CAPITOLO

λ1 STATISTICA IN CAMPO SOCIALE ED ECONOMICO

1 Indagini statistiche in campo sociale ed economico

▶ Esercizi a p. λ32

Dalla domanda all'indagine statistica

Alla base di un'indagine statistica ci sono domande a cui si cerca di rispondere. Esaminiamo un esempio semplificato, con una sola domanda.

Guido, l'addetto alle risorse umane di un'azienda, sostiene che per il posto di lavoro per cui sta facendo i colloqui si presentano soprattutto donne laureate di età compresa tra i 25 e i 35 anni. Ha ragione?

Per rispondere a questa domanda, possiamo svolgere un'**indagine statistica**. Scelti i soggetti interessanti per l'indagine, non poniamo loro la nostra domanda in modo diretto, ma la trasformiamo in **variabili**, cioè domande più semplici le cui risposte varieranno da soggetto a soggetto.

Preparazione dell'indagine

Ogni domanda è associata a un **carattere**. Nella nostra indagine, i caratteri sono il *sesso*, l'*età* e il *titolo di studio* dei candidati al posto.

▶ Stai leggendo di un'indagine che riguarda le fonti per la produzione di energia elettrica e la quantità di energia ricavata da ciascuna fonte. Individua i caratteri dell'indagine stabilendo il tipo di ciascuno.

Il carattere *sesso* si può presentare nelle due **modalità** *maschio* e *femmina*, il *titolo di studio* nelle modalità *nessun titolo di studio*, *licenza elementare*, *licenza media*, *diploma*, *laurea*, *dottorato/master*.

Entrambi i caratteri sono di tipo **qualitativo**, in quanto le modalità sono espresse da parole. Il *titolo di studio* è di tipo **qualitativo ordinale**, perché esiste un ordinamento naturale delle modalità. Il *sesso* è di tipo **qualitativo nominale**.

Il carattere *età*, invece, è di tipo **quantitativo**, perché le modalità sono espresse da numeri, e **discreto**, perché derivano da un'operazione di conteggio.

Ciascun candidato rappresenta un'**unità statistica** e il loro insieme rappresenta la **popolazione statistica**.

Immaginiamo che dai curricula dei dieci candidati che finora si sono presentati si ottengano le seguenti informazioni (con F indichiamo *femmina*, con M *maschio*).

Candidato	Nome	Sesso	Età	Titolo di studio
1	Sandra	F	25	laurea
2	Chris	non dichiarato	23	diploma
3	Luca	non dichiarato	30	laurea
4	Andrea	non dichiarato	25	laurea
5	Francesco	M	31	diploma
6	Xiang	M	40	diploma
7	Giovanna	non dichiarato	28	laurea
8	Angela	F	33	laurea
9	Giuseppe	non dichiarato	33	laurea
10	Francesca	F	32	dottorato

← Tabella 1

Indagini censuarie e indagini campionarie

L'indagine del nostro esempio è un'**indagine censuaria**: di ciascun elemento della popolazione vogliamo conoscere la modalità con cui si presenta ogni carattere.
In un'indagine censuaria è sufficiente utilizzare gli strumenti della **statistica descrittiva**. Le conclusioni a cui si giunge sono vere per la popolazione osservata.

Abbandoniamo per un attimo la nostra indagine e consideriamone un'altra.
Se vogliamo avere informazioni sulla situazione lavorativa di *tutti* gli italiani di età compresa tra i 25 e i 35 anni, suddividendoli in base al sesso e al titolo di studio, i caratteri di interesse possono essere l'età, il titolo di studio, il sesso e l'occupazione.
In questa seconda situazione è difficile ricavare le informazioni necessarie per tutti gli italiani nella fascia di età indicata. Per farlo si può selezionare un sottoinsieme della popolazione in modo opportuno, ottenendo un **campione**. L'indagine così condotta è un'**indagine campionaria**.
In un'indagine campionaria, oltre ai metodi della statistica descrittiva, si devono utilizzare anche gli strumenti della **statistica inferenziale**, che permettono di generalizzare le informazioni ottenute a un insieme più ampio rispetto al campione su cui si sono raccolti i dati.
La scelta fra i due tipi di indagine, censuaria o campionaria, dipende dalla numerosità della popolazione: maggiore è la numerosità più è difficile un'indagine censuaria. Entrambe le indagini, però, possono generare errori:

- **errore campionario**: deriva dal fatto che si osserva solo una parte della popolazione;
- **errore non campionario**: deriva dalla scarsa qualità dei dati (dati mancanti, sbagliati o incompleti).

Le indagini campionarie presentano sempre un errore campionario, dovuto al fatto che non si analizza l'intera popolazione ma solo una parte.
Nelle indagini censuarie l'errore campionario non c'è, mentre quello non campionario può essere molto elevato, soprattutto se la numerosità della popolazione è molto alta o non si ha modo di controllare le risposte o si utilizzano dati provenienti da altre indagini.

Riprendiamo il nostro esempio. Osservando la tabella, notiamo che cinque persone non hanno specificato il sesso nel curriculum; tra queste, Chris e Andrea

► Devi svolgere un'indagine sulla spesa per andare al cinema tra i tuoi compagni di classe, tra tutti i ragazzi della tua scuola, tra tutte le persone della tua città. Scegli caso per caso il tipo di indagine, censuaria o campionaria, motivando la scelta.

hanno nomi che non permettono di sapere qual è il loro sesso e sono perciò da interpretarsi come dati mancanti.

Raccolta dei dati

Nel nostro primo esempio, per ottenere i dati, abbiamo utilizzato le informazioni raccolte in precedenza. Una tale indagine è detta *indiretta*.

DEFINIZIONE

Un'**indagine diretta** è un'indagine in cui i dati vengono raccolti tramite la somministrazione di un questionario.
Un'**indagine indiretta** è un'indagine in cui si utilizzano dati che erano già stati raccolti per altri scopi e sono quindi già disponibili.

Le indagini indirette permettono di ridurre notevolmente i tempi di raccolta dei dati, tuttavia non sempre sono disponibili dati che rispondono precisamente alle necessità dell'indagine.
Per esempio, nel caso dell'indagine sui 10 candidati, non c'è l'obbligo di scrivere il sesso sul curriculum e quindi questa informazione può non essere sempre deducibile.
Quando si realizza un'indagine diretta è necessario scegliere fra tre modalità di intervista quella più appropriata: le **interviste faccia a faccia**, le **interviste telefoniche** e l'**autocompilazione**. Ogni metodo ha vantaggi (velocità o precisione) e svantaggi (inapplicabilità in alcune situazioni, alta possibilità di mancate risposte o cattiva interpretazione delle domande).
La scelta deve essere il giusto compromesso in base alla numerosità della popolazione o del campione, alle risorse, anche economiche, e alla difficoltà delle domande.

