Il campionamento

Obiettivi

- utilizzare le principali tecniche di campionamento
- comprendere il significato di variabile campionaria e di stimatore
- determinare i valori di sintesi di una variabile campionaria

1. Popolazione e campione

Gli esercizi di questo paragrafo sono a pag. 32

1.1 Il campionamento

Le nozioni di statistica acquisite finora permettono di condurre l'analisi di un fenomeno sulla base di un certo numero di dati rilevati. Per esempio, aver analizzato i voti in una materia degli studenti di una scuola al termine del primo quadrimestre, permette di conoscere qual è l'andamento in media in quella materia, quanti studenti sono al di sotto della media, come si disperdono i dati attorno alla media, quanti sono gli studenti insufficienti, in modo da poter prendere decisioni sugli interventi da mettere eventualmente in atto per migliorare la situazione.
Ma l'analisi fatta si ferma a quella particolare scuola e non è possibile generalizzare i risultati estendendoli ad altre. Per avere delle informazioni che possano riguardare una popolazione più vasta, per esempio gli studenti delle scuole superiori di tutto il territorio nazionale, si dovrebbero rilevare i dati di tutti gli studenti di tutte le scuole italiane. E' evidente che una rilevazione del genere comporterebbe costi molto elevati in termini di denaro e di tempo, nonché l'impegno di numerose persone.
In alcune situazioni non si può fare a meno di eseguire rilevazioni sull'intera popolazione, è il caso per esempio dei censimenti (l'ultimo è del 2011), ma nella maggior parte dei casi l'indagine viene condotta su un numero limitato di unità statistiche, sia per limitare i costi, sia, a volte, perché è impossibile l'analisi dell'intera popolazione; per esempio, in una ricerca sulla durata di un elettrodomestico, non è possibile testare l'intera produzione che andrebbe in questo modo persa.
Quello che si fa in questi casi è ricorrere ad un campione.

Campione e tasso di campionamento

Un **campione** è un sottoinsieme proprio di una popolazione P formato solo da alcuni elementi di P scelti in base ad un criterio stabilito.

Se N è il numero di elementi che costituisce P e n è il numero di elementi che forma il campione, il rapporto $\frac{n}{N}$ viene detto **tasso di campionamento**.

Per esempio:

- se P è l'insieme dei 1200 studenti di una scuola e da P si vuole estrarre un campione di ampiezza 60, il tasso di campionamento è $\frac{60}{1200} = 0{,}05$, che, in termini percentuali equivale al 5%;
- viceversa, se si considera un tasso di campionamento del 12% da una popolazione di ampiezza 5000, il campione avrà ampiezza $5000 \cdot \frac{12}{100} = 600$.

Le tecniche di campionamento

Una **procedura di campionamento** consiste nel definire la metodologia, dipendente anche dal tipo di indagine che si deve svolgere, che consente di scegliere le unità della popolazione che devono far parte del campione.
Si può fare una prima distinzione tra *campionamenti probabilistici* e *campionamenti non probabilistici*.

CAMPIONAMENTI PROBABILISTICI

Un campionamento è **probabilistico** se ciascuna unità ha una probabilità nota di essere estratta e quindi di far parte del campione.

Tra i metodi di campionamento probabilistico ricordiamo i seguenti.

Il campionamento casuale semplice

Caratteristica fondamentale di questo metodo è che tutte le unità statistiche hanno la stessa probabilità di essere estratte. Si procede inizialmente assegnando un numero progressivo da 1 a N ad ogni unità della popolazione; l'estrazione delle n unità che costituiscono il campione avviene poi tramite una tavola di numeri casuali, generati per esempio da un computer.
Questo tipo di campionamento è indubbiamente il più semplice, permette di far riferimento ai modelli più elementari di calcolo delle probabilità e garantisce una scelta obiettiva degli elementi del campione, con un rischio minimo di invalidare i risultati; inoltre consente di trasferire in modo relativamente semplice i valori dei parametri calcolati dal campione all'intera popolazione.

Il campionamento casuale stratificato

In base a questo procedimento occorre:

- suddividere la popolazione in fasce il più possibile omogenee al loro interno rispetto al carattere oggetto dell'osservazione (ogni unità statistica può far parte di una sola fascia)
- estrarre casualmente un certo numero di unità da ogni fascia.

Questo metodo si usa frequentemente quando la popolazione presenta caratteristiche molto diverse tra loro. Per esempio, in una indagine sull'incidenza di una certa patologia è opportuno suddividere la popolazione per fasce di età, onde evitare che in una estrazione casuale compaiano prevalentemente persone giovani o prevalentemente persone anziane; all'interno

di ogni fascia si può poi prevedere un campionamento di ampiezza proporzionale alla dimensione della fascia, oppure privilegiare le fasce in cui vi è un'alta differenziazione tra le unità statistiche.
Un campionamento di questo tipo garantisce una migliore rappresentatività rispetto ad uno casuale semplice e genera quindi stime più efficienti; tuttavia è più costoso del precedente ed occorre stare molto attenti nella costruzione delle fasce per evitare di ottenere risultati fuorvianti.

Il campionamento casuale a grappoli

In diversi casi di indagine statistica, la popolazione è già suddivisa in modo naturale in sottogruppi denominati *grappoli* (clusters), che costituiscono una partizione dell'intera popolazione. Per esempio la popolazione degli studenti di una scuola è naturalmente suddivisa in classi, ciascuna delle quali è un grappolo.
Il campionamento avviene selezionando casualmente un certo numero di grappoli e componendo il campione con tutti gli elementi dei grappoli selezionati.
Il pregio di questo metodo è che ha dei costi molto bassi, ma è efficace solo se i grappoli non sono eccessivamente numerosi e se presentano elevata omogeneità tra le unità di ogni grappolo.

Si esegue una partizione di un insieme se si suddividono i suoi elementi in sottoinsiemi a due a due disgiunti la cui unione restituisce l'insieme dato.

Il campionamento sistematico

Si usa con una popolazione i cui elementi sono ordinati e numerabili progressivamente; consiste nell'estrarre gli elementi che sono distanziati da un intervallo costante. Per esempio, nel controllo di una produzione in serie, si può decidere di formare il campione scegliendo il primo elemento a caso e gli altri a distanza di 15 uno dall'altro a partire dal primo scelto.
Questa modalità di campionamento rientra tra quelle casuali in quanto tale è la scelta del primo elemento.

Campionamenti non probabilistici

In molte indagini statistiche non è possibile estrarre campioni casuali sostanzialmente perché la popolazione non è accessibile nella sua totalità; l'indagine si avvale in questi casi delle unità statistiche disponibili.
Pensiamo per esempio alle ricerche in campo medico, dove la sperimentazione di un farmaco avviene di solito su un campione di volontari che non possono essere ovviamente selezionati a caso.
Anche le indagini che avvengono mediante questionari inviati per posta rientrano in questa categoria; infatti, pur essendo il campione iniziale di tipo probabilistico, l'analisi dei dati viene fatta solo sulle persone che rispondono, che hanno quindi caratteristiche diverse da quelle che non hanno risposto.
Il campionamento non probabilistico, proprio per come viene effettuato, non fornisce a ciascuna unità della popolazione la stessa occasione di essere scelta a far parte del campione; al contrario, alcuni gruppi o individui hanno maggiore probabilità di essere scelti. Questo tipo di campionamento non è quindi molto affidabile ed è sconsigliato a meno di ragionevoli giustificazioni.

La scelta di un campione, comunque venga fatta, è un passaggio molto delicato in una indagine statistica, perché i risultati che si ottengono devono poi essere trasferiti all'intera popolazione (la *statistica inferenziale* è quella che si occupa di questi temi) e un errore nella scelta del campione può dare risultati totalmente inaffidabili.
E' quello che succede a volte nei sondaggi, per esempio durante le elezioni, quando il risultato ufficiale smentisce quello previsto. Ricordiamo un errore

clamoroso avvenuto durante le elezioni presidenziali negli Stati Uniti nel 1936, quando un sondaggio condotto per telefono dava Alf Landon vincente contro Franklin Roosevelt con un margine elettorale di 370 voti contro 161; i risultati reali furono di 523 a 8 a favore di Roosvelt. L'errore fu l'aver condotto l'indagine per telefono: dopo la crisi del 1929 solo le persone più agiate potevano permettersi un telefono e questo gruppo di persone votava tendenzialmente per i Repubblicani il cui candidato era Landon.

Il campionamento casuale semplice

Tra quelli evidenziali ci occupiamo in dettaglio di questa tipologia di campionamento perché, come abbiamo già detto, è più semplice delle altre.
In particolare trattiamo il **campionamento bernoulliano** nel quale la formazione del campione avviene mediante estrazione con ripetizione, cioè con la reimmissione dell'unità estratta nella popolazione.
Questo comporta che ogni unità statistica potrebbe essere estratta più volte, ma se la popolazione è sufficientemente grande rispetto al campione, cioè n è molto minore di N, il rischio che si corre è minimo rispetto ai vantaggi che ne derivano.

IL NUMERO DI CAMPIONI ESTRAIBILI

L'estrazione di un campione può essere considerata un evento aleatorio in quanto, una volta fissata l'ampiezza n, di campioni ne possono essere estratti più di uno; ma quanti campioni si possono formare?
Per dare una risposta a questa domanda, ragioniamo col criterio dell'urna.
Poiché ogni elemento, dopo essere stato estratto viene rimesso nell'urna, le estrazioni sono eventi indipendenti; se consideriamo diversi due campioni i cui elementi differiscono non solo per gli elementi estratti, ma anche per l'ordine di estrazione, il numero di campioni che si possono estrarre è uguale al numero di disposizioni con ripetizione di N elementi a gruppi di n, cioè:

numero dei campioni $\qquad D'^{(r)}_{N,n} = N^n$

Per esempio, il numero di campioni bernoulliani di ampiezza 10 estratti da una popolazione di 50 elementi è 50^{10} e ogni campione ha probabilità $\frac{1}{50^{10}}$ di essere estratto.

1.2 Le variabili campionarie

Per comprendere i concetti che dobbiamo esporre, prendiamo in considerazione una popolazione molto semplice, formata da soli 4 elementi:

popolazione: $\qquad \{a, b, c, d\}$

e consideriamo i campioni di ampiezza 2 che si possono estrarre da essa; sappiamo che il loro numero è $4^2 = 16$.
L'estrazione del primo elemento del campione è un esperimento aleatorio che può avere come esito un qualunque elemento della popolazione stessa; il primo elemento è quindi una variabile aleatoria X_1 che può assumere i valori a, b, c, d ciascuno con probabilità $\frac{1}{4}$.
Analogamente, anche l'estrazione del secondo elemento dà origine a una va-

riabile aleatoria X_2 che può anch'essa assumere i valori a, b, c, d ciascuno con la stessa probabilità $\frac{1}{4}$.

Il campione di ampiezza 2 è quindi un vettore aleatorio (X_1, X_2) avente come spazio campionario le coppie (x_i, x_j) che si possono formare dalla popolazione con le regole stabilite dalla procedura di campionamento.
Generalizziamo queste osservazioni e consideriamo una popolazione X costituita da N elementi x_k dalla quale si vuole estrarre un campione di ampiezza n, con $n < N$.

L'esito della i-esima estrazione di un elemento del campione è una variabile aleatoria X_i che può assumere i valori

$x_1 \quad x_2 \quad ..... \quad x_N$ con probabilità $p_1 \quad p_2 \quad ..... \quad p_N$

essendo p_k la frequenza relativa all'elemento x_k della popolazione.
Le n variabili X_i prendono il nome di **variabili campionarie**.

Ogni campione di ampiezza n è dunque un **vettore aleatorio** $(X_1, X_2, ..., X_n)$ le cui componenti sono le variabili aleatorie X_i.

Ogni possibile n-pla di osservazioni $(x_1, x_2, ..., x_n)$, cioè ogni estrazione di un particolare campione, rappresenta la **realizzazione** del campione; lo spazio campionario di tale vettore aleatorio è l'insieme Ω di tutte le realizzazioni campionarie di ampiezza n estraibili dalla popolazione X.

Con le lettere maiuscole si indicano le variabili aleatorie, con le stesse lettere minuscole le realizzazioni del campione.

Tornando all'esempio iniziale, la popolazione ha la seguente distribuzione:

$$X = \begin{Bmatrix} a & b & c & d \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix}$$

Con un campionamento bernoulliano, le variabili campionarie X_1 e X_2 hanno la stessa identica distribuzione:

$$X_1 = \begin{Bmatrix} a & b & c & d \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix} \qquad X_2 = \begin{Bmatrix} a & b & c & d \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix}$$

e una possibile realizzazione del campione è la coppia (c, a), che significa che alla prima estrazione la variabile aleatoria X_1 ha assunto valore c, alla seconda estrazione la variabile aleatoria X_2 ha assunto valore a.

