§19 Stima della media di una variabile statistica

Consideriamo una variabile statistica X di cui intendiamo stimare la media μ (su tutta la popolazione). Scegliamo in modo casuale un campione di n individui, cui sono associati altrettanti valori della variabile X, denotati $x_1, \dots, x_n$. È ragionevole stimare μ con la media di questi valori, detta media campionaria.

$$\hat{\mu} = \mu(\text{campione}) = \frac{x_1 + x_2 + \dots + x_n}{n} \qquad \hat{\clubsuit} = \text{“stima di } \clubsuit\text{”}$$

L'obiettivo principale di questo capitolo è quantificare il probabile errore della stima $\hat{\mu}$. Lo strumento di cui ci serviremo è un concetto importante: l'intervallo di confidenza, che definiremo nella prossima sezione.

Notazione
In questo capitolo μ e σ, senza altre indicazioni, sono la media e la deviazione standard di X su tutta la popolazione, che non conosciamo. Le misure calcolate sul campione portano l'indicazione fra parentesi, come $\mu(\text{campione})$. Il cappello indica una stima: $\hat{\mu}$ è una stima di μ e $\hat{\sigma}$ una stima di σ, qualunque sia la formula con cui le calcoliamo.

§19.1 Metodo per campioni grandi

Questo metodo è valido solo per campioni grandi e popolazioni grandi rispetto al campione; ne presenteremo un altro che, al costo di una maggiore complessità, permette di ovviare al vincolo sul campione. Il metodo di questa sezione ci interessa perché è più semplice, permette di capire meglio le idee, e di fare calcoli approssimativi a memoria ricordando la regola del 68–95–99.7. A seconda dei testi e delle applicazioni, si trovano varie regole circa quali campioni siano *grandi*: una soglia ragionevole potrebbe essere sul centinaio di individui. La sostanza, però, è che il metodo della prossima sezione è sempre più affidabile, quindi, nel dubbio, conviene usare quello.

Si stima la deviazione standard σ di X con la σ del campione

$$\hat{\sigma} = \sigma(\text{campione}) = \sqrt{\frac{(x_1 - \hat{\mu})^2 + \dots + (x_n - \hat{\mu})^2}{n}}$$

Vale il fatto seguente.

Fatto (teorema centrale del limite)
Al variare di tutti i possibili campioni di n individui, le possibili medie campionarie si distribuiscono in modo circa gaussiano (indipendentemente dalla distribuzione di X, a patto che il campione sia abbastanza grande, e la distribuzione di X non abbia varianza *infinita*). La media di questa gaussiana è μ e la deviazione standard

è l'errore standard della media

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{n}} \qquad \textit{“standard error of the mean”}$$

Non sappiamo esattamente σ — va da sé: non sappiamo neanche μ! — però possiamo stimarla con

$$\widehat{\mathrm{SEM}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

Finalmente, siccome la distribuzione delle medie campionarie è gaussiana e ne abbiamo stimato la deviazione standard, possiamo calcolare con quale probabilità la nostra particolare media campionaria $\hat{\mu}$ si trovi in un dato intervallo attorno alla media vera μ. Per esempio,

$\hat{\mu}$ è in $\mu \pm \widehat{\mathrm{SEM}}$	nel 68% dei casi circa
$\hat{\mu}$ è in $\mu \pm 2\,\widehat{\mathrm{SEM}}$	nel 95% dei casi circa
$\hat{\mu}$ è in $\mu \pm 3\,\widehat{\mathrm{SEM}}$	nel 99.7% dei casi circa

Questi valori sono approssimati: cifre più precise si possono facilmente estrarre dalla tavola della gaussiana; per esempio al 95% corrisponde più precisamente l'intervallo $\pm 1.96\,\widehat{\mathrm{SEM}}$.

Per concludere, si osserva che dire

$$a \text{ è in } b \pm \delta \qquad \text{equivale a} \qquad b \text{ è in } a \pm \delta$$

infatti entrambe le frasi dicono che a e b non distano più di δ, quindi

μ è in $\hat{\mu} \pm \widehat{\mathrm{SEM}}$	nel 68% dei casi circa
μ è in $\hat{\mu} \pm 2\,\widehat{\mathrm{SEM}}$	nel 95% dei casi circa
μ è in $\hat{\mu} \pm 3\,\widehat{\mathrm{SEM}}$	nel 99.7% dei casi circa

Questi sono detti intervalli di confidenza e le percentuali sono i rispettivi livelli di confidenza. Valori tipici sono

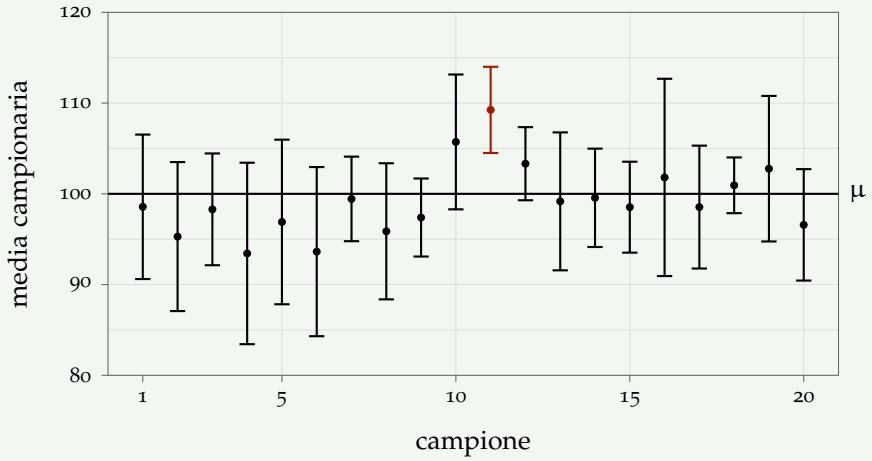
l'intervallo di confidenza al 95% è $\hat{\mu} \pm 1.96\,\widehat{\mathrm{SEM}} \approx \hat{\mu} \pm 2\,\widehat{\mathrm{SEM}}$

l'intervallo di confidenza al 99% è $\hat{\mu} \pm 2.58\,\widehat{\mathrm{SEM}}$

Nota importante
Un livello di confidenza, diciamo, del 95% ci dice che ripetendo molte volte il campionamento, circa il 95% degli intervalli costruiti dai vari campioni ottenuti conterrà il vero valore del parametro. La casualità è quindi quella della scelta del campione. Ho visto applicare questo metodo all'intera popolazione — niente di più stupido: in quel caso la media della popolazione era un valore esatto, non una stima!