ESEMPIO

1. L'**exit poll** è un'indagine statistica effettuata durante i giorni delle elezioni. L'obiettivo dell'indagine è stimare i risultati del voto in corso non appena si chiude la votazione, e quindi prima della divulgazione dei risultati ufficiali. La sua caratteristica principale è la velocità con cui devono essere prodotti i risultati, perciò l'indagine è campionaria e basata su un questionario molto breve.
   La raccolta dei dati avviene tramite interviste faccia a faccia con alcuni elettori che hanno appena espresso il proprio voto. All'uscita dalle urne, gli intervistatori chiedono agli intervistati quale preferenza di voto hanno appena espresso e talvolta pongono altre domande.
2. Il **censimento** della popolazione e delle abitazioni è un'indagine statistica realizzata dall'Istat (Istituto nazionale di statistica). L'obiettivo dell'indagine è produrre informazioni sulle condizioni economiche e sociali della popolazione e sulle caratteristiche delle abitazioni in Italia.
   Fino al 2011, data dell'ultimo censimento, era un'indagine censuaria realizzata con cadenza decennale. Tutte le unità statistiche della popolazione di riferimento, ovvero tutte le famiglie residenti in Italia, dovevano compilare un lungo questionario. A partire dal 2015 le modalità di raccolta dei dati sono state modificate e ora è un'indagine mista, ovvero alcuni dati derivano da indagini indirette, mentre altri da un'indagine campionaria.

2 Frequenze e indici statistici

Frequenze

▶ Esercizi a p. λ33

Riprendiamo il nostro problema iniziale. Sintetizziamo i dati raccolti scrivendo le tabelle delle **frequenze**.

Sesso		
Modalità	**Frequenza assoluta**	**Frequenza relativa**
M	4	0,4 = 40%
F	4	0,4 = 40%
dato mancante	2	0,2 = 20%
Totale	10	1 = 100%

frequenza relativa: $\frac{\text{frequenza assoluta}}{\text{numero totale delle unità}}$ può essere espressa anche in percentuale

← Tabella 2

Età				
Modalità	**Frequenza assoluta**	**Frequenza relativa**	**Frequenza cumulata**	**Frequenza cumulata relativa**
20-24	1	0,1 = 10%	1	0,1 = 10%
25-29	3	0,3 = 30%	4	0,4 = 40%
30-35	5	0,5 = 50%	9	0,9 = 90%
36-40	1	0,1 = 10%	10	1 = 100%
Totale	10	1 = 100%		

← Tabella 3

Titolo di studio				
Modalità	**Frequenza assoluta**	**Frequenza relativa**	**Frequenza cumulata**	**Frequenza cumulata relativa**
diploma	3	0,3 = 30%	3	0,3 = 30%
laurea	6	0,6 = 60%	9	0,9 = 90%
dottorato	1	0,1 = 10%	10	1 = 100%
Totale	10	1 = 100%		

← Tabella 4

Per tutti e tre i caratteri abbiamo scritto le **frequenze assolute** e le **frequenze relative**. Per ogni carattere e per ogni modalità la frequenza relativa è stata calcolata contando il numero di unità statistiche che presentano la modalità e dividendo questo numero, che è la frequenza assoluta, per il totale delle unità statistiche.
La **frequenza cumulata**, invece, è stata calcolata solo per i caratteri *età* e *titolo di studio*. Infatti, per poter calcolare la frequenza cumulata di una modalità è necessario ordinare le modalità e sommare la frequenza di quella modalità alle frequenze di tutte le precedenti. È indispensabile, quindi, che i caratteri siano quantitativi o qualitativi ordinali in modo che ci sia un ordinamento.
Osserviamo anche che le modalità del carattere *età* sono state raggruppate in **classi**. Questo fa perdere alcune informazioni, ma migliora la leggibilità della tabella rendendo più facile l'interpretazione dei dati. Se vogliamo sapere se è vero che la

▶ Una classe è formata da 15 maschi e 8 femmine. Tra i componenti della classe, 6 sono figli unici, 8 hanno un fratello o una sorella, 9 hanno due fratelli. Considera i caratteri di questa statistica e scrivi le tabelle con tutte le frequenze calcolabili per ciascun carattere.

TEORIA

maggior parte dei candidati ha un'età compresa tra i 25 e i 35 anni basta sommare le frequenze relative di due sole classi.
Leggendo queste tre tabelle, possiamo risolvere il problema iniziale: verificare l'affermazione di Guido. Sulla popolazione osservata è vero che la maggior parte dei candidati ha tra i 25 e i 35 anni (sono 30% + 50%, cioè l'80%), è vero che la maggior parte di essi è laureata (sono 60% + 10%, cioè il 70%), mentre non possiamo concludere che siano prevalentemente donne. Quindi, non possiamo dare ragione a Guido.
Nel trarre le conclusioni ci siamo riferiti alla frequenza relativa (espressa in percentuale). Questa, in genere, dà più informazioni della frequenza assoluta. Immaginiamo di fare la stessa indagine in un'altra azienda, in cui i candidati sono 20 anziché 10, e di scoprire che in questo caso i laureati sono 8. Se osserviamo soltanto la frequenza assoluta, saremmo portati a concludere che nella seconda azienda si presentino più laureati che nella prima azienda. La conclusione, però sarebbe errata, perché rispetto al totale dei candidati la percentuale dei laureati che si presentano nella seconda azienda è minore.
Le frequenze danno informazioni dettagliate sul fenomeno, ma nelle indagini statistiche sono utili anche valori sintetici, come quelli che stiamo per esaminare, cioè gli indici di posizione, di variabilità e di concentrazione.

Indici di posizione

▶ Esercizi a p. λ35

Media aritmetica, mediana e moda

Riprendiamo le età dei candidati della tabella 1 e calcoliamo la **media aritmetica**:

$$M = \frac{25 + 23 + 30 + 25 + 31 + 40 + 28 + 33 + 33 + 32}{10} = 30.$$

Questo è il numero che, sostituito a ciascun dato, lascia inalterata la somma: se tutti i candidati avessero 30 anni, la somma delle età sarebbe ancora 300.
Avremmo potuto calcolare l'età media anche a partire dai dati raggruppati in classi della tabella 3. In questo caso, usando il valore centrale della classe e la frequenza di quella classe, avremmo ottenuto:

$$M = \frac{22 \cdot 1 + 27 \cdot 3 + 32{,}5 \cdot 5 + 38 \cdot 1}{10} = 22 \cdot \underbrace{0{,}1}_{\text{frequenza relativa della modalità 22}} + 27 \cdot \underbrace{0{,}3}_{\text{frequenza relativa della modalità 27}} + 32{,}5 \cdot 0{,}5 + 38 \cdot 0{,}1 = 30{,}35.$$

Alla media aritmetica possiamo affiancare anche il calcolo della **mediana**.
Il numero di dati è pari, quindi, dopo aver ordinato le età in modo crescente, calcoliamo la media tra i due valori centrali della sequenza.

▶ Scrivi il numero medio e il numero mediano di fratelli nell'indagine dell'esercizio precedente. [circa 1; 1]

$$\underset{}{23 \quad 25 \quad 25 \quad 28 \quad \overbrace{30 \quad 31}^{\text{valori centrali}} \quad 32 \quad 33 \quad 33 \quad 40} \qquad \frac{30 + 31}{2} = 30{,}5$$

Se il numero di dati fosse stato dispari, avremmo preso il valore centrale.
Come puoi osservare nella sequenza dei dati ordinati, la mediana ha la proprietà di trovarsi al centro della sequenza.
Se consideriamo la suddivisione in classi, la classe mediana è 30-35, perché è la prima classe con frequenza cumulata relativa maggiore o uguale al 50%.
Infine possiamo cercare la **moda**, cioè la modalità con frequenza maggiore. Nel nostro esempio, la moda non è molto significativa: i valori con maggiore frequenza

sono due, 25 e 33, ma la loro frequenza non è molto diversa da quella degli altri valori. Può essere invece interessante osservare, nella tabella 3, che la classe modale è 30-35.