VERIFICA DI COMPRENSIONE

1. Si vuole estrarre un campione da una popolazione di 1500 elementi con un tasso di campionamento del 4%; il campione è composto da:

a. 15 elementi **b.** 40 elementi **c.** 60 elementi **d.** 80 elementi

2. Un'azienda produce 80 pezzi al giorno; in quanti modi si può estrarre un campione bernoulliano di 4 elementi da sottoporre a controllo?

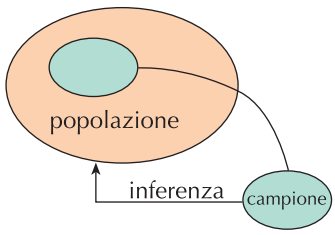
a. 4^{80} **b.** 80^4 **c.** $80 \cdot 4$ **d.** $\frac{80}{4}$

2. PARAMETRI E STIMATORI

Gli esercizi di questo paragrafo sono a pag. 33

Sappiamo che una popolazione statistica X può essere sintetizzata da alcuni indici: un valore centrale (media, moda o mediana), lo scarto quadratico medio, la varianza. Un qualunque indice che caratterizza in modo sintetico una popolazione statistica si dice **parametro** della popolazione.
In generale però, a meno di effettuare l'analisi dell'intera popolazione, non siamo in grado di determinare i suoi parametri; scelto un campione, quello che possiamo fare è trovare i parametri del campione e chiederci in quale misura tali parametri possano dare indicazioni su quelli corrispondenti della popolazione (***figura 1***).

Figura 1



Diciamo che un qualsiasi parametro calcolato su un campione costituisce una **stima** del corrispondente parametro della popolazione.

*La **statistica inferenziale** si occupa di studiare i parametri di una popolazione utilizzando i dati ottenuti su campioni estratti da essa.*

Per esempio, se la media dei pesi di un campione di scatole di zucchero di una certa azienda è 498g, tale valore è una stima della media dei pesi di tutte le scatole di zucchero che costituiscono la produzione di quell'azienda.
Ma, mentre i parametri di una popolazione sono fissi, quelli di un campione sono variabili perché dipendono dal campione scelto; un diverso campione di scatole di zucchero potrebbe dare un valore medio di 502g e un altro ancora di 500g.
Le stime di un parametro sono quindi funzioni delle variabili campionarie X_i.

Chiamiamo **stimatore**, o anche **statistica**, e lo indichiamo con il simbolo T_n, la variabile campionaria generata dalle stime calcolate su tutti i possibili campioni estraibili dalla popolazione.

Uno stimatore, per essere efficace, deve avere alcune caratteristiche che elenchiamo di seguito.

Si dice che uno stimatore T_n è:

- **corretto**, o anche **non distorto**, se il suo valore atteso è uguale al parametro ϑ che deve stimare
- **consistente** se, al crescere dell'ampiezza n del campione, la sua varianza tende a zero, cioè:

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

Se un parametro può essere valutato tramite due stimatori diversi T_{1n} e T_{2n}, diciamo che:

- T_{1n} è **più efficiente** di T_{2n} se la varianza di T_{1n} è minore della varianza di T_{2n} :

$$V(T_{1n}) < V(T_{2n})$$

I principali stimatori dei parametri di una popolazione sono la media, la varianza e la proporzione campionaria.

2.1 La media campionaria

Si chiama media campionaria di un campione casuale X_1, X_2, ..., X_n la quantità $\overline{X_n}$ definita dalla relazione:

$$\overline{X_n} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

La media campionaria dipende dal campione, quindi $\overline{X_n}$ **è una variabile aleatoria** di cui si può facilmente calcolare valore atteso e varianza.
Vediamo prima un semplice esempio. Supponiamo che la popolazione sia costituita dai pesi in quintali del raccolto di cinque campi di ugual superficie coltivati a pomodori:

$$X = \{65,\ 72,\ 70,\ 68,\ 55\}$$

La media e la varianza di questa popolazione sono:

$$\mu = E(X) = 66 \qquad \sigma^2 = 35{,}6$$

Consideriamo i campioni di ampiezza due che si possono ottenere con un campionamento bernoulliano:

65; 65	72; 65	70; 65	68; 65	55; 65
65; 72	72; 72	70; 72	68; 72	55; 72
65; 70	72; 70	70; 70	68; 70	55; 70
65; 68	72; 68	70; 68	68; 68	55; 68
65; 55	72; 55	70; 55	68; 55	55; 55

Il numero dei campioni è $5^2 = 25$ e ciascuno ha probabilità $\frac{1}{25} = 0{,}04$ di essere estratto.

Per costruire la distribuzione di $\overline{X}_2$, dobbiamo calcolare i valori medi di ogni campione dello spazio Ω; ogni valore ottenuto è una stima della media della popolazione

65	68,5	67,5	66,5	60
68,5	72	71	70	63,5
67,5	71	70	69	62,5
66,5	70	69	68	61,5
60	63,5	62,5	61,5	55

Ordinando per questioni di praticità le medie ottenute e considerando la frequenza con cui ognuna di esse compare, abbiamo che $\overline{X}_2$ ha la seguente distribuzione

$\overline{X_2}$	55	60	61,5	62,5	63,5	65	66,5	67,5	68	68,5	69	70	71	72
Frequenza	1	2	2	2	2	1	2	2	1	2	2	3	2	1
Probabilità	0,04	0,08	0,08	0,08	0,08	0,04	0,08	0,08	0,04	0,08	0,08	0,12	0,08	0,04

Calcoliamo il valor medio di questa distribuzione $E(\overline{X}_2) = \sum_{i=1}^{14} \overline{x}_i \cdot p_i = 66$

Si nota subito che il valore trovato coincide con il valor medio della popolazione; questo significa che, all'aumentare del numero dei campioni, il valore medio tende al valore medio della popolazione.

Per calcolare la varianza riscriviamo la tabella aggiungendo una riga per i valori di $\overline{X}_2^2$ (eliminiamo la riga delle frequenze)

$\overline{X}_2$	55	60	61,5	62,5	63,5	65	66,5	67,5	68	68,5	69	70	71	72
$\overline{X}_2^2$	3025	3600	3782,25	3906,25	4032,25	4225	4422,25	4556,25	4624	4692,25	4761	4900	5041	5184
Prob.	0,04	0,08	0,08	0,08	0,08	0,04	0,08	0,08	0,04	0,08	0,08	0,12	0,08	0,04

Allora $E\left(\overline{X}_2^2\right) = \sum_{i=1}^{14} \overline{x}_i^2 \cdot p_i = 4373{,}8$

$$V(\overline{X}_2) = E(\overline{X}_2^2) - \left[E(\overline{X}_2)\right]^2 = 4373{,}8 - 4356 = 17{,}8$$

Osserviamo che il valore trovato per la varianza della media campionaria è la metà della varianza della popolazione.

I risultati ottenuti con questo esempio sono validi in generale; vale infatti il seguente teorema.

Teorema. In un campionamento bernoulliano, il valor medio della media campionaria $\overline{X}_n$ è uguale alla media μ della popolazione e la varianza della media campionaria è uguale alla varianza σ^2 della popolazione divisa per l'ampiezza del campione. In simboli

$$E(\overline{X}_n) = \mu \qquad V(\overline{X}_n) = \frac{\sigma^2}{n}$$

Per indicare il valor medio della media campionaria usiamo il simbolo $\mu_{\overline{X}}$ e per indicare la varianza usiamo $\sigma^2_{\overline{X}}$.

Il risultato a cui siamo giunti con questo teorema indica due aspetti importanti:

- la media campionaria è uno **stimatore corretto** della media della popolazione
- inoltre, essendo la varianza campionaria più piccola del fattore $\frac{1}{n}$ della varianza della popolazione, la dispersione della media campionaria intorno a μ diminuisce al crescere dell'ampiezza n del campione; in altre parole, la varianza della media campionaria tende a zero al crescere di n :

$$\lim_{n\to+\infty} \frac{\sigma^2}{n} = 0$$

e questo significa che la media campionaria è uno **stimatore consistente**.

2.2 La varianza campionaria

Si chiama **varianza campionaria** di un campione casuale X_1, X_2, ..., X_n la quantità S_n^2 definita dalla relazione:

$$S_n^2 = \frac{(X_1 - \overline{X}_n)^2 + (X_2 - \overline{X}_n)^2 + ... + (X_n - \overline{X}_n)^2}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Anche la varianza campionaria dipende dal campione, quindi S_n^2 **è una variabile aleatoria**.

Si dimostra che, nel caso di un campionamento bernoulliano:

data una popolazione X di media μ e varianza σ^2, il valore atteso della distribuzione della variabile aleatoria varianza campionaria è dato dalla relazione

$$E(S_n^2) = \frac{n-1}{n} \cdot \sigma^2$$

La varianza campionaria **non è quindi uno stimatore corretto** della varianza della popolazione.
Per poter lavorare con uno stimatore corretto dobbiamo apportare una correzione a questa statistica; definiamo allora la **varianza corretta** nel seguente modo:

varianza corretta $\widehat{S}_n^2 = \frac{n}{n-1} \cdot S_n^2$ cioè $\widehat{S}_n^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2}{n-1}$

LA VARIANZA CORRETTA

In questo modo risulta: $E(S_n^2) = \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \sigma^2 = \sigma^2$.

La varianza campionaria corretta è uno stimatore non distorto della varianza della popolazione.

Riprendiamo l'esempio del paragrafo precedente in cui la popolazione è il peso del raccolto di pomodori in cinque appezzamenti di terreno e verifichiamo la relazione del teorema.

Campioni	65; 65	72; 65	70; 65	68; 65	55; 65
	65; 72	72; 72	70; 72	68; 72	55; 72
	65; 70	72; 70	70; 70	68; 70	55; 70
	65; 68	72; 68	70; 68	68; 68	55; 68
	65; 55	72; 55	70; 55	68; 55	55; 55
Varianze	0	12,25	6,25	2,25	25
	12,25	0	1	4	72,25
	6,25	1	0	1	56,25
	2,25	4	1	0	42,25
	25	72,25	56,25	42,25	0

La distribuzione della varianza campionaria è dunque la seguente:

S_2^2	0	1	2,25	4	6,25	12,25	25	42,25	56,25	72,25
Frequenza	5	4	2	2	2	2	2	2	2	2
Probabilità	0,2	0,16	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08

Calcoliamo il valore atteso: $E(S_2^2) = 17,8$

Verifichiamo la relazione del teorema: $\frac{n-1}{n} \cdot \sigma^2 = \frac{1}{2} \cdot 35,6 = 17,8$

2.3 La proporzione campionaria

Un'indagine statistica può riguardare il possesso o meno di un certo attributo; per esempio, "possedere una carta di credito" oppure "avere una casa di proprietà" in una popolazione di individui, "essere difettoso" nella popolazione rappresentata dai pezzi prodotti da un'azienda.
Anche in questi casi si analizza un campione della popolazione calcolando la frequenza (proporzione) con cui il carattere compare e poi si cerca di inferire i risultati ottenuti alla popolazione. Uno stimatore della frequenza di un carattere in una popolazione è la frequenza con cui lo stesso carattere compare nel campione. Data dunque una popolazione di ampiezza N, se R è il numero di elementi che possiede un certo carattere, la frequenza p relativa del carattere è il rapporto

$$p = \frac{R}{N}$$

Secondo la definizione classica di probabilità, tale valore rappresenta anche la probabilità che ha il carattere di manifestarsi in un elemento della popolazione. Poiché di solito p non è noto, vogliamo vedere come sia possibile determinare una sua stima dal corrispondente parametro del campione.
Se il campione ha ampiezza n e k è il numero degli elementi che possiede il carattere oggetto di studio, allora

$$f = \frac{k}{n}$$

è una stima del parametro p della popolazione.
Il numero k non è costante, ma varia a seconda del campione ed è quindi una variabile aleatoria che indichiamo con K; tale variabile segue una legge di distribuzione binomiale in cui la probabilità di osservare un particolare valore è:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Il valor medio e la varianza di K sono quindi $E(K) = np$ e $V(K) = np(1-p)$.
Anche il rapporto

$$F = \frac{K}{n}$$

è, di conseguenza, una variabile aleatoria che prende il nome di **proporzione** o **frequenza campionaria**.
La frequenza campionaria segue anch'essa una distribuzione binomiale in cui:

- $E(F) = E\left(\frac{K}{n}\right) = \frac{E(K)}{n} = \frac{np}{n} = p$ e questo significa che

  la frequenza campionaria è uno stimatore corretto della frequenza relativa della popolazione.