Nota

Ecco che cosa succede ripetendo il campionamento. Da una popolazione molto numerosa, in cui una grandezza ha distribuzione gaussiana con media $\mu = 100$ e deviazione standard $\sigma = 10$, ho estratto (con una simulazione al computer) venti campioni di dieci individui, e per ciascuno ho calcolato l'intervallo di confidenza al 95%.



La linea orizzontale rappresenta la media della popolazione. La media campionaria (il pallino) e l'ampiezza dell'intervallo cambiano da un campione all'altro. Come dobbiamo aspettarci, il 95% degli intervalli, 19 su 20, contiene la media vera; uno la manca. In realtà ho eseguito l'esperimento più volte: ogni tanto gli intervalli che mancano sono due o tre, ogni tanto uno, e ogni tanto nessuno. Solo in media ne manca uno su venti. Questa è una simulazione: la realtà differisce perché hai in mano un solo campione. Inevitabilmente, non puoi sapere se il tuo intervallo è uno dei diciannove buoni, o è quello rosso.

§19.2 Metodo più generale

Il metodo precedente, per campioni piccoli, sottostima l'ampiezza degli intervalli di confidenza, soprattutto perché tratta $\hat{\sigma}$ come se fosse σ: con pochi dati la stima di σ è incerta, e di questa incertezza bisogna tener conto. Il metodo di questa sezione vale anche per campioni piccoli (ma sempre con popolazioni grandi rispetto al campione), a patto che la distribuzione di X non si allontani troppo da una gaussiana. Questa condizione sulla distribuzione è necessaria perché il teorema centrale del limite vale solo per n grande. Comunque, se la distribuzione di X è pazzesca, non c'è metodo che tenga.

Innanzitutto rimpiazziamo $\hat{\sigma} = \sigma(\text{campione})$ con

$$\hat{\sigma} = s(\text{campione}) = \sqrt{\underbrace{\frac{(x_1 - \hat{\mu})^2 + \ldots + (x_n - \hat{\mu})^2}{n-1}}_{s^2(\text{campione}) = \text{varianza campionaria}}}$$

Stimiamo l'errore standard della media con

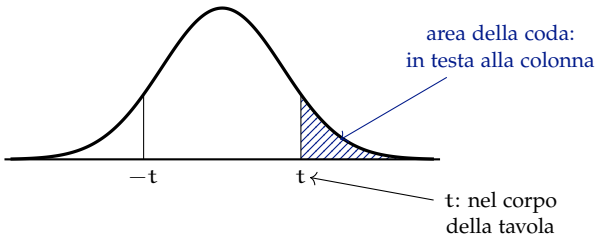
$$\widehat{\mathrm{SEM}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

Ora, la quantità

$$\frac{\hat{\mu} - \mu}{\widehat{\mathrm{SEM}}}$$

si distribuisce, al variare di tutti i possibili campioni, secondo la distribuzione detta t di Student con $n - 1$ gradi di libertà (ce n'è una per ogni numero di gradi di libertà). Se il numero di gradi di libertà è alto, la t di Student coincide, sostanzialmente, con la gaussiana standard. Altrimenti è sottilmente diversa, e queste distribuzioni sono tabulate (vedi l'appendice C).

Ecco un estratto della tavola, con il disegno che ne spiega il significato: per ciascun numero di gradi di libertà (riga) e ciascuna area in testa a una colonna (scritta come numero decimale: 0.025 è il 2.5%), la tavola dà il valore t tale che la coda a destra di t ha proprio quell'area. Siccome la curva è simmetrica, fra $-t$ e t resta il 100% meno il doppio di quell'area: la colonna 0.025 serve per la confidenza al 95%, la colonna 0.005 per la confidenza al 99%.



				conf. 95%		conf. 99%
gradi di libertà	0.25	0.10	0.05	0.025	0.01	0.005
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17

In conclusione, scritto il livello di confidenza come $1 - \alpha$ ($\alpha = 0.05$ per il 95%, $\alpha = 0.01$ per il 99%), e detto $t_{\alpha/2, n-1}$ il valore che la tavola riporta nella colonna $\frac{\alpha}{2}$ alla riga $n - 1$ (il primo indice è l'area della coda, il secondo i gradi di libertà), vale la regola seguente.

Regola (intervallo di confidenza per la media)
L'intervallo di confidenza al livello $1-\alpha$ per la media è

$$\hat{\mu} \pm t_{\alpha/2,\,n-1} \cdot \widehat{\mathrm{SEM}}$$

Al 95%: $\hat{\mu} \pm t_{0.025,\,n-1} \cdot \widehat{\mathrm{SEM}}$. Al 99%: $\hat{\mu} \pm t_{0.005,\,n-1} \cdot \widehat{\mathrm{SEM}}$.

Esempio
Per un campione di $n = 10$ individui, la tavola della t di Student con $9 = 10-1$ gradi di libertà, alla colonna 0.025 (confidenza 95%), dà 2.26. Vediamo quindi che

$$\frac{\hat{\mu}-\mu}{\widehat{\mathrm{SEM}}} \text{ è in } 0 \pm 2.26 \quad \text{nel 95\% dei casi}$$

per cui

$$\mu \text{ è in } \hat{\mu} \pm 2.26\,\widehat{\mathrm{SEM}} \quad \text{nel 95\% dei casi}$$

e questo è il nostro intervallo di confidenza della media (anziché $1.96\,\widehat{\mathrm{SEM}}$ del metodo per campioni grandi).