L'indice di posizione più opportuno

La media aritmetica è calcolabile solo per caratteri quantitativi.
La mediana, invece, può essere calcolata sia per caratteri quantitativi sia per caratteri qualitativi ordinali. Per esempio, considerando l'ordinamento naturale delle modalità del carattere *titolo di studio*, possiamo dire che la mediana è *laurea*.
La moda, infine, può essere determinata per qualunque tipo di carattere.
La media aritmetica è molto influenzata dai valori anomali, cioè quelli più estremi. Tali valori, invece, non cambiano la mediana.

ESEMPIO

La tabella a fianco contiene il reddito annuo del 2016 di dieci lavoratori.
Il reddito annuo medio della popolazione dei 10 lavoratori è di € 16 921,1; quindi, se ridistribuissimo gli stipendi per far avere a tutti lo stesso reddito, ognuno riceverebbe € 16 921,1.
Il reddito annuo mediano è invece di € 15 599,5, quindi il 50% dei lavoratori guadagna non più di € 15 599,5.
Considerando la distribuzione dei redditi della tabella, supponiamo che il reddito di Sofia non sia di € 29 102, ma di € 103 569, e ricalcoliamo il reddito annuo medio e il reddito annuo mediano:

$M = 24\,367{,}8$; mediana $= 15\,599{,}5$.

La media aritmetica è aumentata di € 7446,7, mentre la mediana non è cambiata.

Nome	Reddito (€)
Francesco	6720
Alessandro	8859
Giulia	10 753
Matteo	12 668
Martina	14 560
Lorenzo	16 639
Andrea	19 083
Chiara	22 453
Sara	28 374
Sofia	29 102

È opportuno affiancare alla media anche la mediana nel caso in cui i dati siano raggruppati in classi e ci siano **classi aperte**.
Per esempio, se l'ultima classe del carattere *età*, nella tabella 3, fosse stato «maggiore o uguale a 36», cioè se la classe fosse stata aperta, avremmo dovuto scegliere arbitrariamente il valore centrale da utilizzare nel calcolo della media, a discapito della precisione del risultato. La mediana, invece, non ne avrebbe risentito.

Media ponderata

La **media ponderata** è molto simile, nel significato, alla media aritmetica, ma si utilizza prevalentemente quando ai dati è assegnato un peso, cioè un'importanza diversa, oppure se i dati sono già rappresentati in modo sintetico.

ESEMPIO

1. Alessia, iscritta al corso di laurea in Matematica, dopo gli esami del primo anno ha ottenuto le seguenti valutazioni.

Esame	Algebra	Analisi 1	Fisica 1	Informatica	Geometria
Crediti	9	15	9	12	15
Voto	25	27	20	22	28

Calcoliamo la media ponderata alla fine degli esami del primo anno, considerando come peso i crediti formativi.

$$M_p = \frac{\overbrace{25 \cdot 9 + 27 \cdot 15 + 20 \cdot 9 + 22 \cdot 12 + 28 \cdot 15}^{\text{somma dei prodotti dei valori per i loro pesi}}}{\underbrace{9 + 15 + 9 + 12 + 15}_{\text{somma dei pesi}}} = \frac{1494}{60} = 24{,}9$$

TEORIA

2. In un'indagine sul voto di maturità in 20 licei scientifici è risultato che nell'ultimo anno in 5 licei il voto medio è stato 78,3, in 11 è stato 85,6 e in 4 è stato 93,6. Calcoliamo il voto medio complessivo nei 20 licei:

$$M_p = \frac{5 \cdot 78{,}3 + 11 \cdot 85{,}6 + 4 \cdot 93{,}6}{5 + 11 + 4} = 85{,}375.$$

Media geometrica

La **media geometrica** è il numero che sostituito ai dati lascia inalterato il loro prodotto. È particolarmente indicata per trovare una percentuale media di variazione.

ESEMPIO

La popolazione del Brasile ha registrato dal 2010 al 2014 le seguenti variazioni percentuali: + 11,3%, + 11%, + 8,3% e + 8%. Troviamo la variazione media.
Nel 2010 gli abitanti del Brasile erano 195 200 000 e nel 2014 erano 282 065 630.
La popolazione del Brasile nel 2011 si ottiene con:

$$195\,200\,000 + 195\,200\,000 \cdot \overbrace{0{,}113}^{11{,}3\%} = 195\,200\,000 \cdot (1 + 0{,}113).$$

In modo analogo,

popolazione del 2012: $195\,200\,000 \cdot \overbrace{(1+0{,}113)}^{\text{popolazione del 2011}} \cdot (1+0{,}11)$,

popolazione del 2013: $195\,200\,000 \cdot \overbrace{(1+0{,}113)\cdot(1+0{,}11)}^{\text{popolazione del 2012}} \cdot (1+0{,}083)$,

popolazione del 2014: $195\,200\,000 \cdot \overbrace{(1+0{,}113)\cdot(1+0{,}11)\cdot(1+0{,}083)}^{\text{popolazione del 2013}} \cdot (1+0{,}08)$.

Otteniamo così il valore già noto per il 2014:

$$195\,200\,000(1 + 0{,}113)(1 + 0{,}11)(1 + 0{,}083)(1 + 0{,}08) = 282\,065\,630.$$

La variazione media r è la variazione costante che, se fosse stata registrata in tutti i quattro anni, avrebbe portato alla stessa popolazione nel 2014, ovvero:

$$195\,200\,000(1 + r)(1 + r)(1 + r)(1 + r) = 282\,065\,630.$$

Quindi deve essere:

$$(1 + r)^4 = (1 + 0{,}113)(1 + 0{,}11)(1 + 0{,}083)(1 + 0{,}08) \quad \rightarrow$$

$$1 + r = \overbrace{\sqrt[\overbrace{4}^{\text{numero dei valori}}]{\underbrace{1{,}113 \cdot 1{,}11 \cdot 1{,}083 \cdot 1{,}08}_{\text{prodotto dei valori}}}}^{\text{media geometrica di 1,113; 1,11; 1,083; 1,08}} \simeq 1{,}096.$$

Ricaviamo r:

$$1 + r = 1{,}096 \quad \rightarrow \quad r = 1{,}096 - 1 = 0{,}96.$$

Concludiamo che la percentuale media di variazione è del 9,6%.

▶ La produzione di un bene ha avuto negli ultimi quattro anni le seguenti variazioni percentuali: + 10,2%, −4%, + 15%, + 16%. Calcola la variazione media nei quattro anni. [circa 9%]

Indici di variabilità

▶ Esercizi a p. λ39

Un'indagine sul numero di figli condotta su due campioni diversi, entrambi costituiti da 100 famiglie, ha dato le seguenti distribuzioni di frequenze.