- $V(F) = V\left(\frac{K}{n}\right) = \frac{1}{n^2} V(K) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$

Da quest'ultima relazione possiamo dedurre che la dispersione dei dati attorno alla media può essere resa via via più piccola aumentando l'ampiezza del campione; la **frequenza campionaria è dunque uno stimatore consistente** della frequenza della popolazione.

VERIFICA DI COMPRENSIONE

1. Uno stimatore è corretto se:
- **a.** il suo valore atteso e la sua varianza sono uguali ai rispettivi parametri della popolazione
- **b.** il suo valore atteso è uguale al corrispondente parametro della popolazione
- **c.** il suo valore atteso diviso per l'ampiezza del campione è uguale al corrispondente parametro della popolazione.

2. Degli stimatori media campionaria, varianza campionaria e frequenza campionaria si può dire che:
- **a.** sono tutti stimatori corretti
- **b.** solo la media campionaria è uno stimatore corretto e consistente
- **c.** solo la varianza campionaria non è corretta.

3. IL CASO DELLA DISTRIBUZIONE NORMALE

Gli esercizi di questo paragrafo sono a pag. 37

In molte situazioni, anche se non se ne conoscono i parametri, si sa che una popolazione, rispetto ad un certo carattere, ha una distribuzione normale. In questi casi accade che le medie dei campioni di ampiezza n estratti da tale popolazione si distribuiscono normalmente con la stessa media e varianza pari a $\frac{1}{n}$ di quella della popolazione.

Si dimostra infatti che vale il seguente teorema.

Teorema. Se una popolazione X è distribuita normalmente con media μ e varianza σ^2, anche la variabile aleatoria media campionaria dei campioni di ampiezza n è distribuita normalmente con la stessa media μ e con varianza $\frac{\sigma^2}{n}$.

Inoltre si può dimostrare che:

Teorema (del limite centrale). In qualunque modo sia distribuita una popolazione, purché la media e la varianza siano due valori finiti μ e σ^2, al crescere di n la media campionaria si distribuisce normalmente con media μ e varianza $\frac{\sigma^2}{n}$.

In altri termini, questo teorema afferma che la funzione di distribuzione della variabile aleatoria

$$\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}}}$$

dove $\sigma_{\overline{X}}$ rappresenta lo scarto quadratico medio della media campionaria, approssima la distribuzione normale standardizzata (media 0 e varianza 1).

Con questi due teoremi abbiamo affermato che se la distribuzione di una popolazione è normale, anche la distribuzione della media campionaria lo è; ma anche se la distribuzione della popolazione non è normale, la distribuzione della media campionaria lo diventa a patto di prendere n abbastanza grande, in genere basta prendere $n \geq 30$.

ESEMPI

1. Si sa che la produzione di grano per ettaro in una certa località si distribuisce normalmente con media $\mu = 25$q e scarto quadratico medio $\sigma = 4$q. Vogliamo determinare la probabilità che, estraendo bernoullianamente un campione di 10 campi, la produzione media superi i 27q.

Per un campione bernoulliano abbiamo visto che valgono le relazioni $E(\overline{X}_{10}) = \mu = 25$ e $V(\overline{X}_{10}) = \frac{\sigma^2}{10} = \frac{16}{10} = 1{,}6$; lo scarto quadratico medio del campione è quindi $\sqrt{1{,}6} \approx 1{,}2649$.

Per calcolare la probabilità richiesta dobbiamo passare ad una variabile z distribuita normalmente con media 0 e varianza 1 (normale standardizzata), la cui espressione è

$$z = \frac{\overline{X}_{10} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X}_{10} - 25}{1{,}2649}$$

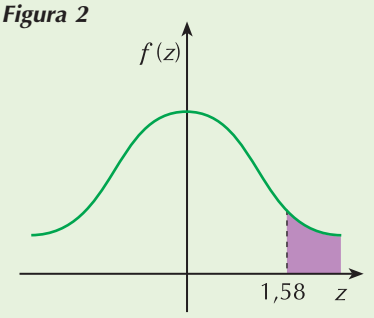
Nel nostro caso, poiché stiamo ricercando la probabilità che il valor medio sia superiore a 27, la variabile z ha valore

$$z = \frac{27 - 25}{1{,}2649} \approx 1{,}58$$

Servendoci delle tavole relative alla gaussiana standardizzata (le trovi online), abbiamo che (***figura 2***)

$$p(z > 1{,}58) = 1 - p(z < 1{,}58) = 1 - 0{,}9429 = 0{,}0571$$

Figura 2



VERIFICA DI COMPRENSIONE

1. In base al teorema del limite centrale si può affermare che la distribuzione della media campionaria:

a. è approssimativamente normale con media μ e varianza $\frac{\sigma^2}{n}$ per n sufficientemente grande solo se la popolazione ha distribuzione normale

b. è approssimativamente normale con media μ e varianza $\frac{\sigma^2}{n}$ per n sufficientemente grande in qualunque modo si distribuisca la popolazione

c. è approssimativamente normale con media μ e varianza σ^2 per n sufficientemente grande in qualunque modo si distribuisca la popolazione

d. è approssimativamente normale con media μ e varianza σ^2 per n sufficientemente grande solo se la popolazione ha distribuzione normale.

4. STIMA PUNTUALE DEI PARAMETRI

Gli esercizi di questo paragrafo sono a pag. 38

Supponiamo di sapere che una popolazione X ha una certa distribuzione di cui conosciamo la forma, ma di cui non conosciamo i valori dei parametri; per esempio potremmo sapere che la distribuzione è normale ma non conoscere μ e σ, oppure binomiale ma non conoscere p. Una stima di questi parametri, come abbiamo più volte detto, può essere fatta tramite un campione e di solito si parla di:

- **stima puntuale** se si vuole determinare un valore per i parametri incogniti dando anche una misura della precisione con cui tale stima viene fatta
- **stima per intervallo** se si vuole determinare un intervallo che, verosimilmente, contiene il valore vero del parametro.

4.1 Stima puntuale della media

Abbiamo visto nei precedenti paragrafi che la media campionaria è uno stimatore corretto e consistente della media della popolazione e si può dimostrare che, rispetto ad altri eventuali stimatori, è anche più efficiente.
L'**errore medio di campionamento**, cioè l'errore che si compie valutando la media della popolazione tramite la media del campione, è rappresentato dallo scarto quadratico medio della media campionaria, cioè, nel caso di campionamento bernoulliano:

$$\text{errore} = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Il problema è che, di solito, lo scarto quadratico medio della popolazione non è noto e si deve ricorrere ad una sua stima valutata sul campione; ricordiamo che una stima corretta della varianza della popolazione è la varianza corretta $\widehat{S}_n^2 = \frac{n}{n-1} \cdot S_n^2$.

Una stima dell'errore medio di campionamento è quindi data da:

$$S_{\overline{x}} = \frac{\widehat{S}_n}{\sqrt{n}}$$

Vediamo alcuni esempi.

ESEMPI

1. Dai 140 studenti che frequentano la classe seconda di un Istituto Tecnico è stato estratto bernoullianamente un campione casuale di 50 elementi ed è stato posto loro un quesito sul tempo, in ore, che ogni giorno dedicano allo studio. I risultati dell'indagine sono riportati nella seguente tabella

Ore	1	2	3	4	5
N. alunni	4	11	22	9	4

Vogliamo stimare il tempo medio che gli studenti di quella scuola dedicano allo studio valutando anche l'errore commesso.

Calcoliamo il valor medio di questo campione $\overline{x} = \frac{1 \cdot 4 + 2 \cdot 11 + 3 \cdot 22 + 4 \cdot 9 + 5 \cdot 4}{50} = 2{,}96$ ore.

Per quanto abbiamo visto, tale media può essere considerata una stima della media della popolazione. Poiché non conosciamo la varianza della popolazione, possiamo stimare l'errore calcolando la varianza corretta del campione e quindi lo scarto quadratico medio:

Varianza del campione:

x_i	f_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot f_i$
1	4	−1,96	3,8416	15,3664
2	11	−0,96	0,9216	10,1376
3	22	0,04	0,0016	0,0352
4	9	1,04	1,0816	9,7344
5	4	2,04	4,1616	16,6464
			$\sum$	**51,92**

$s^2 = \frac{51{,}92}{50} = 1{,}0384$

Calcoliamo ora la varianza corretta che è data da $\hat{s}^2 = \frac{n}{n-1} \cdot s^2 = \frac{50}{49} \cdot 1{,}0384 \approx 1{,}06$

oppure dalla relazione $\hat{s}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot f_i}{n-1} = \frac{51{,}92}{49} \approx 1{,}06$

L'errore medio di campionamento è quindi $s_{\overline{x}} = \frac{\hat{s}}{\sqrt{n}} = \sqrt{\frac{1{,}06}{50}} \approx 0{,}146$

Riassumendo, possiamo dire che gli studenti di quella scuola dedicano allo studio pomeridiano 2,96 ore con un errore medio di circa 0,146 ore, cioè 8,76 minuti.

2. Si è sottoposto un gruppo di 50 bambini che frequentano la stessa classe di una scuola elementare ad un test sul tempo di attenzione ininterrotta. I risultati si possono dedurre dalla seguente tabella:

minuti di attenzione	0 - 5	5 - 8	8 - 10	10 - 12	12 - 16	16 - 20
frequenza	1	3	6	19	17	4

Da tale campione si vuole fare una stima media del tempo di attenzione di tutta la popolazione dei 326 alunni che frequentano quella classe nell'ambito del distretto scolastico e stimare inoltre l'errore di campionamento.

I dati sono raggruppati per classi, associamo quindi ad ogni classe il valore centrale e procediamo completando la seguente tabella:

valore centrale x_i	frequenza f_i	$x_i \cdot f_i$	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot f_i$
2,5	1	2,5	−9,4	88,36	88,36
6,5	3	19,5	−5,4	29,16	87,48
9	6	54	−2,9	8,41	50,46
11	19	209	−0,9	0,81	15,39
14	17	238	2,1	4,41	74,97
18	4	72	6,1	37,21	148,84
	$\sum$	**595**			**465,5**

Tenendo presente che $N = 326$ e $n = 50$:

- calcoliamo la media ponderata $\overline{x} = \frac{595}{50} = 11{,}9$
- la varianza corretta del campione è $\hat{s}^2 = \frac{465{,}5}{49} = 9{,}5$
- l'errore medio di campionamento è $s_{\overline{x}} = \frac{\sqrt{9{,}5}}{\sqrt{50}} = 0{,}4359$ cioè circa 26 secondi

Possiamo concludere dicendo che, mediamente, il livello di attenzione senza interruzione per la popolazione presa in esame è di circa 12 minuti con un errore di 26 secondi.

3. Esamina con attenzione il seguente esempio in cui affronteremo un problema di **stima puntuale di un totale**.
In un distretto scolastico, in un certo anno, si sono costituite 184 classi di prima media; da una indagine su un campione di 30 classi è risultato che il numero medio di alunni per classe è 23 con uno scarto quadratico medio di 5 alunni. Si vuole fare una stima del numero degli alunni che compongono la popolazione.

I dati del problema ci dicono che, relativamente al campione analizzato, $\overline{x} = 23$ e $s = 5$.

Prendiamo come stimatore del totale della popolazione la funzione $T_{30} = N \cdot \overline{X}_{30}$ dove N è il numero delle classi che costituiscono la popolazione; nel nostro caso, un valore di T_{30}, che è anche una stima del totale della popolazione, è

$$t_{30} = 184 \cdot 23 = 4232$$

Possiamo dire che il numero totale degli alunni è stimato in 4232 unità.

Per trovare l'errore di campionamento operiamo in questo modo:

- ricaviamo la varianza del campione $s^2 = (5)^2 = 25$
- la varianza corretta è quindi $\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{30}{29} \cdot 25 \approx 25{,}86$
- da cui l'errore di campionamento è $s_{\overline{x}} = \frac{\hat{s}}{\sqrt{n}} = \frac{\sqrt{25{,}86}}{\sqrt{30}} \approx 0{,}93$

Per trovare la misura dell'errore relativo a tutta la popolazione, moltiplichiamo per N i valori trovati ottenendo:

$$N \cdot s_{\overline{x}} = 184 \cdot 0{,}93 \approx 171$$

Possiamo allora concludere dicendo che la stima per il totale degli alunni è di 4232 unità con un errore di 171 alunni.