Esercizio
Delle 36 studentesse che hanno partecipato alla raccolta delle stature (capitolo 17) ne sono state estratte a caso 10, con le stature seguenti, in centimetri.

167 178 168 165 170 160 159 170 165 160

Stima la statura media delle 36 studentesse con un intervallo di confidenza al 95%.

Soluzione. Con il metodo per campioni grandi:

$$\hat{\mu} = 166.2\,\mathrm{cm} \qquad \sigma(\text{campione}) \approx 5.51\,\mathrm{cm} \qquad \widehat{\mathrm{SEM}} = \frac{5.51\,\mathrm{cm}}{\sqrt{10}} \approx 1.74\,\mathrm{cm}$$

e l'intervallo di confidenza al 95% è $(166.2 \pm 1.96 \cdot 1.74)\,\mathrm{cm} = (166.2 \pm 3.4)\,\mathrm{cm}$.

Con il metodo più generale:

$$s(\text{campione}) \approx 5.81\,\mathrm{cm} \qquad \widehat{\mathrm{SEM}} = \frac{5.81\,\mathrm{cm}}{\sqrt{10}} \approx 1.84\,\mathrm{cm}$$

e, siccome $\frac{\hat{\mu}-\mu}{\widehat{\mathrm{SEM}}}$ è in ± 2.26 nel 95% dei casi,

$$\mu = \hat{\mu} \pm 2.26\,\widehat{\mathrm{SEM}} = (166.2 \pm 4.2)\,\mathrm{cm}$$

In questo caso conosciamo anche la media vera delle 36 stature: $\mu = 164.8\,\mathrm{cm}$, che in effetti cade nell'intervallo. Qui, però, il campione è quasi un terzo della popolazione, che quindi non è grande rispetto al campione: in casi come questo gli intervalli che otteniamo sono più larghi del necessario, cioè prudenti. □

§19.3 Misure ripetute

Un caso importante di stima della media è quello delle misure sperimentali (vedi la teoria degli errori, capitolo 9), quando l'unico modo di conoscere l'incertezza di una misura è ripeterla. Se ripetiamo una medesima misura n volte, possiamo vedere i risultati $x_1, \dots, x_n$ come un campione della popolazione (infinita) costituita da tutte le misure che potremmo fare. In assenza di errori sistematici, la media μ di questa popolazione è il valore vero della grandezza misurata: stimare il valore vero è quindi stimare μ. Apprezza la vacuità dell'affermazione precedente: dire che non ci sono errori sistematici equivale per definizione a dire che la media della popolazione delle misure è il valore della grandezza misurata.

Studiando la teoria degli errori abbiamo imparato a riportare i risultati nella forma $x = \hat{\mu} \pm \delta x$. Ora possiamo tentare una definizione precisa dell'errore assoluto δx:

$$\delta x = \widehat{\mathrm{SEM}} = \frac{s(\text{campione})}{\sqrt{n}}$$

cioè l'intervallo di incertezza è, all'incirca, l'intervallo di confidenza al 68% (esattamente, solo se le misure sono molte). Questa non è l'unica definizione adottata in metrologia: per esempio anche l'intervallo di confidenza al 95% sarebbe una scelta sensata. Se leggiamo $x = x_0 \pm \delta x$ $(k = 2)$, quel valore di k fra parentesi indica un multiplo dell'errore standard: in questo caso ci dice che $\delta x = 2\,\widehat{\mathrm{SEM}}$.

Esercizio
Supponi di misurare dieci volte la medesima lunghezza, ottenendo

42.5 mm	42.6 mm	42.2 mm
42.8 mm	42.4 mm	42.5 mm
42.2 mm	42.6 mm	42.3 mm
42.4 mm		

Che stima della lunghezza puoi riportare?

Soluzione. Otteniamo:

$$\hat{\mu} = 42.45\,\mathrm{mm} \qquad s(\text{campione}) \approx 0.190\,\mathrm{mm} \qquad \widehat{\mathrm{SEM}} = \frac{0.190\,\mathrm{mm}}{\sqrt{10}} \approx 0.060\,\mathrm{mm}$$

Quindi, seguendo la regola di arrotondare l'incertezza ad una cifra significativa

$$x = (42.45 \pm 0.06)\,\mathrm{mm}$$

Osserviamo che, benché le singole misure fossero arrotondate ai decimi di millimetro, il risultato che abbiamo ottenuto si arrotonda al centesimo. □

§19.4 Stima di proporzioni

La stima delle medie si può applicare anche alle variabili statistiche binarie, quelle che ci dicono se un individuo ha o no una certa caratteristica: se fuma, se guarisce, se una compressa è fuori specifica. In questi casi, vogliamo stimare la proporzione p, o la percentuale $p \cdot 100\%$, degli individui che hanno la caratteristica su tutta la popolazione. Se, in un campione di n individui, k la hanno, è ragionevole stimare

$$\hat{p} = \frac{k}{n}$$

Vorremmo associare a questa stima un intervallo di confidenza.

Questo si può vedere come un caso particolare della stima della media. Associamo a ciascun individuo la variabile X che vale 1 se ha la caratteristica e 0 se non la ha: la media μ di X è proprio il totale degli uni fratto il numero di individui nella popolazione, ossia la proporzione p. Possiamo quindi calcolare un intervallo di confidenza per p, stimata da $\hat{p} = \hat{\mu}$, col metodo per campioni grandi che conosciamo.

