% valgono: $P \pm z_C \sqrt{\frac{pq}{n}} = \frac{22}{50} \pm 1{,}96 \sqrt{\frac{0{,}5 \cdot 0{,}5}{50}} = 0{,}44 \pm 0{,}14$

I limiti al 98% valgono: $P \pm z_C \sqrt{\frac{pq}{n}} = \frac{22}{50} \pm 2{,}33 \sqrt{\frac{0{,}5 \cdot 0{,}5}{50}} = 0{,}44 \pm 0{,}16$

Stima di somme e differenze

Consideriamo due statistiche S_1 e S_2 con distribuzione normale (o approssimativamente normale); se consideriamo due campioni indipendenti, i limiti di confidenza:

della somma dei parametri della popolazione valgono:	$S_1 + S_2 \pm z_c \sigma_{S_1+S_2} = S_1 + S_2 \pm z_c \sqrt{\sigma_{S_1}^{\ 2} + \sigma_{S_2}^{\ 2}}$
della differenza dei parametri della popolazione valgono:	$S_1 - S_2 \pm z_c \sigma_{S_1-S_2} = S_1 - S_2 \pm z_c \sqrt{\sigma_{S_1}^{\ 2} + \sigma_{S_2}^{\ 2}}$

Stima dello scarto quadratico medio

Una volta stimato lo scarto quadratico medio σ di una popolazione per mezzo di quello (s) di un suo campione, i suoi limiti di confidenza sono: $s \pm z_c \sigma_C = s \pm z_c \frac{\sigma}{\sqrt{2n}}$

Esempio 1

Trovare i limiti di confidenza al 95% e al 99% per l'intera produzione di una ditta di lampadine per le quali è stato calcolato uno scarto quadratico medio delle durate pari a 100 ore considerando un campione di 200 pezzi.

Soluzione

I limiti di confidenza al 95% sono: $s \pm z_c \sigma_C = s \pm z_c \frac{\sigma}{\sqrt{2n}} = 100 \pm 1{,}96 \frac{100}{\sqrt{2 \cdot 200}} = 100 \pm 9{,}8$.

I limiti di confidenza al 98% sono: $s \pm z_c \sigma_C = s \pm z_c \frac{\sigma}{\sqrt{2n}} = 100 \pm 2{,}58 \frac{100}{\sqrt{2 \cdot 200}} = 100 \pm 12{,}8$.

Esempio 2

Si misura un pezzo di metallo cinque volte, ottenendo le seguenti lunghezze di $6{,}32, 6{,}33, 6{,}36, 6{,}37$ e $6{,}37$ in centimetri.

A. Si calcoli la media della popolazione con una stima corretta e inefficiente
B. Si calcolino la media e la varianza della popolazione con una stima corretta e efficiente

Soluzione A

Una stima corretta e inefficiente della media della popolazione è la mediana. Essa vale $M_e = 6{,}36$.

Soluzione B

Una stima corretta ed efficiente della media della popolazione è la media, cioè:

$$M = \frac{6{,}32 + 6{,}33 + 6{,}36 + 6{,}37 + 6{,}37}{5} = 5 .$$

Una stima corretta ed efficiente della varianza della popolazione è la varianza corretta, cioè:

$$\bar{s}^2 = \frac{\sum (x_i - M)^2}{n-1} = \frac{(6{,}32 - 6{,}35)^2 + (6{,}33 - 6{,}35)^2 + (6{,}36 - 6{,}35)^2 + (6{,}37 - 6{,}35)^2 + (6{,}37 - 6{,}35)^2}{5-1} =$$

$$= 0{,}00055 .$$