Generalizzando l'ultimo problema visto negli esempi, per risolvere un **problema di stima puntuale di un totale** si assume come stimatore la funzione $T_n = N \cdot \overline{X}_n$ che risulta essere uno stimatore corretto del totale della popolazione; infatti

$$E(T_n) = E(N \cdot \overline{X}_n) = N \cdot E(\overline{X}_n) = N \cdot \mu$$

Una stima del totale della popolazione è il valore $N \cdot \overline{x}$.
L'errore medio di campionamento si ottiene moltiplicando per N lo scarto quadratico medio del campione:

$$N \cdot s_{\overline{x}} = N \cdot \frac{\hat{s}}{\sqrt{n}}$$

4.2 Stima puntuale della frequenza

Sappiamo che la frequenza campionaria $F = \frac{K}{n}$ è uno stimatore corretto della frequenza p del carattere oggetto di studio valutato sull'intera popolazione.

Una stima di p è quindi data dal rapporto

$$f = \frac{k}{n}$$

dove k è il numero di elementi del campione che possiede il carattere.
Per valutare l'**errore medio di campionamento** si considera la popolazione X come una variabile aleatoria avente distribuzione binomiale con valor medio $E(X) = np$ e varianza $V(X) = np(1 - p)$.
Il valore medio e la varianza della frequenza campionaria sono (lo abbiamo visto in uno dei precedenti paragrafi)

$$E(F) = np \qquad V(F) = \frac{p(1-p)}{n}$$

Poiché il parametro p non è noto, possiamo sostituire ad esso la frequenza f del campione (purché il campione sia sufficientemente numeroso, almeno $n > 30$) e stimare l'errore medio di campionamento con la relazione

$$s_f = \sqrt{\frac{f(1-f)}{n}}$$

Vediamo un esempio.

ESEMPI

1. Si vuole stimare la percentuale delle donne che, in un data regione ed in un fissato intervallo di tempo, hanno avuto, nell'arco della loro vita, almeno un parto gemellare. Da una analisi effettuata su di un campione casuale di 500 donne è risultato che 32 possiedono questo attributo.
Determiniamo una stima della percentuale con cui il carattere si presenta nella popolazione e diamo una stima dell'errore di campionamento.

Abbiamo visto che una stima del dato percentuale relativo alla presenza del carattere nella popolazione è il rapporto

$$f = \frac{32}{500} = 0{,}064$$

Stimiamo l'errore medio di campionamento:

$$s_f = \sqrt{\frac{f(1-f)}{n}} = \sqrt{\frac{0{,}064(1-0{,}064)}{500}} = 0{,}01095$$

Possiamo concludere che il 6,4% delle donne di quella regione ha avuto almeno un parto gemellare con un errore medio di campionamento dell'1,095%.

VERIFICA DI COMPRENSIONE

1. La durata in ore di cinque componenti elettrici estratti da un certo lotto è di 526, 532, 520, 525, 534. Una stima della durata media m in ore dei componenti con il relativo errore e è:

a. $m = 524{,}5 \quad e = 1{,}54$
b. $m = 520{,}2 \quad e = 0{,}35$
c. $m = 527{,}4 \quad e = 0{,}26$
d. $m = 527{,}4 \quad e = 2{,}52$

5. STIME PER INTERVALLO: EFFICACIA DI UN PRODOTTO O DI UN SERVIZIO

Gli esercizi di questo paragrafo sono a pag. 42

Nelle produzioni industriali è spesso più significativo, rispetto ad avere una stima puntuale, conoscere un intervallo entro cui cade il valore vero del parametro con una certa probabilità; questo permette infatti di **verificare l'efficacia** di un processo produttivo e di esercitare un **controllo sulla produzione**.
Per esempio un'azienda alimentare può ritenere che i macchinari usati funzionino in modo corretto se il peso medio del mais contenuto nelle scatolette che produce cade nell'intervallo (496; 505) grammi con una probabilità del 95%.
Un problema di stima per intervallo si può sintetizzare in questo modo:

trovare un intervallo (z_1, z_2) entro il quale si trova il parametro incognito ϑ di una popolazione con una probabilità fissata a priori.

L'intervallo (z_1, z_2) viene detto **intervallo di fiducia** o anche **intervallo di confidenza**.
La probabilità che ϑ appartenga all'intervallo (z_1, z_2) viene detta **livello di fiducia** e si indica con il simbolo $1 - \alpha$.
Il margine di errore che si è disposti ad accettare nel compiere la stima viene detto **livello di rischio** ed è pari ad α.
In questo paragrafo ci occupiamo di determinare una stima per intervallo della media di una popolazione e della frequenza, supponendo che il campione estratto sia un grande campione.

Affinché un campione di ampiezza n possa essere considerato un grande campione è sufficiente che sia $n > 30$.
In questo caso, per il teorema del limite centrale, la media campionaria $\overline{X}$ ha distribuzione normale comunque sia distribuita la popolazione.

5.1 Stima per intervallo della media

Stima per intervallo della media di un campione

In una produzione aziendale si eseguono spesso controlli di qualità sulla produzione; in questi casi sono di solito note le caratteristiche degli impianti ed è quindi possibile risalire alla media e alla varianza della popolazione dei pezzi prodotti e si vuole determinare in quale intervallo deve essere compresa la media di un campione per considerare accettabile la produzione ad un certo livello di confidenza.

Un problema di questo tipo si può sintetizzare in questo modo:

determinare l'intervallo di confidenza a livello $1 - \alpha$ in cui si può trovare la media di un campione di ampiezza n, supposto che siano note la media μ e la varianza σ^2 della popolazione.

Se il campione, come abbiamo supposto, è grande, la variabile aleatoria media campionaria ha distribuzione normale con media μ e varianza $\frac{\sigma^2}{n}$, cioè deviazione standard $\frac{\sigma}{\sqrt{n}}$.
La variabile aleatoria standardizzata

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

è allora distribuita normalmente con media 0 e deviazione standard 1.

Il livello di fiducia $1-\alpha$ di appartenenza della variabile z ad un intervallo $(-a, a)$, simmetrico rispetto alla media 0, è dato da (segui la ***figura 3***)

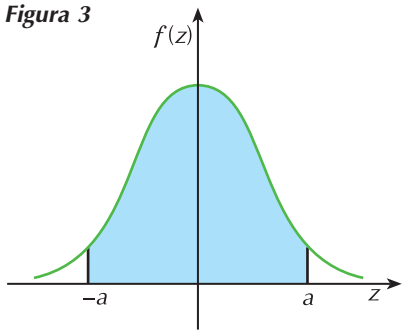
$$1-\alpha = p(-a < z < a) = 2p(0 < z < a) = 2[p(z < a) - p(z > 0)] =$$

$$= 2[p(z < a) - 0{,}5] = 2p(z < a) - 1$$

Con questa relazione possiamo calcolare il livello di fiducia, ma possiamo anche risalire al valore di a una volta noto quello di $1-\alpha$.
Ad esempio, fissato $1-\alpha = 0{,}95$, dobbiamo risolvere l'equazione

$$2p(z < a) - 1 = 0{,}95 \qquad \text{da cui ricaviamo che} \qquad p(z < a) = 0{,}975$$

Figura 3



Ricercando nelle tavole della gaussiana standardizzata il valore 0,975, troviamo per a il valore 1,96. La variabile z cade dunque nell'intervallo $(-1{,}96;\ 1{,}96)$ con una probabilità del 95%.

Le tavole della gaussiana sono disponibili a pagina 50.

I punti a sono dunque i valori che soddisfano l'equazione

$$p(z < a) = 1 - \frac{\alpha}{2}$$

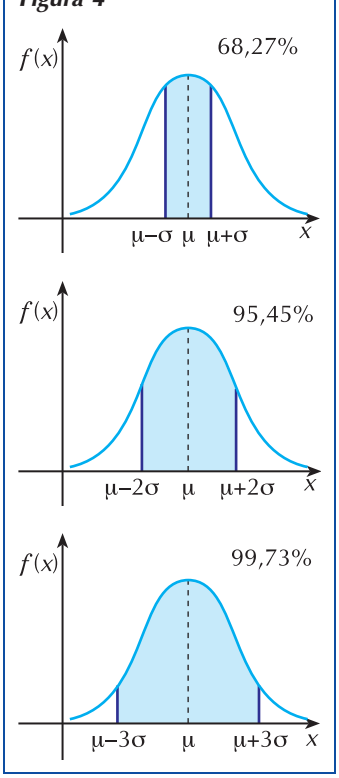
Essi si dicono **punti critici** e si indicano con il simbolo $z_{1-\frac{\alpha}{2}}$.

Alcuni punti critici che vengono spesso utilizzati sono i seguenti

livello fiducia	99,90%	99,73%	99%	98%	96%	95,44%	95%	90%	80%	68,26%
valore critico	3,29	3	2,58	2,33	2,05	2	1,96	1,645	1,28	1

I valori di probabilità in corrispondenza dei valori critici 1, 2, 3 sono quelli che indicano scostamenti dalla media di σ, 2σ, 3σ (***figura 4***).

Figura 4



Riprendendo l'esempio precedente, possiamo scrivere che

$$p(-1{,}96 < z < 1{,}96) = 0{,}95$$

Operando la sostituzione inversa, cioè risostituendo alla variabile z il valore $\dfrac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ si ha:

$$p\left(-1{,}96 < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < 1{,}96\right) = 0{,}95$$

cioè, operando in modo opportuno

$$p\left(\mu - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} < \overline{X}_n < \mu + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

L'intervallo di fiducia al livello del 95% è dunque l'insieme

$$\left(\mu - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}};\ \mu + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

I risultati ottenuti significano che il 95% dei campioni estratti ha un valor medio che cade nell'intervallo di confidenza, il 5% ha un valor medio che cade al di

fuori e precisamente il 2,5% nella zona esterna di sinistra ed il 2,5% nella zona esterna di destra (***figura 5***).

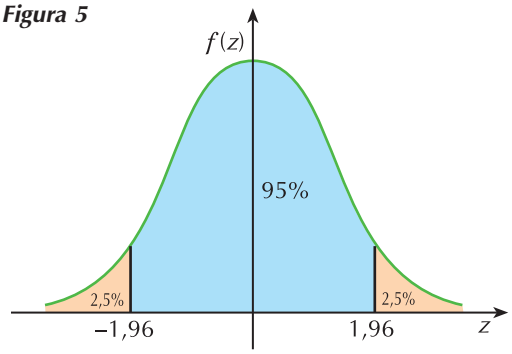
In generale, se $z_{1-\frac{\alpha}{2}}$ rappresenta il punto critico corrispondente al livello di fiducia $1-\alpha$, si ha che

$$p\left(\mu - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \overline{X}_n < \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

e l'intervallo di confidenza assume quindi la forma

$$\left(\mu - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\,;\, \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Figura 5



ESEMPI

1. Un macchinario produce filtri per condizionatori dello spessore medio di 3cm, con uno scarto quadratico medio di 0,2cm. Estratto bernoullianamente un campione di 100 filtri, calcoliamo l'intervallo di fiducia in cui dovrebbe trovarsi il valor medio dello spessore del campione esaminato, con un livello di fiducia del 98%.

In questo caso sono noti sia la media della popolazione che lo scarto quadratico medio: $\mu = 3$ e $\sigma = 0{,}2$.

Il livello $1-\alpha$ di fiducia è 0,98 a cui corrisponde il valore critico $z_{1-\frac{\alpha}{2}} = 2{,}33$; l'intervallo di fiducia in cui si troverà il valor medio del campione con una probabilità del 98% è quindi

$$\left(3 - 2{,}33 \cdot \frac{0{,}2}{\sqrt{100}}\,;\, 3 + 2{,}33 \cdot \frac{0{,}2}{\sqrt{100}}\right) \quad \text{cioè} \quad (2{,}9534;\, 3{,}0466)$$

I risultati ottenuti possono essere interpretati dalla direzione dell'azienda in questo modo: se la media del campione estratto non rientra in questo intervallo, il lotto di produzione non è presumibilmente affidabile ed occorre procedere ad una analisi del processo produttivo per individuare eventuali criticità.

Stima per intervallo della media della popolazione

A differenza del caso precedente, in situazioni di questo tipo non si conosce la media della popolazione e si vuole trovare una sua stima a partire dalla media del campione.
Ricordiamo che il teorema del limite centrale afferma che, qualunque sia la distribuzione della popolazione, purché abbia media μ e varianza σ^2, le medie dei campioni, al crescere dell'ampiezza n, tendono ad una distribuzione normale con media μ e varianza $\frac{\sigma^2}{n}$.