Per farlo dobbiamo calcolare la σ del campione. Nel campione ci sono k valori uguali a 1 e $n - k$ uguali a 0; usando l'identità $\hat{\sigma}^2 = (x_1^2 + \cdots + x_n^2)/n - \hat{\mu}^2$ (capitolo 17),

$$\hat{\sigma}^2 = \frac{\overbrace{1^2 + \cdots + 1^2}^{k}}{n} - \hat{\mu}^2 = \hat{p} - \hat{p}^2 = \hat{p}\,(1 - \hat{p}) \qquad \hat{\sigma} = \sqrt{\hat{p}\,(1 - \hat{p})}$$

Di conseguenza, stimando p con $\hat{p}$, il metodo per campioni grandi dà

Regola (intervallo di confidenza per una proporzione)

$$\widehat{\mathrm{SEM}} = \sqrt{\frac{\hat{p}\,(1 - \hat{p})}{n}} \qquad p \text{ è in } \hat{p} \pm 1.96\,\widehat{\mathrm{SEM}} \quad \text{nel } 95\% \text{ dei casi}$$

Vale per campioni grandi (in pratica, almeno una decina di individui con la caratteristica e una decina senza) e popolazioni grandi rispetto al campione.

Esercizio
In una sperimentazione, 52 degli 80 pazienti di un campione, trattati con un farmaco, guariscono entro una settimana. Stima, con un intervallo di confidenza al 95%, la percentuale di guarigioni entro una settimana.

Soluzione.

$$\hat{p} = \frac{52}{80} = 0.650 \qquad \widehat{\mathrm{SEM}} = \sqrt{\frac{0.650 \cdot 0.350}{80}} \approx 0.0533 \qquad 1.96 \cdot 0.0533 = 0.1045$$

Arrotondando, $p = 0.65 \pm 0.10$: fra il 55% e il 75% dei pazienti guarisce entro una settimana, al 95% di confidenza. Verifica delle condizioni: 52 guariti e 28 no, entrambi ben più di una decina. □

§19.5 Esercizi

Esercizio 101 (soluzione a p. 311)
I valori assunti da una variabile statistica su un campione di sei individui sono

$$14.1 \quad 7.5 \quad 17.9 \quad 12.2 \quad 12.9 \quad 15.8$$

1. Calcola media, varianza (σ^2) e varianza campionaria (s^2) di questi valori.
2. Con che intervallo di confidenza al 95% possiamo stimare la media della popolazione?

Esercizio 102 (soluzione a p. 312)
Si estrae un campione casuale di dieci individui da una popolazione molto vasta. La variabile statistica X assume, su questi dieci individui, i seguenti valori.

$$28.6 \quad 33.1 \quad 36.5 \quad 36.9 \quad 37.6 \quad 41.7 \quad 43.0 \quad 44.7 \quad 45.9 \quad 51.6$$

1. Calcola media, mediana e differenza interquartile di X sul campione.
2. Se stimiamo la media di X su tutta la popolazione con la media del campione, che intervallo di confidenza al 95% abbiamo?
3. Usando un campione più grande, potremmo determinare la media della popolazione con un intervallo di confidenza al 95% di ± 1.0: quanto dovrebbe essere grande questo campione?

Esercizio 103 (soluzione a p. 313)
Il seguente diagramma ramo-foglia rappresenta i valori assunti da una variabile statistica X su un campione di 100 individui, estratto casualmente da una vasta popolazione.

10	5
11	26
12	268
13	0014555889
14	000122233333334567777888
15	000112234447888899
16	000113333555677788889
17	223455578899
18	112489
19	58

Questi valori sono quindi

$$x_1 = 105 \quad x_2 = 112 \quad x_3 = 116 \quad \ldots \quad x_{99} = 195 \quad x_{100} = 198$$

e, per convenienza, sono già state calcolate le somme

$$x_1 + x_2 + \cdots + x_{100} = 15\,482 \qquad x_1^2 + x_2^2 + \cdots + x_{100}^2 = 2\,428\,902$$

1. Determina media μ(campione), deviazione standard σ(campione), mediana $\tilde{X}$ e differenza interquartile IQR di X sul campione.
2. Rappresenta i dati per mezzo di un istogramma, dividendo l'asse delle ascisse in classi di ampiezza 10, e indicando sull'asse delle ordinate la densità di probabilità.
3. Rappresenta sul medesimo diagramma una distribuzione gaussiana con media μ(campione) e deviazione standard σ(campione).
4. Calcola un intervallo di confidenza al 95% per la media μ di X sulla popolazione.

Esercizio 104 (soluzione a p. 315)
Su un campione di compresse di un lotto si stima il contenuto medio di principio attivo, con l'intervallo di confidenza al 95% da 249 mg a 253 mg. Quali delle affermazioni seguenti sono corrette?

1. Il 95% delle compresse del lotto contiene fra 249 mg e 253 mg di principio attivo.
2. Intervalli costruiti nello stesso modo, a partire da campioni diversi, contengono il contenuto medio vero del lotto nel 95% dei casi.
3. Se si ripetesse il campionamento, il 95% delle nuove medie campionarie cadrebbe fra 249 mg e 253 mg.
4. Con un campione più grande l'intervallo sarebbe, di regola, più stretto.

Esercizio 105 (soluzione a p. 316)
Da un campione grande si ottiene, per la media di una variabile, l'intervallo di confidenza al 95% $\mu = 12.0 \pm 1.0$. Senza rifare i conti sui dati, trova gli intervalli di confidenza al 99% e al 68%.

Esercizio 106 (soluzione a p. 316)
Con 25 misure ripetute di una concentrazione si ottiene l'intervallo di confidenza al 95% (6.4 ± 0.8) mg/L. Circa quante misure servirebbero, con lo stesso strumento, per ottenere un intervallo di ± 0.4 mg/L?

Esercizio 107 (soluzione a p. 316)
Un articolo riporta, per 100 pazienti, "peso medio 72 ± 3 kg", senza precisare che cosa sia il ± 3 kg. Fra quali valori sta, grosso modo, il peso del 95% dei pazienti, se

1. 3 kg è la deviazione standard dei pesi;
2. 72 ± 3 kg è l'intervallo di confidenza al 95% della media?

Supponi che i pesi siano distribuiti in modo circa gaussiano.

Esercizio 108 (soluzione a p. 317)
Su 250 compresse prodotte da una macchina, scelte a caso e controllate, 12 risultano fuori specifica. Stima, con un intervallo di confidenza al 95%, la percentuale di compresse fuori specifica prodotte dalla macchina.