Campione A				
Numero di figli	0	1	2	3
Frequenza	10	30	45	15

Campione B				
Numero di figli	0	1	2	3
Frequenza	31	18	6	45

Il numero medio di figli sui due campioni è lo stesso, infatti:

$$M_A = \frac{0 \cdot 10 + 1 \cdot 30 + 2 \cdot 45 + 3 \cdot 15}{10 + 30 + 45 + 15} = \frac{165}{100} = 1{,}65,$$

$$M_B = \frac{0 \cdot 31 + 1 \cdot 18 + 2 \cdot 6 + 3 \cdot 45}{31 + 18 + 6 + 45} = \frac{165}{100} = 1{,}65.$$

Anche le mediane coincidono, e sono uguali a 2.
Le distribuzioni osservabili in tabella, però, sono molto diverse, cosa di cui non potremmo renderci conto se conoscessimo solo la media e la mediana.
Per questo, per descrivere meglio le distribuzioni, utilizziamo anche i seguenti **indici di dispersione**.

- Il **campo di variazione** è la differenza tra la modalità massima rilevata e la modalità minima rilevata: $x_{max} - x_{min}$.
- Lo **scarto semplice medio** S è la media aritmetica dei valori assoluti degli scarti dalla media aritmetica del carattere delle modalità rilevate:
$$S = \frac{|x_1 - M| + |x_2 - M| + \ldots + |x_n - M|}{n}.$$
- La **deviazione standard** σ è la radice quadrata della media aritmetica dei quadrati degli scarti dalla media aritmetica del carattere delle modalità rilevate:
$$\sigma = \sqrt{\frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}}.$$

Calcoliamo questi indici sui due campioni già considerati.
Campo di variazione di entrambi i campioni:

$$x_{max} - x_{min} = 3 - 0 = 3.$$

Scarti semplici medi:

$$S_A = \frac{|0-1{,}65| \cdot 10 + |1-1{,}65| \cdot 30 + |2-1{,}65| \cdot 45 + |3-1{,}65| \cdot 15}{100} = 0{,}720,$$

$$S_B = \frac{|0-1{,}65| \cdot 31 + |1-1{,}65| \cdot 18 + |2-1{,}65| \cdot 6 + |3-1{,}65| \cdot 45}{100} = 1{,}257.$$

Deviazioni standard:

$$\sigma_A = \sqrt{\frac{(0-1{,}65)^2 \cdot 10 + (1-1{,}65)^2 \cdot 30 + (2-1{,}65)^2 \cdot 45 + (3-1{,}65)^2 \cdot 15}{100}} \simeq 0{,}853,$$

$$\sigma_B = \sqrt{\frac{(0-1{,}65)^2 \cdot 31 + (1-1{,}65)^2 \cdot 18 + (2-1{,}65)^2 \cdot 6 + (3-1{,}65)^2 \cdot 45}{100}} \simeq 1{,}322.$$

Lo scarto semplice medio e la deviazione standard sul campione A sono minori dello scarto semplice medio e della deviazione standard sul campione B. Ciò vuol dire che sul campione A il carattere è meno disperso, cioè i dati sono più concentrati intorno alla media (la maggior parte delle unità statistiche presentano le modalità 1 figlio o 2 figli), mentre sul campione B il carattere è più disperso (ci sono molte unità statistiche che presentano modalità estreme).

La deviazione standard è influenzata dall'ordine di grandezza dei dati e ha la stessa unità di misura dei dati: è perciò poco utile se si vogliono confrontare gruppi diversi oppure grandezze diverse. Per eliminare questa dipendenza si può utilizzare il **coefficiente di variazione**, che è uguale al rapporto tra la deviazione standard e la media aritmetica. Lo indichiamo con CV:

$$CV = \frac{\sigma}{M}.$$

TEORIA

ESEMPIO

Sono stati rilevati pesi e lunghezze di cinque neonati.

Peso (kg)	3,6	3,5	2,9	3	3,5
Lunghezza (cm)	49	50	48	49	52

Se calcoliamo le medie aritmetiche e le deviazioni standard dei due caratteri, otteniamo:

$$M_p = 3{,}3\text{ kg}, \quad \sigma_p \simeq 0{,}29\text{ kg}; \qquad M_l = 49{,}6\text{ cm}, \quad \sigma_l \simeq 1{,}36\text{ cm}.$$

Essendo $0{,}29 < 1{,}36$, possiamo dire che la lunghezza ha variabilità maggiore? No, perché i due valori riguardano grandezze diverse e hanno unità di misura diverse, quindi non sono confrontabili. Se calcoliamo i coefficienti di variazione,

$$CV_p = \frac{0{,}29}{3{,}3} \simeq 0{,}09 \quad \text{e} \quad CV_l = \frac{1{,}36}{49{,}6} \simeq 0{,}03,$$

possiamo confrontarli e dedurre che è maggiore la variabilità dei pesi.

▶ Quelli che seguono sono i prezzi di cinque tipi di pasta e di tre tipi di liquori.
Pasta: 0,90; 2,15; 1,5; 1; 0,99.
Liquori: 6,78; 7,90; 8,2.
È maggiore la variabilità del prezzo della pasta o quella del prezzo dei liquori? [pasta]

Indici di concentrazione

▶ Esercizi a p. λ40

Gli **indici di concentrazione** permettono di dare una misura sull'equità di ripartizione di un totale su una popolazione. Tali indici, perciò, hanno senso solo per caratteri **quantitativi trasferibili**, cioè per i quali si può pensare che un'unità ceda parte del carattere che possiede a un'altra. Per esempio, mentre il reddito di una persona è trasferibile, l'età non lo è. L'indice di concentrazione più utilizzato è **l'indice di Gini**, che confronta la distribuzione rilevata con la distribuzione teorica, cioè quella in cui c'è equidistribuzione.
Si vuole indagare su come è distribuito il reddito familiare mensile degli abitanti di una città in relazione a due contesti abitativi differenti. Si scelgono due quartieri diversi della città e in ciascuno un campione di cinque famiglie, e si ottiene la seguente tabella.

Quartiere	Redditi (€)					Totale
A	2300	1000	2000	1500	10 000	16 800
B	3300	3400	3500	3200	3000	16 400

Il reddito complessivo sui due campioni è circa lo stesso. I singoli dati, però, suggeriscono che è ripartito diversamente nelle cinque famiglie di ciascun campione.

Calcoliamo l'indice di Gini sul campione del quartiere A.
Se la distribuzione dei redditi nel quartiere A fosse quella teorica, tutte le famiglie avrebbero lo stesso reddito, pari al reddito medio: $\frac{16\,800}{5} = 3360$. Ciascuna famiglia avrebbe la stessa percentuale del reddito totale, cioè il 20%. La percentuale cumulata per due famiglie sarebbe 40%, per tre famiglie 60% e per quattro famiglie 80%.

Elaboriamo i dati relativi al quartiere A nel seguente modo.

- Ordiniamo le modalità rilevate in ordine crescente:

  € 1000, € 1500, € 2000, € 2300, € 10 000.