D'altra parte, se la popolazione ha distribuzione normale, anche la distribuzione della media campionaria è normale qualunque sia il valore di n.
Nella nostra analisi ci occuperemo del caso di grandi campioni con una distribuzione della popolazione qualsiasi e del caso di piccoli campioni in cui si sa però che la distribuzione della popolazione è normale o quasi.
Quello che differenzierà i nostri studi sarà la conoscenza o meno della varianza σ^2 della popolazione.

I caso: è noto il valore di σ^2

Supponiamo di aver estratto bernoullianamente dalla popolazione un campio-

ne di ampiezza n di cui possiamo calcolare il valor medio $\overline{x}$ e la varianza s^2; a partire da queste informazioni vogliamo stimare la media μ della popolazione, supponendo nota la varianza σ^2. Riprendiamo allora la relazione del precedente paragrafo

$$p\left(-z_{1-\frac{\alpha}{2}} < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

che possiamo scrivere in questo modo $p\left(-z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \overline{X}_n - \mu < z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$

Dalla disuguaglianza di sinistra si ricava che $\mu < z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} + \overline{X}_n$

Dalla disuguaglianza di destra si ricava che $\mu > -z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} + \overline{X}_n$

In definitiva, possiamo quindi scrivere che:

$$\boldsymbol{p\left(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha} \qquad \textbf{(A)}$$

Questo significa che l'intervallo di fiducia del parametro μ, con livello di fiducia $1 - \alpha$, è

$$\boldsymbol{\left(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\,;\, \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)} \qquad \textbf{(B)}$$

Gli estremi di questo intervallo sono simmetrici rispetto a μ e, dato che σ e n sono noti, possono essere calcolati a partire dai dati del campione una volta che si sia fissato il livello di fiducia.
Ad esempio, la scrittura

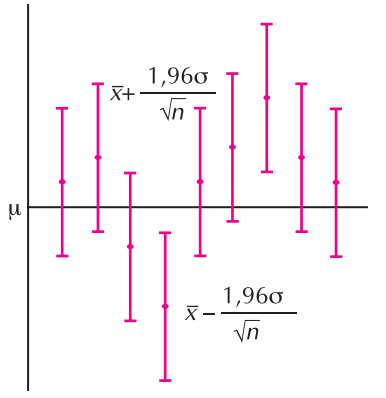
$$p\left(\overline{X}_n - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

significa che, nel 95% dei campioni bernoulliani, la media μ appartiene all'intervallo

$$\left(\overline{X}_n - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\,;\, \overline{X}_n + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Attenzione! Quello che abbiamo detto è che l'affermazione "μ appartiene all'intervallo $\left(\overline{X}_n - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\,;\, \overline{X}_n + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right)$" è vera al 95%, cioè 95 volte su 100 in una lunga sequenza di prove. Ciò non equivale a dire che con una probabilità del 95% la media μ cade in questo intervallo, perché μ non è una variabile: in ogni caso particolare μ giace all'interno oppure all'esterno dell'intervallo.
Bisogna pertanto immaginare una serie di ripetuti campionamenti casuali da una popolazione con un dato valore di μ (***figura 6***); nella lunga sequenza, il 95% degli intervalli di confidenza conterrà μ ed in quei casi l'affermazione risulterà vera. Cinque volte su cento $\overline{x}$ disterà da μ più di

Figura 6



1,96 volte l'errore standard (come nel quarto e nel settimo campione della figura) e l'intervallo di confidenza non conterrà μ: l'affermazione risulterà in quei casi falsa.
Se, in un particolare problema, si deve calcolare un intervallo di confidenza, si può essere tanto sfortunati da imbattersi in uno di quei cinque casi in cui l'affermazione è falsa; tuttavia si ha una probabilità del 95% di estrarre un campione in cui essa è vera.

ESEMPI

1. Da un lotto di prosciutti messi a stagionare se ne estrae un campione bernoulliano formato da 80 pezzi e se ne controlla il peso. Il valore medio del peso del campione è di 2,4kg. Da valutazioni precedenti si sa che la varianza di tutti i prosciutti messi a stagionare è $0{,}9\text{kg}^2$. Vogliamo determinare l'intervallo di fiducia per la stima della media di tutta la popolazione, con un livello di fiducia del

a. 95%; **b.** 96%; **c.** 99%.

Il campione è sufficientemente numeroso perché $n = 80$, quindi possiamo supporre che la distribuzione della media campionaria sia normale. Conosciamo il peso medio del campione, $\overline{x} = 2{,}4\text{kg}$ e la varianza della popolazione, $\sigma^2 = 0{,}9\text{kg}^2$.
Il valore critico che corrisponde ai vari livelli di fiducia è

a. $1 - \alpha = 0{,}95$ $\qquad z_{1-\frac{\alpha}{2}} = 1{,}96$

b. $1 - \alpha = 0{,}96$ $\qquad z_{1-\frac{\alpha}{2}} = 2{,}05$

c. $1 - \alpha = 0{,}99$ $\qquad z_{1-\frac{\alpha}{2}} = 2{,}58$

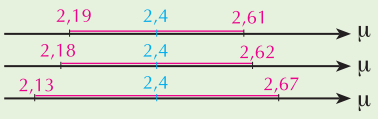
Calcoliamo l'intervallo di confidenza nei tre casi

a. $p\left(2{,}4 - 1{,}96 \cdot \sqrt{\frac{0{,}9}{80}} < \mu < 2{,}4 + 1{,}96 \cdot \sqrt{\frac{0{,}9}{80}}\right) = 0{,}95$ vale a dire che l'intervallo di confidenza per quel campione è $(2{,}19;\ 2{,}61)$.
Questo significa che l'affermazione "μ cade nell'intervallo $(2{,}19;\ 2{,}61)$", oppure "$\mu = 2{,}4 \pm 0{,}21$" è vera con una probabilità del 95%.

b. $p\left(2{,}4 - 2{,}05 \cdot \sqrt{\frac{0{,}9}{80}} < \mu < 2{,}4 + 2{,}05 \cdot \sqrt{\frac{0{,}9}{80}}\right) = 0{,}96$ vale a dire che l'intervallo di confidenza per quel campione è $(2{,}18;\ 2{,}62)$. L'affermazione "$\mu = 2{,}4 \pm 0{,}22$" è vera con una probabilità del 96%.

c. $p\left(2{,}4 - 2{,}58 \cdot \sqrt{\frac{0{,}9}{80}} < \mu < 2{,}4 + 2{,}58 \cdot \sqrt{\frac{0{,}9}{80}}\right) = 0{,}99$ vale a dire che l'intervallo di confidenza per quel campione è $(2{,}13;\ 2{,}67)$. L'affermazione "$\mu = 2{,}4 \pm 0{,}27$" è vera con una probabilità del 99%.

Figura 7



In ***figura 7*** puoi vedere gli intervalli di confidenza al variare del livello di fiducia.

Confrontando i tre casi analizzati nell'esempio precedente (puoi riferirti ancora alla ***figura 7***), vediamo che, all'aumentare del grado di fiducia, aumenta anche l'ampiezza dell'intervallo di confidenza; questo significa avere una stima intervallare meno precisa della media μ.

Se vogliamo avere una stima più precisa, l'unica possibilità che abbiamo è dunque quella di restringere l'intervallo di fiducia, cioè diminuirne l'ampiezza che è pari a $2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

Osserviamo però che, per diminuire tale ampiezza, non possiamo agire su σ che è fisso; dobbiamo perciò aumentare il valore di n, vale a dire scegliere campioni più grandi.

Ad esempio, se nel caso **a.** dell'esempio precedente avessimo preso un campione di 200 prosciutti (supponiamo con lo stesso valor medio), a un livello di fiducia del 95% avremmo trovato che $\mu = 2{,}4 \pm 0{,}13$, quindi un intervallo di confidenza più piccolo del precedente.

Il caso: il valore di σ^2 non è noto

Questo caso capita di frequente nelle indagini statistiche, perché 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.

Nel caso di grandi campioni ($n > 30$), si ha così che

$$p\left(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{S}}{\sqrt{n}} < \mu < \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{S}}{\sqrt{n}}\right) = 1 - \alpha$$

L'intervallo di fiducia del parametro μ, con livello di fiducia $1 - \alpha$, è quindi

$$\left(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{S}}{\sqrt{n}}; \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{S}}{\sqrt{n}}\right) \quad \text{cioè} \quad \left(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}; \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}\right)$$

Nel caso di piccoli campioni, non possiamo stimare la varianza della popolazione con la varianza corretta.

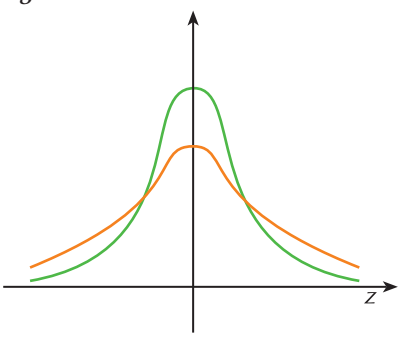
In pratica allora, anziché considerare la variabile z definita all'inizio di questo paragrafo, si considera la variabile aleatoria

$$t = \frac{\overline{X}_n - \mu}{\frac{\hat{S}}{\sqrt{n}}} \quad \text{o anche} \quad t = \frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n-1}}}$$

in cui si è sostituito σ con la stima campionaria della deviazione standard. Tale variabile prende il nome di ***t* di Student** dallo statistico W.S. Gosset, che usava lo pseudonimo Student per pubblicare i suoi lavori.

La variabile t segue una distribuzione non molto diversa da quella normale standardizzata a condizione che n sia sufficientemente grande perché, in tal caso, la deviazione standard del campione s è una buona approssimazione di σ; quando n è piccolo, s può invece differire considerevolmente da σ. Tutto questo fa sì che t abbia una variabilità casuale più grande di quella di z, da cui la maggior dispersione di questa distribuzione; in ***figura 8*** puoi vedere il confronto fra la gaussiana standardizzata (in verde) e la t di Student (in arancio).

Figura 8



Si rende necessario a questo punto fare una precisazione. Quando si calcola lo scarto quadratico medio s del campione, i valori $x_i - \overline{x}$ sono n e sono, in generale, uno diverso dall'altro; essi hanno però una caratteristica: la loro somma è zero. Questo significa che $n-1$ di essi sono arbitrari, ma uno dipende dagli altri valori. Si esprime questo fatto dicendo che le differenze $x_i - \overline{x}$ hanno $n-1$ **gradi di libertà**.

La variabile aleatoria t ha quindi anch'essa $n-1$ gradi di libertà che si indicano di solito con il simbolo ν; al crescere di ν, per quanto abbiamo detto, la distribuzione t tende alla normale standardizzata; i suoi valori critici possono essere tabulati e, a pagina 51, troverai i valori di questa distribuzione per alcuni valori di ν e di $1-\alpha$.
Noterai che, per $\nu \to +\infty$, si ritrova il valore critico 1,96 in corrispondenza di un livello di confidenza del 95%.

Con considerazioni del tutto analoghe a quelle fatte nel caso precedente, possiamo allora scrivere che

$$p\left(\overline{X}_n - {}_\nu t_{1-\alpha} \cdot \frac{S}{\sqrt{n-1}} < \mu < \overline{X}_n + {}_\nu t_{1-\alpha} \cdot \frac{S}{\sqrt{n-1}}\right) = 1-\alpha$$

dove con il simbolo ${}_\nu t_{1-\alpha}$ abbiamo indicato i valori critici della distribuzione di Student. L'intervallo di fiducia del parametro μ al livello di fiducia $1-\alpha$ è quindi:

$$\left(\overline{X}_n - {}_\nu t_{1-\alpha} \cdot \frac{S}{\sqrt{n-1}};\quad \overline{X}_n + {}_\nu t_{1-\alpha} \cdot \frac{S}{\sqrt{n-1}}\right)$$

Quest'ultima relazione differisce da quella trovata nel caso di grandi campioni per la sostituzione dei valori critici della distribuzione normale con quelli della distribuzione t che, come si vede confrontando le due tavole, sono più alti.