Esercizio 109 (soluzione a p. 317)
Un sondaggio su 1000 persone scelte a caso stima che il 40% della popolazione assuma regolarmente integratori alimentari, con un intervallo di confidenza al 95% di $\pm 3\%$.

1. Verifica che l'intervallo $\pm 3\%$ è corretto.
2. Con 4000 intervistati, e la stessa proporzione nel campione, quale sarebbe l'intervallo?

Esercizio 101 (p. 142)
I valori assunti da una variabile statistica su un campione di sei individui sono

$$14.1 \quad 7.5 \quad 17.9 \quad 12.2 \quad 12.9 \quad 15.8$$

1. Calcola media, varianza (σ^2) e varianza campionaria (s^2) di questi valori.
2. Con che intervallo di confidenza al 95% possiamo stimare la media della popolazione?

Soluzione.

Soluzione della domanda 1. Calcoliamo prima la media,

$$\hat{\mu} = \frac{14.1 + 7.5 + 17.9 + 12.2 + 12.9 + 15.8}{6} = \frac{80.4}{6} = 13.4$$

e poi, per le due varianze, gli scarti dalla media e i loro quadrati.

x_i	14.1	7.5	17.9	12.2	12.9	15.8
$x_i - \hat{\mu}$	0.7	−5.9	4.5	−1.2	−0.5	2.4
$(x_i - \hat{\mu})^2$	0.49	34.81	20.25	1.44	0.25	5.76

I quadrati degli scarti hanno somma 63.00, che dividiamo per $n = 6$ per avere la varianza e per $n - 1 = 5$ per la varianza campionaria:

$$\sigma^2 = \frac{63.00}{6} = 10.5 \qquad s^2 = \frac{63.00}{5} = 12.6$$

Gli scarti si controllano facilmente, perché la loro somma deve essere nulla, e infatti $0.7 - 5.9 + 4.5 - 1.2 - 0.5 + 2.4 = 0$. □

Soluzione della domanda 2. Con sei individui usiamo il metodo più generale, con la t di Student, e calcoliamo la deviazione standard campionaria e l'errore standard della media:

$$s = \sqrt{12.6} = 3.550 \qquad \widehat{\mathrm{SEM}} = \frac{s}{\sqrt{n}} = \frac{3.550}{\sqrt{6}} = 1.449$$

La soglia è $t_{0.025,5} = 2.57$ (tavola in appendice, riga dei $6 - 1 = 5$ gradi di libertà, colonna 0.025), e quindi

$$\mu = \hat{\mu} \pm 2.57 \cdot \widehat{\mathrm{SEM}} = 13.4 \pm 3.7 \qquad \text{cioè fra 9.7 e 17.1, al 95\% di confidenza}$$

Se usassimo il metodo per campioni grandi, con 1.96 al posto di 2.57 e la deviazione standard con n al denominatore, $\sqrt{10.5} = 3.240$, otterremmo 13.4 ± 2.6. L'intervallo sarebbe troppo stretto, perché con sei soli individui la t di Student è ben diversa dalla gaussiana. □

Esercizio 102 (p. 142)
Si estrae un campione casuale di dieci individui da una popolazione molto vasta. La variabile statistica X assume, su questi dieci individui, i seguenti valori.

$$28.6 \quad 33.1 \quad 36.5 \quad 36.9 \quad 37.6 \quad 41.7 \quad 43.0 \quad 44.7 \quad 45.9 \quad 51.6$$

1. Calcola media, mediana e differenza interquartile di X sul campione.
2. Se stimiamo la media di X su tutta la popolazione con la media del campione, che intervallo di confidenza al 95% abbiamo?
3. Usando un campione più grande, potremmo determinare la media della popolazione con un intervallo di confidenza al 95% di ± 1.0: quanto dovrebbe essere grande questo campione?

Soluzione.

Soluzione della domanda 1. Sommando i dieci valori otteniamo la media

$$\hat{\mu} = \frac{28.6 + 33.1 + \ldots + 51.6}{10} = \frac{399.6}{10} = 39.96$$

I dati sono già in ordine crescente, e dato che sono dieci la mediana è la media dei due centrali,

$$\tilde{X} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{37.6 + 41.7}{2} = 39.65$$

Per i quartili la regola dà $k = 1 + 0.25 \cdot 9 = 3.25$ e $k = 1 + 0.75 \cdot 9 = 7.75$, da cui

$$Q_1 = x_{(3)} + 0.25\,(x_{(4)} - x_{(3)}) = 36.5 + 0.25 \cdot 0.4 = 36.6$$
$$Q_3 = x_{(7)} + 0.75\,(x_{(8)} - x_{(7)}) = 43.0 + 0.75 \cdot 1.7 = 44.275$$

e la differenza interquartile è $IQR = 44.275 - 36.6 = 7.675 \approx 7.7$. □

Soluzione della domanda 2. Dieci individui sono pochi, e ci serve il metodo più generale. Cominciamo dagli scarti dalla media e dai loro quadrati,

$x_i - \hat{\mu}$	−11.36	−6.86	−3.46	−3.06	−2.36	1.74	3.04	4.74	5.94	11.64
$(x_i - \hat{\mu})^2$	129.05	47.06	11.97	9.36	5.57	3.03	9.24	22.47	35.28	135.49

e la somma dei quadrati è 408.52, da cui

$$s = \sqrt{\frac{408.52}{9}} = 6.737 \qquad \widehat{\mathrm{SEM}} = \frac{6.737}{\sqrt{10}} = 2.131$$

Con 9 gradi di libertà la soglia della t di Student è $t_{0.025,9} = 2.26$ (tavola in appendice, riga 9, colonna 0.025), per cui

$$\mu = \hat{\mu} \pm 2.26 \cdot \widehat{\mathrm{SEM}} = 40.0 \pm 4.8 \qquad \text{cioè fra } 35.1 \text{ e } 44.8$$