- Dividiamo ogni reddito per il totale. Otteniamo la percentuale del reddito totale che compete a ogni famiglia.

Reddito (€)	1000	1500	2000	2300	10 000
% rispetto al totale	$\frac{1000}{16\,800} \simeq 6{,}0\%$	$\frac{1500}{16\,800} \simeq 8{,}9\%$	$\frac{2000}{16\,800} \simeq 11{,}9\%$	$\frac{2300}{16\,800} \simeq 13{,}7\%$	$\frac{10\,000}{16\,800} \simeq 59{,}5\%$

- Determiniamo le percentuali cumulate di ogni unità statistica, sommando la percentuale di ogni unità con le percentuali delle unità precedenti.

Reddito (€)	1000	1500	2000	2300	10 000
% cumulate	6,0%	14,9%	26,8%	40,5%	100%

Ciò vuol dire che la famiglia con reddito minore ha il 6% del reddito totale, le due famiglie più povere hanno il 14,9% del reddito totale e così via.

- Calcoliamo l'indice di Gini considerando le prime quattro unità statistiche, sommando le differenze tra le percentuali cumulate teoriche e quelle rilevate e dividendo il risultato per la somma delle percentuali cumulate teoriche, cioè:

$$R_A = \frac{(20-6)+(40-14{,}9)+(60-26{,}8)+(80-40{,}5)}{20+40+60+80} = 0{,}559.$$

Se calcoliamo l'indice di Gini anche per il quartiere B, otteniamo:

$R_B = 0{,}035.$

Osserviamo che sul campione del quartiere B, dove i redditi sono tutti molto simili, l'indice di Gini è molto vicino a 0: il reddito è all'incirca equidistribuito tra le cinque famiglie. Sul campione del quartiere A, dove una famiglia ha un reddito molto superiore rispetto alle altre, l'indice di Gini è più vicino a 1: il reddito è distribuito in modo meno uniforme.
In generale, diamo la seguente definizione.

DEFINIZIONE

Siano $x_1, x_2, \ldots, x_n$ i dati ordinati relativi a un carattere quantitativo trasferibile e sia S la loro somma. Indichiamo con $q_i = \frac{x_i}{S}$ e $p_i = \frac{1}{n}$ le percentuali rilevate e teoriche di ogni unità rispetto al totale e con $Q_i = q_1 + q_2 + \ldots + q_i$ e $P_i = p_1 + p_2 + \ldots + p_i$ le percentuali cumulate rilevate e teoriche. L'**indice di Gini** è:

$$R = \frac{(P_1 - Q_1)+(P_2 - Q_2)+\ldots+(P_{n-1} - Q_{n-1})}{P_1 + P_2 + \ldots + P_{n-1}}.$$

L'indice di Gini è:

- un numero compreso tra 0 e 1;
- uguale a 0, quando il totale è equamente diviso tra le unità statistiche;
- uguale a 1, quando una sola unità possiede tutto il totale.

Animazione

Nell'animazione ci sono tutti i calcoli relativi anche agli indici di Gini per il quartiere B.

MATEMATICA E SCIENZE SOCIALI

Distribuzione dei redditi nel mondo e in Italia

I redditi sono ripartiti in modo più uniforme negli Stati Uniti o in Germania? Negli anni, in Italia la distribuzione dei redditi è sempre stata la stessa? Per rispondere a domande come queste puoi cercare tabelle, mappe e grafici relativi all'indice di Gini.

Cerca nel Web: stati uguaglianza reddito, mappa mondiale Gini, Gini Italia anni

3 Rapporti statistici socioeconomici

Tassi demografici

▶ Esercizi a p. λ41

I **fenomeni demografici** sono quelli che modificano la dimensione o la composizione di una popolazione di persone in una certa area geografica.

TEORIA

I rapporti statistici utilizzati per analizzare i fenomeni demografici prendono il nome di **tassi demografici**.

ESEMPIO

Consideriamo i dati Istat (anno 2013) sulla natalità in Liguria, riportati nella tabella a fianco.

Provincia	Nascite	Popolazione
Imperia	1475	214 290
Savona	1849	280 837
Genova	5785	851 283
La Spezia	1563	218 717
Totale	10 672	1 565 127

Da una lettura frettolosa della tabella potremmo essere portati a pensare che a Genova la natalità sia molto più alta che nelle altre province.

Tale conclusione però è errata, perché invece dei valori assoluti dobbiamo confrontare i tassi di natalità, cioè i rapporti:

$$n = \frac{\text{numero di nascite}}{\text{popolazione}} \cdot 1000.$$

Calcoliamoli nelle quattro province:

$$n_{\text{Imperia}} = \frac{1475}{214\,290} \cdot 1000 = 6{,}88; \qquad n_{\text{Savona}} = \frac{1849}{280\,837} \cdot 1000 = 6{,}58;$$

$$n_{\text{Genova}} = \frac{5785}{851\,283} \cdot 1000 = 6{,}80; \qquad n_{\text{La Spezia}} = \frac{1563}{218\,717} \cdot 1000 = 7{,}15.$$

Quindi la natalità più alta della regione è a La Spezia e non a Genova.

Riassumiamo nella seguente tabella i principali tassi demografici.

Principali tassi demografici	
Tasso di natalità	$n = \frac{\text{numero di nascite}}{\text{popolazione}} \cdot 1000$
Tasso di mortalità	$m = \frac{\text{numero di morti}}{\text{popolazione}} \cdot 1000$
Tasso di immigrazione	$i = \frac{\text{numero di persone immigrate}}{\text{popolazione}} \cdot 1000$
Tasso di emigrazione	$e = \frac{\text{numero di persone emigrate}}{\text{popolazione}} \cdot 1000$
Tasso di nuzialità	$s = \frac{\text{numero di matrimoni}}{\text{popolazione}} \cdot 1000$

Esistono due tipologie di tassi demografici.

- **Tassi generici**: il numeratore e il denominatore sono riferiti all'intera popolazione.
- **Tassi specifici**: il numeratore e il denominatore sono riferiti a un sottoinsieme della popolazione identificato in base alla modalità di un certo carattere (per esempio sesso, età, stato civile ecc.).

I tassi generici si possono calcolare anche come media ponderata dei tassi specifici usando come pesi il numero di individui in ogni sottoinsieme della popolazione.

ESEMPIO

La tabella riporta i tassi di mortalità specifici per età in Italia in base ai dati Istat (anno 2015).

Età	Popolazione (migliaia)	Tasso di mortalità
0-19	11 247	0,24
20-39	14 231	0,40
40-59	18 468	2,19
60-79	12 872	14,93
80+	3 977	102,21

Il tasso di mortalità generico calcolato a partire dai tassi di mortalità specifici è:

$$m = \frac{11247 \cdot 0,24 + 14231 \cdot 0,40 + 18468 \cdot 2,19 + 12872 \cdot 14,93 + 3977 \cdot 102,21}{11247 + 14231 + 18468 + 12872 + 3977} \simeq$$

$$\frac{647\,504}{60\,795} = 10,65.$$

▶ Nella sezione A di una scuola ci sono 25 studenti in prima, 26 in seconda, 24 in terza, 27 in quarta e 20 in quinta. Alla fine dell'anno sono stati bocciati 2 ragazzi in prima, 3 in seconda, 3 in terza, 1 in quarta e 2 in quinta.
Calcola i tassi specifici dei bocciati in ogni classe e il tasso generico dei bocciati nella sezione A.