LE REGOLE

Data una popolazione normale ed un campionamento di tipo bernoulliano:

- se è noto il valore di σ^2, l'intervallo di fiducia a livello $1-\alpha$ è

$$\left(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}};\ \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

- se non è noto il valore di σ^2, ma si ha a che fare con un grande campione $(n > 30)$, l'intervallo di fiducia a livello $1-\alpha$ è

$$\left(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}};\ \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n-1}}\right)$$

- se non è noto il valore di σ^2 ed il campione è piccolo $(n \leq 30)$, l'intervallo di fiducia a livello $1-\alpha$ è

$$\left(\overline{X}_n - {}_\nu t_{1-\alpha} \cdot \frac{S}{\sqrt{n-1}};\ \overline{X}_n + {}_\nu t_{1-\alpha} \cdot \frac{S}{\sqrt{n-1}}\right)$$

ESEMPI

1. Esaminando un campione bernoulliano di 15 travi da costruzione, si trova che la lunghezza media è di 2,52m con uno scarto quadratico medio di 0,3m. Stimiamo l'intervallo di fiducia al 90% nel caso in cui la distribuzione della popolazione sia normale.

Si tratta di un piccolo campione estratto da una popolazione normale; dobbiamo quindi servirci della distribuzione t. Il numero di gradi di libertà è 14 e $1-\alpha = 0{,}90$. Il valore critico che leggiamo sulla tavola di Student in corrispondenza di questi valori è ${}_{14}t_{0,90} = 1{,}761$; possiamo quindi dire che

$$p\left(2{,}52 - 1{,}761 \cdot \frac{0{,}3}{\sqrt{14}} < \mu < 2{,}52 + 1{,}761 \cdot \frac{0{,}3}{\sqrt{14}}\right) = 0{,}90$$

cioè che $\mu = 2{,}52 \pm 0{,}14$ o anche che $\mu \in (2{,}38;\ 2{,}66)$ al livello di fiducia del 90%.

5.2 Stima per intervallo di una frequenza

Consideriamo la distribuzione F della frequenza campionaria; se il campione è grande e se p non è prossimo a 0 oppure a 1, la distribuzione binomiale si può approssimare con quella normale e, nel caso di campionamento bernoulliano, si verifica che:

$$p\left(p - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} < f < p + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

Il problema che dobbiamo affrontare è però quello di trovare un intervallo di stima per p in funzione del valore di f, frequenza relativa del campione.

Dall'equazione $f = p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$, con una serie di passaggi algebrici che omettiamo per semplicità, si ricava che

$$p = f \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}}$$

Allora risulta che $p\left(f - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}} < p < f + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{f(1-f)}{n}}\right) = 1-\alpha$

E' possibile anche fare una **valutazione di prudenza** della frequenza della popolazione considerando lo scarto quadratico medio della popolazione

$$S = \sqrt{\frac{p(1-p)}{n}}$$

Sappiamo che la deviazione standard rappresenta l'errore medio di campionamento che, in questo caso, raggiunge il suo valore massimo per $p = 0{,}5$ (puoi trovare il massimo con i metodi dell'analisi). Il massimo errore di campionamento è quindi $\frac{1}{2\sqrt{n}}$ (basta sostituire 0,5 nell'espressione precedente).

Allora, per valori di p vicini a 0,5, si ha che $p = f \pm z_{1-\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}}$, valutazione che, essendo indipendente da f (f non compare nell'espressione dopo il doppio segno), è applicabile a qualsiasi popolazione.

Tralasciamo il caso dei piccoli campioni in cui la binomiale non può essere approssimata dalla gaussiana, lasciando la trattazione ad uno studio più approfondito.

ESEMPI

1. Su un campione bernoulliano di 250 alunni di sesso maschile di una scuola si è trovato che il 68% è solito fumare sigarette, mentre su un campione analogo di alunni di sesso femminile si è trovata una percentuale del 62% di fumatrici. Diamo una valutazione ad un livello di fiducia del 95% dell'intervallo di fiducia rispettivamente delle percentuali dei fumatori e delle fumatrici.

Per i maschi abbiamo una frequenza data da $f_1 = 0{,}68$ e $z_{1-\frac{\alpha}{2}} = 1{,}96$ quindi

$$p\left(0{,}68 - 1{,}96 \cdot \sqrt{\frac{0{,}68(1-0{,}68)}{250}} < p_1 < 0{,}68 + 1{,}96 \cdot \sqrt{\frac{0{,}68(1-0{,}68)}{250}}\right) = 0{,}95 \quad \text{cioè} \quad 0{,}62 < p_1 < 0{,}74$$

Per le femmine abbiamo una frequenza data da $f_2 = 0{,}62$ e $z_{1-\frac{\alpha}{2}} = 1{,}96$ quindi

$$p\left(0{,}62 - 1{,}96 \cdot \sqrt{\frac{0{,}62(1-0{,}62)}{250}} < p_2 < 0{,}62 + 1{,}96 \cdot \sqrt{\frac{0{,}62(1-0{,}62)}{250}}\right) = 0{,}95 \quad \text{cioè} \quad 0{,}56 < p_2 < 0{,}68$$

VERIFICA DI COMPRENSIONE

1. In un processo industriale vengono riempiti vasetti di marmellata; si sa che i pesi dei vasetti si distribuiscono normalmente con media μ e deviazione standard $15g$. Un campione casuale di 25 confezioni ha dato un peso medio di $362,3g$. Un intervallo di confidenza con livello di fiducia del 95% per la media μ è:

a. (348,2; 363,8) **b.** (356,4; 368,2) **c.** (354,1; 365,9) **d.** (358,3; 364,6)

2. Dall'esame di un campione di 100 famiglie si è osservato che 24 hanno usato l'aereo per le loro ultime vacanze. Una stima della percentuale di famiglie che hanno usato l'aereo ad un livello di fiducia del 99,73% è l'intervallo:

a. (11%; 37%) **b.** (12%; 25%) **c.** (15%; 35%) **d.** (20%; 28%)

6. LA VERIFICA DELLE IPOTESI

Gli esercizi di questo paragrafo sono a pag. 45

6.1 Le ipotesi statistiche: controllo dell'efficacia di un prodotto o servizio

Le indagini statistiche si fanno per avere informazioni dirette su un fenomeno in modo da poter prendere decisioni motivate.
Per esempio, un laboratorio che pensa di aver selezionato un tipo di erba da giardino che non cresce più di 3cm al mese, può fare dei controlli seminando quel tipo di erba in un certo numero di terreni campione e osservando la crescita; uno zuccherificio che vuole monitorare costantemente la produzione, effettua dei controlli a campione sulle confezioni in modo da controllare il peso.

In entrambi gli esempi, la situazione può essere sintetizzata con un'affermazione

"l'erba non cresce più di 3cm al mese" "il peso delle confezioni è di 1kg"

e si deve stabilire, con un prefissato margine di errore, se quanto detto corrisponde al vero.
In situazioni di questo tipo le affermazioni di cui si vuole stabilire il valore di verità sono relative a un parametro della distribuzione di una variabile aleatoria e per stabilire se sono vere o meno ci si affida a un campione. Relativamente ai due esempi:

- variabile aleatoria: crescita dell'erba
  parametro da testare: valore medio μ uguale a 3cm
- variabile aleatoria: peso di una confezione
  parametro da testare: valore medio μ uguale a 1kg

Si dice **ipotesi statistica** una qualunque proposizione inerente un parametro di una variabile aleatoria X di cui si conosce la distribuzione.

Un'ipotesi statistica normalmente chiede di verificare che non ci sia discordanza tra il parametro osservato sul campione e quello corrispondente della popolazione; riferendoci agli stessi esempi, l'ipotesi da verificare è che sia $\mu = 3$ nel primo caso, $\mu = 1$ nel secondo, nel limite previsto di errore, cioè che, selezionato un campione dalla popolazione, il valore medio osservato non sia diverso da quello previsto.

In generale, indicato con ϑ il parametro della popolazione, l'ipotesi $H_0 : \vartheta = \vartheta_0$ viene detta **ipotesi nulla** ed esprime la condizione che non c'è differenza tra il valore ϑ_0 e il corrispondente valore calcolato sul campione.
Qualunque altra ipotesi diversa da quella nulla si dice **ipotesi alternativa** e si indica con H_1.

Data l'ipotesi nulla $H_0 : \vartheta = \vartheta_0$, le ipotesi alternative più significative sono le seguenti:

$$H_1 : \vartheta \neq \vartheta_0 \qquad H_1 : \vartheta > \vartheta_0 \qquad H_1 : \vartheta < \vartheta_0$$

Relativamente ai precedenti esempi:

- ipotesi nulla: $H_0 : \mu = 3$ $H_0 : \mu = 1$
- possibili ipotesi alternative: $H_1 : \mu > 3$ $H_1 : \mu \neq 1$

6.2 Le regole di decisione

Consideriamo una popolazione X, un suo parametro ϑ e formuliamo l'ipotesi che sia $\vartheta = \vartheta_0$ (ipotesi nulla). Estratto da X un campione di ampiezza n ci chiediamo se le osservazioni sul campione confermano o contraddicono l'ipotesi fatta.
Per esempio, se l'erba cresce mediamente 3,5cm possiamo considerare che l'ipotesi $\mu = 3$ sia vera? Analogamente, se la media dei pesi delle scatole di zucchero è 1,05kg, possiamo considerare che l'ipotesi $\mu = 1$ sia vera?

La risposta a questo tipo di domanda può essere data in base a un **test di significatività**, cioè un test che, in base alla misura della discrepanza tra quanto si osserva nel campione e quanto è previsto dall'ipotesi nulla, stabilisca se accettare o rifiutare l'ipotesi.
Un test di questo tipo si basa sulla determinazione dell'intervallo di confidenza entro il quale, con una data probabilità, cade il parametro. Basta allora ricordare quello che abbiamo imparato a questo riguardo nei precedenti paragrafi. Dobbiamo però distinguere due tipi di test a seconda di come si configura l'ipotesi alternativa.

I caso: il test a due code

Supponiamo di voler testare l'ipotesi nulla

$H_0 : \vartheta = \vartheta_0$ e sia $H_1 : \vartheta \neq \vartheta_0$ l'ipotesi alternativa

Estratto dalla popolazione un campione di ampiezza n, sia T_n uno stimatore del parametro ϑ; se T_n ha una distribuzione normale, si può trasformare il valore della stima nella variabile standardizzata z e determinare l'intervallo in cui è compresa la differenza fra ϑ_0 e la sua stima. Fissato un livello di fiducia $1 - \alpha$ e, di conseguenza, il valore critico $z_{1-\frac{\alpha}{2}}$, si ha che:

- vale $1 - \alpha$ la probabilità che il valore z della variabile standardizzata appartenga all'intervallo $(-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}})$;
- vale α la probabilità che z cada al di fuori di tale intervallo.

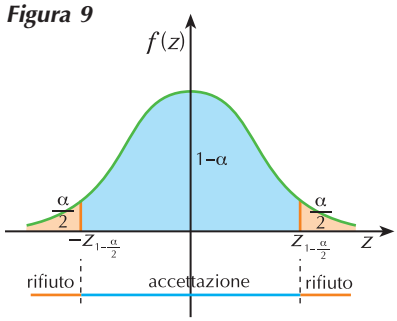
La regola di decisione basata su queste osservazioni è dunque la seguente:

- si accetta l'ipotesi nulla (o meglio non la si rifiuta) se $|z| < z_{1-\frac{\alpha}{2}}$; in altre parole non si ritiene significativa la differenza fra il valore stimato e ϑ_0 e si attribuisce la differenza a fluttuazioni casuali del campione;
- si rifiuta l'ipotesi nulla se $|z| \geq z_{1-\frac{\alpha}{2}}$; in questo caso si ritiene che la differenza fra il valore stimato e ϑ_0 sia significativa ed attribuibile ad altre circostanze piuttosto che alla naturale variabilità del campione.

Il valore di α rappresenta quindi il **livello di significatività** della differenza fra stima e valore del parametro. I livelli di significatività che si usano in genere sono $\alpha = 0{,}05$ e $\alpha = 0{,}01$ anche se possono essere usati valori diversi.

In ***figura 9*** abbiamo rappresentato la zona di accettazione dell'ipotesi e quella di rifiuto; per la forma che assume la zona di rifiuto, questa regola di decisione prende il nome di **test di significatività bilaterale o test a due code**. Quando il parametro ϑ cade nella zona di rifiuto, si dice che il **test è significativo**.

Figura 9



Quando si ha a che fare con **piccoli campioni** e la varianza della popolazione non è nota sappiamo poi che, per determinare l'intervallo di confidenza, dobbiamo riferirci alla distribuzione t di Student.