□

Soluzione della domanda 3. Il campione che cerchiamo sarà grande, quindi l'intervallo di confidenza al 95% sarà $\hat{\mu} \pm 1.96 \frac{\sigma}{\sqrt{n}}$, e vogliamo che

$$1.96 \frac{\sigma}{\sqrt{n}} = 1.0 \qquad \Longrightarrow \qquad n = (1.96\,\sigma)^2$$

La deviazione standard σ della popolazione non la conosciamo, ma possiamo usare al suo posto la stima migliore che abbiamo, $s = 6.737$, e otteniamo

$$n = (1.96 \cdot 6.737)^2 = 13.20^2 \approx 174$$

quindi servono circa 200 individui. È una stima grossolana, perché s viene da soli dieci individui, per cui possiamo arrotondare a una sola cifra significativa. Inoltre s è solo una stima: al variare del campione oscilla attorno alla σ vera e, grosso modo, metà delle volte sarà sopra e metà sotto. Per questo, con un campione di 200 individui ci aspettiamo solo una probabilità del 50%, più o meno, di ottenere un intervallo di ± 1.0 o meno. Per avere una ragionevole certezza, invece, sceglierei un campione almeno due volte più grande (per esempio 500 individui). □

Esercizio 103 (p. 142)
Il seguente diagramma ramo-foglia rappresenta i valori assunti da una variabile statistica X su un campione di 100 individui, estratto casualmente da una vasta popolazione.

10	5
11	26
12	268
13	0014555889
14	000122233333344567777888
15	000112234447888899
16	000113333555677788889
17	223455578899
18	112489
19	58

Questi valori sono quindi

$$x_1 = 105 \quad x_2 = 112 \quad x_3 = 116 \quad \dots \quad x_{99} = 195 \quad x_{100} = 198$$

e, per convenienza, sono già state calcolate le somme

$$x_1 + x_2 + \cdots + x_{100} = 15\,482 \qquad x_1^2 + x_2^2 + \cdots + x_{100}^2 = 2\,428\,902$$

1. Determina media μ(campione), deviazione standard σ(campione), mediana $\tilde{X}$ e differenza interquartile IQR di X sul campione.
2. Rappresenta i dati per mezzo di un istogramma, dividendo l'asse delle ascisse in classi di ampiezza 10, e indicando sull'asse delle ordinate la densità di probabilità.

3. Rappresenta sul medesimo diagramma una distribuzione gaussiana con media μ(campione) e deviazione standard σ(campione).
4. Calcola un intervallo di confidenza al 95% per la media μ di X sulla popolazione.

Soluzione.

Soluzione della domanda 1. Dalle due somme otteniamo media e deviazione standard, usando l'identità $\sigma^2 = \frac{x_1^2+\cdots+x_n^2}{n} - \mu^2$ (qui applicata al campione):

$$\mu(\text{campione}) = \frac{15\,482}{100} = 154.82$$
$$\sigma^2(\text{campione}) = \frac{2\,428\,902}{100} - 154.82^2 = 319.79 \qquad \sigma(\text{campione}) = \sqrt{319.79} = 17.88$$

Per mediana e quartili contiamo le foglie di ciascuna riga, e i totali progressivi:

riga	10	11	12	13	14	15	16	17	18	19
foglie	1	2	3	10	25	19	20	12	6	2
fino a qui	1	3	6	16	41	60	80	92	98	100

La mediana è la media di $x_{(50)}$ e $x_{(51)}$, che stanno nella riga 15 (posti dal 42 al 60) e ne sono la nona e la decima foglia, entrambe 4, quindi $\tilde{X} = 154$.

Per i quartili calcoliamo $k = 1 + 0.25 \cdot 99 = 25.75$ e $k = 1 + 0.75 \cdot 99 = 75.25$. I valori $x_{(25)}$ e $x_{(26)}$ sono la nona e la decima foglia della riga 14, entrambe 3, e $x_{(75)}$ e $x_{(76)}$ la quindicesima e la sedicesima foglia della riga 16, entrambe 7, quindi non abbiamo nulla da interpolare e

$$Q_1 = 143 \qquad Q_3 = 167 \qquad \text{IQR} = 167 - 143 = 24$$

□

Soluzione della domanda 2. Le classi di ampiezza 10 sono le righe del diagramma: 100–110, 110–120, &c. In ciascuna cade un numero di individui pari al numero delle foglie, che su 100 individui è anche la percentuale. La densità di probabilità è questa percentuale divisa per l'ampiezza della classe: per esempio, nella riga 14, $\frac{25\%}{10} = 2.5\,\%/\text{unità}$.

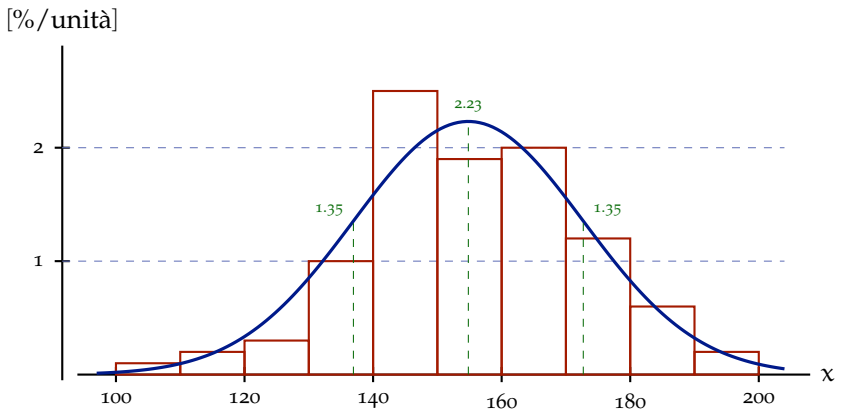
classe da	100	110	120	130	140	150	160	170	180	190
%/unità	0.1	0.2	0.3	1.0	2.5	1.9	2.0	1.2	0.6	0.2

Disegneremo l'istogramma nella figura della domanda seguente, insieme alla gaussiana.