Indici dei prezzi al consumo

▶ Esercizi a p. λ42

Un'altra applicazione dei rapporti statistici è lo studio dell'**inflazione**, cioè dell'aumento dei prezzi nel tempo con conseguente diminuzione del potere di acquisto della moneta. Per analizzare l'inflazione si usano gli **indici dei prezzi al consumo**.

DEFINIZIONE

Un **indice dei prezzi al consumo** I_t è il rapporto percentuale tra il prezzo P_t di un bene o un insieme di beni a una certa data e il prezzo P_0 dello stesso bene o insieme di beni a una data scelta come periodo di riferimento.

$$I_t = \frac{P_t}{P_0} \cdot 100$$

Gli indici dei prezzi al consumo possono essere:

- **indici a base fissa** se i prezzi ai diversi tempi vengono divisi per il prezzo a una data fissata e poi moltiplicati per 100;
- **indici a base mobile** se ogni prezzo viene diviso per il prezzo alla data precedente e poi moltiplicato per 100.

ESEMPIO

La tabella contiene il prezzo di 1 kg di pane in quattro anni diversi.

Anno	2013	2014	2015	2016
Prezzo (€)	2,62	2,71	2,63	2,80

Calcoliamo gli indici a base fissa del prezzo del pane con base il 2013.

$$I_{2013} = \frac{2,62}{2,62} \cdot 100 = 100; \qquad I_{2014} = \frac{2,71}{2,62} \cdot 100 = 103,44;$$

$$I_{2015} = \frac{2,63}{2,62} \cdot 100 = 100,38; \qquad I_{2016} = \frac{2,80}{2,62} \cdot 100 = 106,87.$$

Calcoliamo gli indici a base mobile del prezzo del pane.

TEORIA

Non conoscendo il prezzo nel 2012, non possiamo calcolare l'indice del 2013.

$$I_{2014} = \frac{2,71}{2,62} \cdot 100 = 103,44; \qquad I_{2015} = \frac{2,63}{2,71} \cdot 100 = 97,05;$$

$$I_{2016} = \frac{2,80}{2,63} \cdot 100 = 106,46.$$

Gli indici dei prezzi al consumo godono delle seguenti proprietà:

- sono sempre positivi;
- se $I > 100$, il prezzo è aumentato rispetto a quello nel periodo di riferimento (serve più denaro per acquistare lo stesso bene);
- se $I = 100$, il prezzo non è cambiato (serve la stessa quantità di denaro per acquistare lo stesso bene);
- se $I < 100$, il prezzo è diminuito (serve meno denaro per acquistare lo stesso bene);

La quantità $I - 100$ rappresenta la variazione percentuale del prezzo tra i due periodi. Riferendoci all'esempio precedente, la variazione percentuale del prezzo del pane tra il 2013 e il 2016 è $106,87\% - 100\% = 6,87\%$, quella tra il 2015 e il 2016 è $106,46\% - 100\% = 6,46\%$.
Gli indici dei prezzi al consumo sono statisticamente interessanti quando rilevati e calcolati per più anni in modo da poter calcolare l'**indice medio dei prezzi**.
Consideriamo la seguente tabella, che riporta i dati per un indice dei prezzi al consumo a base mobile in Italia per un certo prodotto.

Anno	2010	2011	2012
Indice	101,6	102,7	103,0

Supponiamo che nel 2010 il prezzo del prodotto fosse € 50.
Se chiamiamo rispettivamente P_{2011} e I_{2011} il prezzo e l'indice del 2011, in base a come viene calcolato l'indice a base mobile, per l'indice I_{2011} abbiamo:

$$I_{2011} = 102,7 = \frac{P_{2011}}{50} \cdot 100 \rightarrow P_{2011} = 50 \cdot \overbrace{\frac{102,7}{100}}^{I_{2011}} = 51,35.$$

Ragionando analogamente per P_{2012}, otteniamo:

$$P_{2012} = P_{2011} \cdot \frac{103,0}{100} = \underbrace{50 \cdot \overbrace{\frac{102,7}{100}}^{I_{2011}}}_{P_{2011}} \cdot \overbrace{\frac{103,0}{100}}^{I_{2012}} = 52,89.$$

Pertanto, otteniamo la seguente relazione, che lega P_{2012} a P_{2010}:

$$\overbrace{50}^{P_{2010}} \cdot \frac{I_{2011}}{100} \cdot \frac{I_{2012}}{100} = \overbrace{52,89}^{P_{2012}}.$$

Vogliamo calcolare l'indice medio dei prezzi tra il 2010 e il 2012, cioè quel valore che sostituito a I_{2011} e I_{2012} restituisce lo stesso costo per la spesa nel 2012. Poiché nella relazione precedente i due indici sono moltiplicati fra loro, la media da utilizzare è la media geometrica. Otteniamo:

$$I_m = \sqrt{I_{2011} \cdot I_{2012}} = \sqrt{102,7 \cdot 103,0} \simeq 102,850.$$

Infatti:

$$50 \cdot \frac{I_m}{100} \cdot \frac{I_m}{100} \simeq 52,89.$$

Il valore di I_m trovato ci dice che i prezzi sono aumentati in media ogni anno del 2,85%.

In generale, l'**indice medio dei prezzi** è la media geometrica degli indici dei prezzi a base mobile.

$$I_m = \sqrt[n]{I_1 \cdot I_2 \cdot \ldots \cdot I_n}.$$

4 Statistica inferenziale

In una scuola è stata effettuata un'indagine censuaria per avere informazioni sull'altezza dei ragazzi che la frequentano. Sono state perciò misurate le altezze dei 500 studenti della scuola, ottenendo i risultati in tabella.

Altezza (cm)	Frequenza relativa
155-164	0,12
165-174	0,35
175-184	0,4
185-194	0,1
195-204	0,03

Da questi dati possiamo ricavare l'altezza media,

$$M = 159{,}5 \cdot 0{,}12 + 169{,}5 \cdot 0{,}35 + 179{,}5 \cdot 0{,}4 + 189{,}5 \cdot 0{,}1 + 199{,}5 \cdot 0{,}03 = 175{,}2,$$

e la deviazione standard,

$$\sigma = \sqrt{15{,}7^2 \cdot 0{,}12 + 5{,}7^2 \cdot 0{,}35 + 4{,}3^2 \cdot 0{,}4 + 14{,}3^2 \cdot 0{,}1 + 24{,}3^2 \cdot 0{,}03} \simeq 9{,}3.$$

L'altezza media, la deviazione standard e tutti gli altri indici che possiamo calcolare a partire da questi dati rappresentano dei **parametri** che sintetizzano sulla popolazione il carattere *altezza*.

DEFINIZIONE

I **parametri** della popolazione sono i valori di sintesi calcolati conoscendo la modalità del carattere di ogni unità statistica della popolazione.