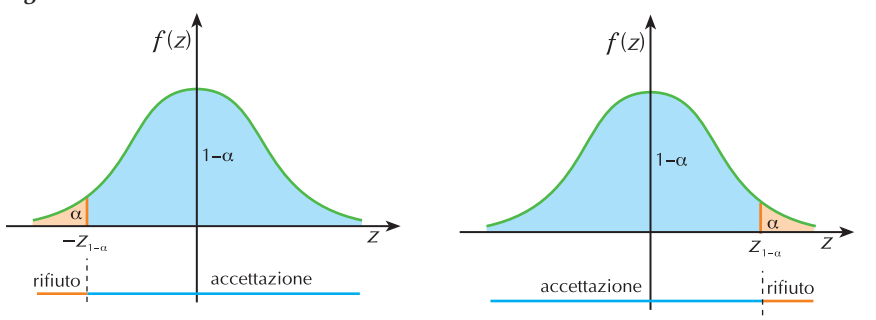
II caso: il test a una coda

Supponiamo adesso di voler testare l'ipotesi nulla

$H_0 : \vartheta = \vartheta_0$ e sia $H_1 : \vartheta > \vartheta_0$ l'ipotesi alternativa (oppure $H_1 : \vartheta < \vartheta_0$)

In questo caso si considerano come significative quelle differenze che vanno solo in una direzione (***figura 10***) e la regola di decisione prende il nome di **test di significatività unilaterale** o **test a una coda**.

Figura 10



a. Test unilaterale sinistro

b. Test unilaterale destro

Per determinare il valore critico dobbiamo risolvere l'equazione (riferisciti ancora alla ***figura 10***)

$$p(Z < a) = 1 - \alpha$$

Ricercando nelle tavole della gaussiana standardizzata il valore $1 - \alpha$, troviamo il valore critico a che indicheremo questa volta con il simbolo $z_{1-\alpha}$.

Ad esempio, se per un test bilaterale al 5% il valore critico è $z_{1-\frac{\alpha}{2}} = 1{,}96$, per un test unilaterale allo stesso livello di significatività il valore critico è $z_{1-\alpha} \approx 1{,}645$.

Riportiamo nella seguente tabella i valori critici $z_{1-\frac{\alpha}{2}}$ per test bilaterali e $z_{1-\alpha}$ per test unilaterali per alcuni valori di α fra i più usati

livello di significatività	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}005$	$\alpha = 0{,}002$
$z_{1-\frac{\alpha}{2}}$ (test a due code)	1,645	1,96	2,58	2,81	3,08
$z_{1-\alpha}$ (test a una coda)	1,28	1,645	2,33	2,58	2,88

La regola di decisione per un test a una coda è dunque la seguente:

- si accetta o comunque non si rifiuta l'ipotesi H_0 con $H_1 : \vartheta > \vartheta_0$ se $z < z_{1-\alpha}$ (oppure, nel caso dell'ipotesi $H_1 : \vartheta < \vartheta_0$, se $z > -z_{1-\alpha}$);
- si rifiuta l'ipotesi H_0 con $H_1 : \vartheta > \vartheta_0$ se $z \geq z_{1-\alpha}$ (oppure, nel caso dell'ipotesi $H_1 : \vartheta < \vartheta_0$, se $z \leq -z_{1-\alpha}$).

ESEMPI

1. Un'azienda produce delle pellicole per radiografie che hanno uno spessore medio di 5mm, con scarto quadratico medio di 1,4mm. Un controllo periodico di qualità porta all'estrazione di un campione bernoulliano di 100 pellicole e accerta uno spessore medio di 5,6mm. Ci chiediamo se, ad un livello di significatività del 5%, la produzione è sotto controllo.

La produzione è sotto controllo se la variazione dello spessore medio è del tutto casuale e non è impu-

tabile agli impianti di produzione. L'ipotesi che dobbiamo verificare è quindi

$$H_0 : \mu = 5 \qquad \text{con l'ipotesi alternativa} \qquad H_1 : \mu \neq 5$$

e, se tale affermazione risulta vera ai livelli di fiducia del 95%, potremo accettare l'ipotesi.

Le informazioni di cui disponiamo sono:

- la numerosità del campione $n = 100$ (poiché $n > 30$ possiamo parlare di grande campione)
- la media della popolazione $\mu = 5\text{mm}$
- lo scarto quadratico medio della popolazione $\sigma = 1{,}4\text{mm}$
- la media del campione $\overline{x} = 5{,}6\text{mm}$
- il livello di significatività $\alpha = 0{,}05$

Visto che conosciamo media e varianza della popolazione possiamo valutare l'intervallo entro cui deve cadere il valor medio del campione (paragrafo 5.1) e decidere in base ai risultati ottenuti se accettare o rifiutare l'ipotesi H_0, oppure applicare il test bilaterale.

I modo: intervallo in cui deve cadere $\overline{x}$

Il punto critico ad un livello di fiducia del 95% è 1,96; l'intervallo di confidenza in cui dovrebbe trovarsi il valor medio del campione per poter accettare l'ipotesi nulla ha quindi per estremi i valori $\mu \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, cioè, nel nostro caso, $5 \pm 1{,}96\frac{1{,}4}{\sqrt{100}}$; l'intervallo è quindi $(4{,}7256;\ 5{,}2744)$.

Poiché $\overline{x} = 5{,}6$ non cade nell'intervallo, dobbiamo rifiutare H_0 e concludere che la produzione non è sotto controllo.

II modo: test bilaterale

Per prima cosa trasformiamo il valore medio del campione $\overline{x}$ nel corrispondente valore della variabile normale standardizzata

$$|z| = \left|\frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right| \qquad \text{cioè} \qquad |z| = \left|\frac{5{,}6 - 5}{\frac{1{,}4}{\sqrt{100}}}\right| \approx 4{,}2857$$

per $\alpha = 0{,}05$ abbiamo $z_{1-\frac{\alpha}{2}} = 1{,}96$.

Poiché $|z| > 1{,}96$ l'ipotesi è da rifiutare.

VERIFICA DI COMPRENSIONE

1. Nella verifica di un'ipotesi statistica:

a. l'ipotesi nulla esprime la condizione che il parametro ϑ sia uguale a zero V F

b. la sola ipotesi alternativa è che sia $\vartheta \neq \vartheta_0$ V F

c. un'ipotesi che si ritiene accettabile a livello di significatività del 5%, può essere rifiutata a livello del 3% V F

d. la scelta del test di significatività a una coda oppure a due code dipende dalla formulazione dell'ipotesi alternativa. V F

I concetti e le regole

Popolazione e campione

Lo studio di un fenomeno statistico si basa sull'analisi di un **campione** della popolazione; è poi l'inferenza statistica che si occupa di stabilire le regole in base alle quali estendere le osservazioni fatte sull'intera popolazione.
Il rapporto tra la numerosità del campione e l'intera popolazione si chiama **tasso di campionamento** e si esprime di solito in forma percentuale.
Le modalità di estrazione di un campione da una popolazione sono diverse; la più semplice da trattare è quella che riguarda il **campionamento bernoulliano**; questo tipo di campionamento si rifà al criterio dell'urna e delle estrazioni con reimmissione.

Le variabili campionarie

L'esito della i-esima estrazione di un elemento del campione è una variabile aleatoria X_i che può assumere un qualsiasi valore della popolazione. Ogni campione di ampiezza n è un **vettore aleatorio** $(X_1, X_2, ..., X_n)$ le cui componenti sono le variabili aleatorie X_i; il suo spazio campionario è l'insieme Ω di tutte le realizzazioni campionarie di ampiezza n estraibili dalla popolazione.

Parametri e stimatori

Un **parametro** di una popolazione è un qualsiasi indice che caratterizza in modo sintetico una popolazione. Di solito i parametri di una popolazione non sono noti, di essi si può però fare una stima mediante il corrispondente parametro del campione; il parametro del campione prende il nome di **stimatore**.
Uno stimatore si dice:

- **corretto** se il suo valore atteso è uguale al parametro ϑ che deve stimare
- **consistente** se, al crescere dell'ampiezza n del campione, la sua varianza tende a zero
- **più efficiente** di un altro stimatore se la sua varianza è minore della varianza dell'altro stimatore.

I principali stimatori dei parametri di una popolazione sono la media, la varianza e la proporzione campionaria e si verifica che:

- la **media campionaria**, definita dalla relazione $\overline{X_n} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ è uno stimatore corretto e consistente della media della popolazione
- la **varianza campionaria**, definita dalla relazione $S_n^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X_n})^2$ non è uno stimatore corretto della varianza della popolazione
  per avere uno stimatore corretto si deve ricorrere alla varianza campionaria corretta definita dalla relazione
  $$\widehat{S}_n^2 = \frac{n}{n-1} \cdot S_n^2$$
- la **proporzione** o **frequenza campionaria**, definita dalla relazione $F = \frac{K}{n}$ è uno stimatore corretto e consistente della frequenza relativa della popolazione.

La distribuzione normale

Se una popolazione ha distribuzione normale con media μ e varianza σ^2, anche la variabile aleatoria media campionaria ha distribuzione normale con la stessa media μ e con varianza $\frac{\sigma^2}{n}$ essendo n l'ampiezza del campione.
Inoltre, qualunque sia il tipo di distribuzione di una popolazione con media μ e varianza σ^2, la media campionaria ha distribuzione normale con la stessa media μ e con varianza $\frac{\sigma^2}{n}$ se n è sufficientemente grande.

La stima dei parametri

La stima di un parametro può essere:

- **puntuale** e in questo caso restituisce un valore preciso per il parametro incognito, corredato da un errore
  - la stima di una media viene fatta mediante la media campionaria e l'errore medio di campionamento è dato da:

    $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ se è noto σ $\qquad$ $S_{\overline{x}} = \frac{\hat{S}}{\sqrt{n}}$ se non è noto σ

  - la stima di una frequenza viene fatta mediante la frequenza campionaria e l'errore medio di campionamento è dato da $s_f = \sqrt{\frac{f(1-f)}{n}}$
- **per intervallo** e in questo caso restituisce un intervallo che, con un fissato margine di errore, contiene il valore vero del parametro.

La verifica di ipotesi

Una ipotesi statistica è una affermazione che viene fatta in merito ad un parametro ϑ di una variabile aleatoria X. Si chiama **ipotesi nulla** l'affermazione $H_0 : \vartheta = \vartheta_0$; una qualunque ipotesi H_1 diversa da quella nulla viene detta **ipotesi alternativa**.
Tra le ipotesi alternative, le più significative sono:

- $H_1 : \vartheta \neq \vartheta_0$ che dà origine a un test a due code
- $H_1 : \vartheta > \vartheta_0$ oppure $H_1 : \vartheta < \vartheta_0$ che danno origine a un test a una coda

Il campionamento

POPOLAZIONE E CAMPIONE

la teoria è a pag. 1

Comprensione

1 Barra vero o falso.

a. Un campione è un qualunque sottoinsieme proprio della popolazione. V F

b. Il tasso di campionamento è un numero che indica quanti campioni si possono estrarre dalla popolazione. V F

c. Un campionamento è probabilistico solo se tutti gli elementi della popolazione hanno la stessa probabilità di essere estratti. V F

d. Nelle indagini statistiche, laddove è possibile, è opportuno che il campione sia di tipo probabilistico. V F

2 In un campionamento casuale semplice:

a. ogni unità statistica ha la stessa probabilità delle altre di essere estratta

b. le unità statistiche hanno diverse probabilità di essere estratte

c. tutte le unità statistiche hanno probabilità $\frac{1}{2}$ di essere estratte

d. tutte le precedenti affermazioni sono errate.

Qual è la sola affermazione corretta tra le precedenti?

3 Barra vero o falso. Per effettuare un campionamento sistematico:

a. la popolazione deve essere dapprima ordinata in base a un qualche criterio V F

b. si seleziona il campione scegliendo le unità che si preferiscono V F

c. si seleziona il campione prelevando una unità ogni k, essendo k un numero fissato a priori. V F

Applicazione

4 Completa:

a. ampiezza popolazione: 1300
ampiezza campione: 35
tasso di campionamento: .....

b. ampiezza popolazione: 12000
ampiezza campione: 300
tasso di campionamento: .....

c. ampiezza popolazione: 840
ampiezza campione: 200
tasso di campionamento: .....

d. ampiezza popolazione: 15800
ampiezza campione: 79
tasso di campionamento: .....