□

Soluzione della domanda 3. La gaussiana ha il massimo in 154.8 e i flessi a una deviazione standard di distanza, in 154.8 ± 17.9, ossia in 136.9 e 172.7, e le quote corrispondenti sono

$$\frac{39.89\%}{17.88} = 2.23\,\frac{\%}{\text{unità}} \qquad \frac{24.20\%}{17.88} = 1.35\,\frac{\%}{\text{unità}}$$



La gaussiana segue abbastanza bene l'istogramma, e possiamo verificarlo contando gli individui. Entro una deviazione standard dalla media, fra 136.9 e 172.7, ne cadono 69 su 100, ed entro due deviazioni standard, fra 119.1 e 190.6, ne cadono 95, in buon accordo con la regola del 68–95–99.7. □

Soluzione della domanda 4. Con 100 individui il campione è grande, e possiamo usare il metodo per campioni grandi, in cui la stima di σ è $\hat{\sigma} = \sigma(\text{campione})$:

$$\widehat{\text{SEM}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{17.88}{\sqrt{100}} = 1.788 \qquad 1.96 \cdot 1.788 \approx 3.50$$

$$\mu = 154.8 \pm 3.5 \qquad \text{cioè fra 151.3 e 158.3, al 95\% di confidenza}$$

Con il metodo più generale non cambierebbe quasi nulla. La deviazione standard campionaria è $s = 17.88 \cdot \sqrt{100/99} = 17.97$, la soglia $t_{0.025,99} \approx 1.98$ prende il posto di 1.96, e otterremmo l'intervallo 154.8 ± 3.6. □

Esercizio 104 (p. 143)
Su un campione di compresse di un lotto si stima il contenuto medio di principio attivo, con l'intervallo di confidenza al 95% da 249 mg a 253 mg. Quali delle affermazioni seguenti sono corrette?

1. Il 95% delle compresse del lotto contiene fra 249 mg e 253 mg di principio attivo.
2. Intervalli costruiti nello stesso modo, a partire da campioni diversi, contengono il contenuto medio vero del lotto nel 95% dei casi.
3. Se si ripetesse il campionamento, il 95% delle nuove medie campionarie cadrebbe fra 249 mg e 253 mg.
4. Con un campione più grande l'intervallo sarebbe, di regola, più stretto.

Soluzione.

1. L'affermazione è falsa, perché l'intervallo che abbiamo calcolato riguarda la *media* del lotto e non le singole compresse, che sono molto più disperse: la loro deviazione standard è $\sqrt{n}$ volte l'errore standard della media.

2. È corretta, ed è proprio il significato del livello di confidenza — a cambiare per caso è il campione che scegliamo, e con lui l'intervallo.
3. Anche questa è falsa: il livello di confidenza del 95% ci garantisce semplicemente che il 95% degli intervalli di confidenza contengono la media vera, ma, sul singolo intervallo — in questo caso (251 ± 2) mg — non ci offre alcuna garanzia. Inoltre, se ripetessimo il campionamento, le nuove medie campionarie si distribuirebbero attorno alla media vera μ, non attorno alla nostra stima $\hat{\mu} = 251$ mg, che è a sua volta spostata da μ di una quantità casuale che non conosciamo.
4. È corretta, perché l'errore standard della media è $\sigma/\sqrt{n}$ e diminuisce al crescere di n.

L'errore del primo punto — confondere la media con i singoli individui — è di gran lunga il più grossolano e il più comune.

Esercizio 105 (p. 143)
Da un campione grande si ottiene, per la media di una variabile, l'intervallo di confidenza al 95% $\mu = 12.0 \pm 1.0$. Senza rifare i conti sui dati, trova gli intervalli di confidenza al 99% e al 68%.

Soluzione. Per un campione grande l'intervallo al 95% è $\hat{\mu} \pm 1.96\,\widehat{\mathrm{SEM}}$, quindi

$$\widehat{\mathrm{SEM}} = \frac{1.0}{1.96} \approx 0.51$$

Al 99% il coefficiente è 2.58, al 68% è 1:

$$\mu = 12.0 \pm 1.3 \quad \text{al } 99\% \qquad\qquad \mu = 12.0 \pm 0.5 \quad \text{al } 68\%$$

Più confidenza vuol dire meno precisione: un intervallo più sicuro è anche più largo.

Esercizio 106 (p. 143)
Con 25 misure ripetute di una concentrazione si ottiene l'intervallo di confidenza al 95% (6.4 ± 0.8) mg/L. Circa quante misure servirebbero, con lo stesso strumento, per ottenere un intervallo di ± 0.4 mg/L?

Soluzione. L'ampiezza dell'intervallo è proporzionale all'errore standard della media, $\sigma/\sqrt{n}$, e con lo stesso strumento σ non cambia. Per dimezzare l'intervallo dobbiamo quindi raddoppiare $\sqrt{n}$, ossia quadruplicare n, e ci servono circa $4 \cdot 25 = 100$ misure. Il conto è approssimato, perché la s del nuovo campione non sarà esattamente quella del primo, e anche il coefficiente della t di Student cambia un poco ($t_{0.025,\,24} = 2.06$, $t_{0.025,\,99} \approx 1.98$), ma l'ordine di grandezza resta quello.

Esercizio 107 (p. 143)
Un articolo riporta, per 100 pazienti, "peso medio 72 ± 3 kg", senza precisare che cosa sia il ± 3 kg. Fra quali valori sta, grosso modo, il peso del 95% dei pazienti, se

1. 3 kg è la deviazione standard dei pesi;
2. 72 ± 3 kg è l'intervallo di confidenza al 95% della media?

Supponi che i pesi siano distribuiti in modo circa gaussiano.

Soluzione.