Supponiamo ora che non sia invece possibile misurare le altezze di tutti gli studenti e che si debba perciò eseguire un'indagine campionaria.
Il primo passo è scegliere opportunamente il **campione**, cioè un sottoinsieme della popolazione su cui effettuare l'indagine.
Le unità che faranno parte del campione devono essere scelte in modo casuale, senza che influiscano elementi soggettivi.
Per esempio, se si scegliessero solo gli studenti che praticano pallavolo o pallacanestro oppure tutti quelli di prima o quelli di quinta, non avremmo un campione che rappresenta tutta la popolazione.

Esistono diversi metodi per scegliere il campione casualmente: il più elementare è il **campionamento con ripetizione** o **bernoulliano**.
Il campionamento con ripetizione si può realizzare mediante l'estrazione a sorte di palline da un'urna. Si assegna a ogni unità della popolazione un numero progressivo, si inseriscono in un'urna altrettante palline numerate e si estraggono le palline una alla volta, reinserendole nell'urna dopo ogni estrazione, finché non si è estratto un numero di palline uguale alla numerosità del campione.
Il campione così estratto è casuale, infatti:

- ogni unità della popolazione ha la stessa probabilità di essere scelta;
- tutti i campioni di uguale numerosità hanno la stessa probabilità di essere estratti.

L'insieme di tutti i campioni di dimensione n estraibili da una certa popolazione costituisce lo **spazio dei campioni** (o **campionario**) di dimensione n.
Si può calcolare che, se la popolazione è formata da N unità statistiche, il numero di possibili campioni di dimensione n è N^n.

TEORIA

▶ Quanti sono i campioni di due unità estratti da una popolazione di quattro unità, A, B, C e D? Scrivili tutti.

Considerando la scuola con 500 studenti, lo spazio dei campioni di dimensione 30 è formato da 500^{30} campioni.

Prendiamo uno di questi campioni e supponiamo che le unità statistiche così selezionate presentino le seguenti modalità, espresse in centimetri.

163	167	171	175	164	180
176	180	175	178	169	159
171	176	181	159	178	179
169	176	177	178	178	167
171	168	176	177	176	179

Su questo campione la media aritmetica delle altezze, approssimata all'unità, è:

$$\bar{x} = \frac{159 \cdot 2 + 163 + 164 + 167 \cdot 2 + \ldots + 180 \cdot 2 + 181}{30} \simeq 173.$$

Il valore ottenuto è valido solo sul campione selezionato ed è una **stima** dell'altezza media di tutta la popolazione.

DEFINIZIONE

Una **stima** di un parametro è un valore calcolato conoscendo le modalità del carattere solo per le unità statistiche del campione a disposizione ed è un'approssimazione del parametro della popolazione.

A differenza dei parametri, che sono valori costanti, le stime variano da campione a campione.
Supponiamo, per esempio, di aver selezionato un altro campione di 30 studenti che presentano le seguenti modalità, espresse in centimetri.

177	167	171	175	180	179
179	175	175	175	178	169
179	171	176	179	178	179
179	176	177	178	168	180
178	168	180	177	176	179

Su questo campione la media aritmetica delle altezze è diversa da quella calcolata sul campione precedente:

$$\bar{x} = \frac{167 + 168 \cdot 2 + 169 + 171 \cdot 2 + 175 \cdot 4 + \ldots + 180 \cdot 3}{30} \simeq 176.$$

Analogamente, se avessimo calcolato le deviazioni standard sui due campioni, avremmo ottenuto due risultati diversi tra loro ed entrambi diversi dal parametro deviazione standard calcolato sulla popolazione.

La **statistica inferenziale** permette di utilizzare le stime ottenute su un campione al posto dei corrispondenti parametri, qualora questi ultimi non fossero calcolabili, controllando o valutando l'errore campionario che si commette.

Stima puntuale della media

▶ Esercizi a p. λ44

Torniamo all'esempio degli studenti della scuola e consideriamo tutti i 500^{30} possibili campioni formati da 30 unità statistiche. Su ciascun campione calcoliamo la media aritmetica dei dati (approssimiamo all'unità, dal momento che non ci serve la precisione dei millimetri) ottenendo 500^{30} medie i cui valori con le corrispondenti frequenze relative sono riportati nella tabella 5. Ogni media su un campione è una **stima puntuale** dell'altezza media della popolazione.

Se stiamo svolgendo un'indagine campionaria, vuol dire che abbiamo a disposizione un solo campione. Come facciamo a valutare l'attendibilità e la precisione della stima su un solo campione?

L'insieme X delle medie calcolate sui campioni che hanno la stessa numerosità (nel nostro caso 30 unità statistiche) si chiama **media campionaria**.
Calcoliamo il valore medio della media campionaria:

$$M_X = 167 \cdot 0{,}01 + 168 \cdot 0{,}03 + 171 \cdot 0{,}05 + 172 \cdot 0{,}09 + 173 \cdot 0{,}12 + 175 \cdot 0{,}26 +$$
$$176 \cdot 0{,}15 + 177 \cdot 0{,}12 + 178 \cdot 0{,}1 + 179 \cdot 0{,}04 + 180 \cdot 0{,}03 = 175.$$

Il valore medio della media campionaria è uguale all'altezza media calcolata sull'intera popolazione. Questo vuol dire che, anche se le stime sui singoli campioni sono diverse, mediamente otteniamo il valore del parametro da stimare. Statisticamente questo si esprime dicendo che la **stima** è **corretta**.

DEFINIZIONE
Una **stima** è **corretta** se il valore medio delle stime su tutti i campioni di uguale numerosità è uguale al parametro che si vuole stimare.

Errore campionario

Anche se mediamente la media campionaria è uguale al parametro della popolazione e quindi ogni stima è un'approssimazione accettabile, può essere interessante valutare la sua variabilità al variare dei campioni e quindi l'**errore campionario** commesso considerando la stima al posto del parametro.

Per farlo usiamo la **deviazione standard della media campionaria**.

$$\sigma_X = \sqrt{(167-175)^2 \cdot 0{,}01 + (168-175)^2 \cdot 0{,}03 + \ldots + (180-175)^2 \cdot 0{,}03} \simeq 1{,}7.$$

Osserviamo che otteniamo lo stesso risultato calcolando:

$$\sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{9{,}3}{\sqrt{30}} \simeq 1{,}7.$$

(σ: deviazione standard sulla popolazione; n: numerosità del campione)

Vale infatti il seguente teorema.

TEOREMA
Se il campionamento è con ripetizione, la **deviazione standard della media campionaria** σ_X è uguale al rapporto tra la deviazione standard del carattere nella popolazione e la radice quadrata del numero n di unità nel campione.

$$\sigma_X = \frac{\sigma}{\sqrt{n}}$$

La deviazione standard della media campionaria è anche detta **errore standard**.