1. Il 95% dei pazienti sta entro due deviazioni standard dalla media, quindi fra

$$72\,\text{kg} - 6\,\text{kg} = 66\,\text{kg} \qquad \text{e} \qquad 72\,\text{kg} + 6\,\text{kg} = 78\,\text{kg}$$

2. In questo caso $1.96\,\widehat{\text{SEM}} = 3\,\text{kg}$, da cui ricaviamo $\widehat{\text{SEM}} \approx 1.53\,\text{kg}$ e la deviazione standard dei pesi,

$$\hat{\sigma} = \sqrt{n} \cdot \widehat{\text{SEM}} = 10 \cdot 1.53\,\text{kg} = 15.3\,\text{kg}$$

   Il 95% dei pazienti sta allora fra circa $72 - 30 = 42\,\text{kg}$ e $72 + 30 = 102\,\text{kg}$.

Questo ipotetico articolo sarebbe quindi scritto coi piedi, perché lo stesso "± 3" può descrivere due popolazioni molto diverse a seconda di cosa rappresenta.

Esercizio 108 (p. 144)
Su 250 compresse prodotte da una macchina, scelte a caso e controllate, 12 risultano fuori specifica. Stima, con un intervallo di confidenza al 95%, la percentuale di compresse fuori specifica prodotte dalla macchina.

Soluzione. La proporzione nel campione è

$$\hat{p} = \frac{12}{250} = 0.048$$

e l'errore standard

$$\widehat{\text{SEM}} = \sqrt{\frac{0.048 \cdot 0.952}{250}} \approx 0.0135$$

Al 95% di confidenza, $p = 0.048 \pm 1.96 \cdot 0.0135 = 0.048 \pm 0.026$, ossia

$$p = (4.8 \pm 2.6)\% \qquad \text{cioè fra il 2.2\% e il 7.4\%}$$

Le condizioni sono appena soddisfatte: 12 compresse fuori specifica, poco più di una decina. Con un campione più piccolo il metodo non sarebbe affidabile.

Esercizio 109 (p. 144)
Un sondaggio su 1000 persone scelte a caso stima che il 40% della popolazione assuma regolarmente integratori alimentari, con un intervallo di confidenza al 95% di $\pm 3\%$.

1. Verifica che l'intervallo $\pm 3\%$ è corretto.
2. Con 4000 intervistati, e la stessa proporzione nel campione, quale sarebbe l'intervallo?

Soluzione.

1. Con $\hat{p} = 0.40$,

$$\widehat{\text{SEM}} = \sqrt{\frac{0.40 \cdot 0.60}{1000}} \approx 0.0155 \qquad 1.96 \cdot 0.0155 \approx 0.030$$

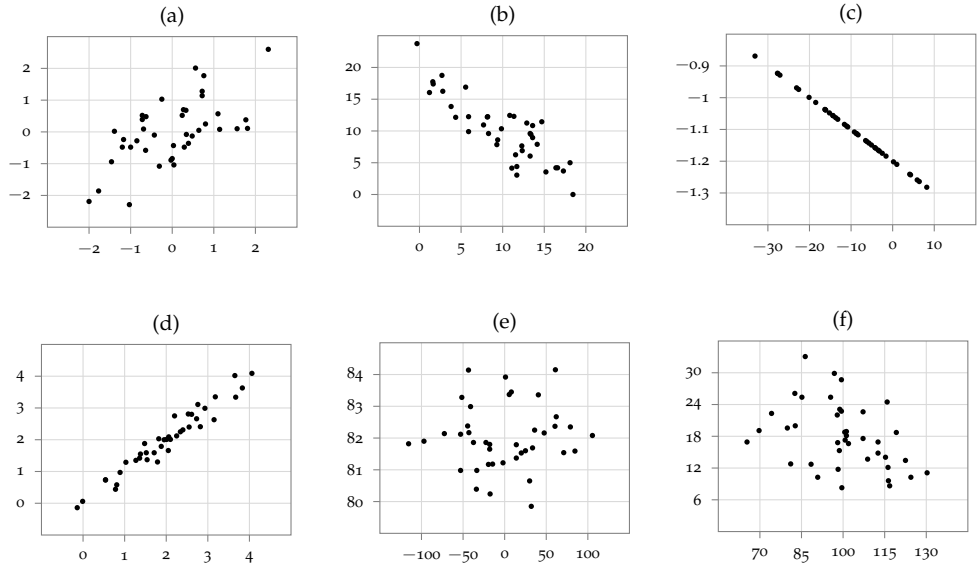
   cioè $\pm 3\%$.

2. Con quattro volte gli intervistati l'errore standard si dimezza: $\pm 1.5\%$.

Esercizio 110 (p. 154)
I coefficienti di correlazione corrispondenti ai seguenti diagrammi di dispersione sono

$$-0.85 \qquad -0.38 \qquad -1.00 \qquad 0.06 \qquad 0.97 \qquad 0.62$$

Associa ad ogni diagramma il suo coefficiente.



I valori indicati sugli assi sono spaziati di una deviazione standard delle rispettive variabili, con la media posta al centro. Per ciascuno dei diagrammi, calcola l'equazione della retta di regressione e traccia questa retta sul diagramma stesso.

Soluzione. Il valore -1.00 va al diagramma (c), i cui punti stanno esattamente su una retta decrescente. Fra gli altri, il segno si legge dall'inclinazione della nuvola, il valore assoluto da quanto la nuvola è schiacciata:

(a) 0.62 (b) -0.85 (c) -1.00 (d) 0.97 (e) 0.06 (f) -0.38

Per le rette usiamo $B = \mu_B + r\,\frac{\sigma_B}{\sigma_A}\,(A - \mu_A)$, leggendo medie e deviazioni standard sugli assi:

	μ_A	σ_A	μ_B	σ_B	retta
(a)	0	1	0	1	$B = 0.62\,A$
(b)	10	5	10	5	$B = 18.5 - 0.85\,A$
(c)	-10	10	-1.1	0.1	$B = -1.2 - 0.01\,A$
(d)	2	1	2	1	$B = 0.06 + 0.97\,A$
(e)	0	50	82	1	$B = 82 + 0.0012\,A$
(f)	100	15	18	6	$B = 33.2 - 0.152\,A$