Come ci aspettiamo, maggiore è la numerosità n del campione, minore è la variabilità delle stime e quindi l'errore standard. Nel caso limite, cioè se la numerosità del campione è uguale alla numerosità della popolazione, lo spazio dei campioni è formato da un solo elemento: tutta la popolazione. La stima in tal caso coincide con il parametro della popolazione, quindi non c'è errore campionario.
Viceversa, a parità di n, maggiore è la deviazione standard calcolata sui dati della popolazione, maggiore è l'errore. Infatti, se la deviazione standard sulla popolazione è grande, vuol dire che i dati sulla popolazione sono più dispersi, quindi è

Media sul campione (cm)	Frequenza relativa
167	0,01
168	0,03
171	0,05
172	0,09
173	0,12
175	0,26
176	0,15
177	0,12
178	0,1
179	0,04
180	0,03

↑ Tabella 5

▶ Su una popolazione di 4 unità sono stati rilevati i seguenti redditi annui (in migliaia di euro): 13, 15, 26, 28. Calcola la media M e la varianza σ^2 dei redditi sulla popolazione. Considerando poi tutti i campioni di ampiezza 2 estraibili, scrivi i valori che assume la media campionaria e verifica che il suo valore medio è M e la sua varianza è $\frac{\sigma^2}{2}$.

TEORIA

possibile che il campione estratto contenga dati anomali che si discostano dalla media della popolazione, alterandone così la stima.

Considerando la stima ottenuta sul primo campione e l'errore standard, possiamo valutare che l'altezza media dei ragazzi della scuola è 173 cm con un errore, in più o in meno, di 1,7 cm.

Tuttavia, se il nostro scopo è trovare una stima del parametro media, a partire da un campione, non abbiamo a disposizione l'altezza media di tutta la popolazione e quindi neppure la deviazione standard per poter determinare l'errore standard. Per risolvere questo problema possiamo procedere in due modi:

- dedurre la deviazione standard sulla popolazione da precedenti indagini;
- utilizzare una stima della deviazione standard a partire dal campione a disposizione.

DEFINIZIONE

Siano $x_1, \ldots, x_n$ i dati rilevati su un campione di numerosità n estratto da una popolazione e sia $\bar{x}$ la media calcolata sul campione. Una **stima corretta della deviazione standard** è:

$$s_c = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}}.$$

Utilizzando la stima corretta della deviazione standard, l'errore standard diventa:

$$s_{\bar{x}} = \frac{s_c}{\sqrt{n}}.$$

Nel nostro caso, avendo a disposizione solo i dati sul primo campione, avremmo, per esempio:

$$s_c = \sqrt{\frac{(159-173)^2\cdot 2+(163-173)^2+\ldots+(180-173)^2\cdot 2+(181-173)^2}{30-1}} \simeq 6{,}25.$$

e quindi errore standard $s_{\bar{x}} = \frac{6{,}25}{\sqrt{30}} \simeq 1{,}14$.

ESEMPIO

Per avere una stima delle ore di permesso mensili richieste dai dipendenti di un'azienda sono stati intervistati dieci impiegati ottenendo i risultati in tabella.

Ore di permesso	0	2	4	5	6
Frequenza	2	4	2	1	1

Una stima puntuale del numero medio delle ore di permesso è:

$$\bar{x} = \frac{0\cdot 2+2\cdot 4+4\cdot 2+5\cdot 1+6\cdot 1}{10} = 2{,}7.$$

Determiniamo l'errore standard. Calcoliamo la stima della deviazione standard utilizzando i dati del campione,

$$s_c = \sqrt{\frac{(0-2{,}7)^2\cdot 2+(2-2{,}7)^2\cdot 4+(4-2{,}7)^2\cdot 2+(5-2{,}7)^2\cdot 1+(6-2{,}7)^2\cdot 1}{9}} \simeq 2,$$

e da questa l'errore standard:

$$s_{\bar{x}} = \frac{2}{\sqrt{10}} \simeq 0{,}63.$$

► Su un campione di 20 persone è stato rilevato che il 5% di esse percorre 20 km al giorno per recarsi a lavoro, il 20% 10 km, il 30% solo 5 km e il resto 30 km. Determina una stima della media dei kilometri percorsi per andare al lavoro, una stima della deviazione standard e l'errore commesso con la stima puntuale della media.

Animazione

Possiamo stimare che il numero medio delle ore mensili richieste dai dipendenti è $2{,}7 \pm 0{,}63$.

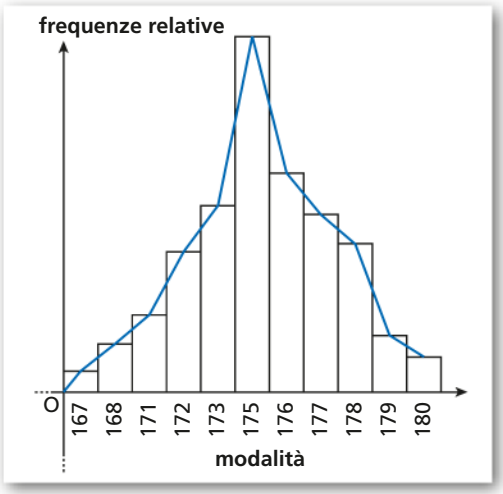
Stima della media per intervallo

▶ Esercizi a p. λ46

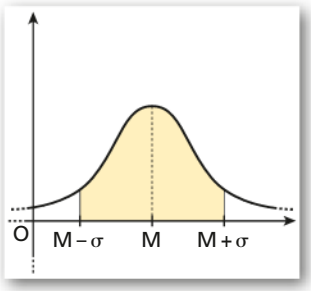
Un diverso tipo di stima è quella per intervallo: non è un numero ma un intervallo in cui ci aspettiamo, con un certo **grado di fiducia** deciso a priori, che si trovi la media della popolazione. Lo chiamiamo **intervallo di fiducia**.

Per capire come ottenere un intervallo di fiducia a partire dai dati su un campione, ripartiamo dalla media campionaria.

Consideriamo il poligono delle frequenze costruito usando le frequenze relative della tabella 5 di pagina λ17. Sulle ascisse mettiamo le modalità e sulle ordinate le frequenze relative di ogni modalità.



Il poligono delle frequenze ha una forma che, in modo approssimato, è simile a quella della curva di Gauss, rappresentata a fianco.



Vale infatti il seguente teorema.

TEOREMA

Se il numero di unità nel campione è abbastanza grande ($n \geq 30$), la **media campionaria** ha una distribuzione ben approssimata da una curva di Gauss simmetrica rispetto alla media della distribuzione.

Nel nostro esempio e nella maggior parte delle indagini campionarie, cioè per tutte le indagini che si occupano di caratteri quantitativi per cui si vuole determinare il valore medio e i cui campioni hanno numerosità maggiore di 30, possiamo dunque pensare che la distribuzione della media campionaria sia gaussiana, cioè abbia un poligono delle frequenze approssimabile con la curva di Gauss e quindi goda delle proprietà di questa curva.

Proprietà della curva di Gauss

Possiamo pensare alla curva di Gauss come al poligono delle frequenze relative di una distribuzione continua: è quella a cui ci avviciniamo sempre più se aumentiamo la numerosità dei campioni e la precisione delle misurazioni nel caso delle altezze. Ciò vuol dire che:

- la frequenza maggiore si ha in corrispondenza della media della distribuzione;
- se ci discostiamo in più o in meno della stessa quantità dalla media, abbiamo la stessa frequenza relativa;
- l'area sotto la curva e sopra l'asse delle ascisse rappresenta la somma di tutte le frequenze relative, per cui è uguale a 1.