5 Da una popolazione di 1200 unità si vuole estrarre un campione con un tasso di campionamento del 15%; qual è l'ampiezza del campione? [180]

6 Da una certa popolazione si è estratto un campione di ampiezza 135 in base a un tasso di campionamento del 15%. Qual è l'ampiezza della popolazione? [900]

7 Da una popolazione di 800 unità si vuole estrarre un campione di 85 elementi; qual è il tasso di campionamento? [10,625%]

8 Da una popolazione di 3000 unità si vuole estrarre un campione bernoulliano formato da 150 unità statistiche. Quanti sono i possibili campioni? $[3000^{150}]$

9 Una classe è composta da 28 studenti; quanti sono i campioni bernoulliani di ampiezza 3? [21952]

10 Da una produzione aziendale di 3 500 pezzi si estrae un campione con un tasso di campionamento del 2%. Quanti sono i possibili campioni? $[3\,500^{70}]$

11 Considera la popolazione i cui elementi sono quelli che appartengono all'insieme $\{a, b, c, d, e\}$; dopo aver stabilito quanti sono i possibili campioni bernoulliani di ampiezza 2, elencali.

12 Una popolazione è formata da soli quattro elementi ed è $\{1, 2, 3, 4\}$; elenca i possibili campioni di ampiezza 3.

13 Da una popolazione di 50 elementi si esegue un campionamento con un tasso del 10%; quanti campioni bernoulliani si possono costruire? $[50^5]$

14 Un'urna contiene quattro palline contrassegnate dai numeri 10, 15, 20, 25; si estraggono campioni di due palline. Determina:
a. lo spazio campionario e il numero dei campioni bernoulliani
b. la probabilità di estrarre un campione di due palline aventi somma 30. $\left[16;\ \frac{3}{16}\right]$

15 Al 31 Dicembre un'azienda rileva i giorni di assenza dei suoi dipendenti e i dati sono riassunti nella seguente tabella:

giorni assenza	3	4	5	6	7	8	9
n. dipendenti	4	16	28	46	18	5	3

Dopo aver determinato quanti campioni di 4 elementi si possono estrarre dalla popolazione, determina la probabilità di costruire un campione di dipendenti che hanno tutti fatto 6 giorni di assenza.

$\left[120^4,\ \left(\frac{23}{60}\right)^4\right]$

PARAMETRI E STIMATORI

la teoria è a pag. 6

RICORDA

In ogni campionamento bernoulliano:

- data una popolazione avente media μ e varianza σ^2:
  - la variabile aleatoria media campionaria $\overline{X_n} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ ha valor medio uguale a μ e varianza uguale a $\frac{\sigma^2}{n}$
  - la variabile aleatoria varianza campionaria $S_n^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X_n})^2$ ha valor medio uguale a $\frac{n}{n-1} \cdot \sigma^2$.

- quando invece in una indagine statistica si analizza il possedere o meno un determinato carattere, si usa la frequenza campionaria definita dal rapporto $F = \frac{K}{n}$; se p è la frequenza con cui compare il carattere nella popolazione, si verifica che $E(F) = p$.

Comprensione

16 Un parametro di una popolazione è:
a. il dato che ha la maggiore frequenza tra quelli rilevati
b. un indice che rappresenta mediamente i dati rilevati sulla popolazione
c. un indice che esprime in modo sintetico i valori rilevati su un campione della popolazione.
d. un indice che caratterizza in modo sintetico una popolazione.

17 Se ϑ è un parametro di una popolazione, una stima di ϑ è:
a. uno dei possibili valori che ϑ può assumere
b. il valore di ϑ calcolato sul campione preso in esame
c. il valore di ϑ calcolato su tutti i possibili campioni che si possono estrarre dalla popolazione
d. il valore di ϑ calcolato su tutti i campioni di ampiezza n che si possono estrarre dalla popolazione.

18 Uno stimatore si dice corretto se:
a. il suo valore atteso è uguale al corrispondente parametro della popolazione
b. al crescere dell'ampiezza n del campione si avvicina al valore del corrispondente parametro della popolazione
c. non cambia il suo valore al variare del campione
d. il suo valore si discosta di poco da quello del corrispondente parametro della popolazione.

19 Uno stimatore si dice consistente se, al crescere dell'ampiezza n del campione, tende a zero:
a. il suo valore medio **b.** la sua varianza **c.** il suo valore

20 Uno stimatore T_1 è più efficiente di uno stimatore T_2 se, al crescere dell'ampiezza n del campione:
a. $V(T_1) < V(T_2)$ **b.** $V(T_1) > V(T_2)$ **c.** $V(T_1) \to 0 \wedge V(T_2) \neq 0$

21 Sono stimatori corretti e consistenti:
a. la media campionaria
b. la varianza campionaria
c. la frequenza campionaria.

Applicazione

22 **ESERCIZIO GUIDA**

Assegnata la popolazione statistica individuata dai seguenti redditi annuali di 5 persone (espressi in migliaia di euro)

persone	A	B	C	D	E
redditi	15	17	18	20	25

determina:
a. la media e la varianza dei redditi della popolazione;

b. lo spazio campionario dei campioni di due elementi che possono essere estratti con estrazione bernoulliana;

c. la media e la varianza della media campionaria, verificando le relazioni con la media e la varianza della popolazione.

a. La media dei redditi della popolazione è data da

$$\mu = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{15 + 17 + 18 + 20 + 25}{5} = 19$$

mentre la varianza è $\sigma^2 = E(X^2) - [E(X)]^2 = \frac{15^2 + 17^2 + 18^2 + 20^2 + 25^2}{5} - 19^2 = 11,6$

b. Il numero dei campioni di due elementi estratti con estrazione bernoulliana è

$$D'^{(r)}_{5,2} = 5^2 = 25$$

lo spazio campionario è costituito dalle coppie:

AA	AB	AC	AD	AE
BA	BB	BC	BD	BE
CA	CB	CC	CD	CE
DA	DB	DC	DD	DE
EA	EB	EC	ED	EE

i cui redditi sono rispettivamente:

15; 15	15; 17	15; 18	15; 20	15; 25
17; 15	17; 17	17; 18	17; 20	17; 25
18; 15	18; 17	18; 18	18; 20	18; 25
20; 15	20; 17	20; 18	20; 20	20; 25
25; 15	25; 17	25; 18	25; 20	25; 25

c. Troviamo i valori medi di ogni campione:

15	16	16,5	17,5	20
16	17	17,5	18,5	21
16,5	17,5	18	19	21,5
17,5	18,5	19	20	22,5
20	21	21,5	22,5	25

La distribuzione della media campionaria è quindi

$\overline{X}_2$	15	16	16,5	17	17,5	18	18,5	19	20	21	21,5	22,5	25
Freq.	1	2	2	1	4	1	2	2	3	2	2	2	1
Prob.	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

Il valore medio di $\overline{X}_2$ è

$$E(\overline{X}_2) = \sum_{i=1}^{13} x_i \cdot p_i =$$

$$= 15 \cdot \frac{1}{25} + 16 \cdot \frac{2}{25} + 16{,}5 \cdot \frac{2}{25} + 17 \cdot \frac{1}{25} + 17{,}5 \cdot \frac{4}{25} + 18 \cdot \frac{1}{25} + 18{,}5 \cdot \frac{2}{25} + 19 \cdot \frac{2}{25} +$$

$$+20 \cdot \frac{3}{25} + 21 \cdot \frac{2}{25} + 21{,}5 \cdot \frac{2}{25} + 22{,}5 \cdot \frac{2}{25} + 25 \cdot \frac{1}{25} = 19.$$

Risulta pertanto verificato che $E(\overline{X}_2) = \mu$ nelle estrazioni bernoulliane.

La varianza di $\overline{X}_2$ è $V(\overline{X}_2) = E(\overline{X}_2^2) - \left[E(\overline{X}_2)\right]^2$.

Costruiamo la tabella dei valori di $\overline{X}_2^2$

$\overline{X}_2$	15	16	16,5	17	17,5	18	18,5	19	20	21	21,5	22,5	25
$\overline{X}_2^2$	225	256	272,25	289	306,25	324	342,25	361	400	441	462,25	506,25	625
Prob.	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

Anche in questo caso si ritrova che $V(\overline{X}) = \frac{\sigma^2}{n} = \frac{11{,}6}{2} = 5{,}8$ valida per estrazioni bernoulliane.

23 Sia X la popolazione dei possibili esiti del lancio di un dado. Determina:
a. la media e la varianza di tale popolazione; [3,5; 2,91$\overline{6}$]
b. lo spazio campionario dei campioni di due elementi che possono essere estratti con estrazione bernoulliana; [36]
c. la media e la varianza della media campionaria. [3,5; 1,458]

24 Una popolazione è costituita dai seguenti elementi: 0 5 10 15 20
Determina la media e la varianza della popolazione. Successivamente calcola la media e la varianza della distribuzione campionaria delle medie, con campionamento bernoulliano, nel caso in cui i campioni siano di ampiezza:
a. $n = 2$; **b.** $n = 3$; **c.** $n = 5$.
Osserva infine che, aumentando il numero degli elementi del campione, la distribuzione campionaria delle medie assume una varianza minore.
[popolazione: media = 10, varianza = 50; media : 10; varianze : 25; 16,67; 10]

25 Considera la popolazione statistica individuata dalle altezze degli N alunni della tua classe. Attraverso l'uso di programmi applicativi o mediante una tabella di numeri casuali, estrai 4 campioni di 10 elementi ciascuno e determina
a. le medie campionarie;
b. la media delle medie campionarie;
c. il grafico della distribuzione delle medie dei campioni.

26 Ripeti l'esercizio 25 considerando campioni di ampiezza 5.

27 Ripeti l'esercizio 25 considerando campioni di ampiezza 15.

28 Confronta i grafici ottenuti nei precedenti esercizi 25, 26, 27 e rileva differenze ed analogie.

29 Data la popolazione costituita dai seguenti elementi: 2 3 6 8 12
considera tutti i campioni di ampiezza 2 costruiti mediante estrazione con ripetizione dalla assegnata popolazione. Determina

a. la media della popolazione; [6,2]

b. la varianza della popolazione; [12,96]

c. la media della distribuzione campionaria della media; [6,2]

d. la varianza della distribuzione campionaria della media. [6,48]

IL CASO DELLA DISTRIBUZIONE NORMALE

la teoria è a pag. 11

Applicazione

30 **ESERCIZIO GUIDA**

La sezione media di alcune sbarre di ferro è 0,70cm con una deviazione standard di 0,01cm; qual è la probabilità che selezionando un campione casuale bernoulliano di 50 elementi si ottenga una sezione media del campione:

a. superiore a 0,701cm;

b. inferiore a 0,698cm.

La sezione media è $\mu = 0{,}70\text{cm}$, mentre $\sigma_{\overline{X}_{50}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0{,}01}{\sqrt{50}} = 0{,}0014$.

a. Si ha $P(\overline{X}_{50} \geq 0{,}701) = P(Z \geq z_0)$

con $z_0 = \dfrac{\overline{x} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \dfrac{0{,}701 - 0{,}70}{0{,}0014} \approx 0{,}71$

da cui $P(\overline{X}_{50} \geq 0{,}701) = P(Z \geq 0{,}71) = 1 - 0{,}7611 = 0{,}2389$.

b. Analogamente si ha $P(\overline{X} \leq 0{,}698) = P(Z \leq z_0)$

con $z_0 = \dfrac{\overline{x} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \dfrac{0{,}698 - 0{,}70}{0{,}0014} = -1{,}43$

da cui $P(\overline{X}_{50} \leq 0{,}698) = P(Z \leq -1{,}43) = 1 - 0{,}9236 = 0{,}0764$.

31 Un'azienda deve accettare degli anellini di alluminio proposti da un fornitore. Viene pertanto selezionato un campione casuale bernoulliano di 50 pezzi. Si stabilisce che siano acquistati solamente gli oggetti il cui diametro medio sia compreso tra 15,90cm e 16,20cm. Determina qual è la probabilità di accettare la fornitura avente un diametro medio di

a. 15,50cm e deviazione standard di 1cm; [0,002327]

b. 16,00cm e deviazione standard di 0,8cm. [0,7722]

32 Una ditta riceve degli elementi da un fornitore. Viene selezionato un campione casuale di 50 pezzi. Si stabilisce che si acquistino solamente gli oggetti la cui lunghezza media sia compresa tra 44,50cm e 45,00cm. Qual è la probabilità di accettare la fornitura avente una lunghezza media di 44,25cm e deviazione standard di 0,9cm? [0,025]

33 Per vagliare un acquisto si selezionano 40 elementi della partita di una fornitura; si decide di accettare

solo i pezzi il cui diametro medio sia compreso tra 123,00cm e 125,00cm. Determina la probabilità di accettare la fornitura i cui elementi abbiano un diametro medio di 122,50cm e deviazione standard di 1cm. [0,0008]

34 Considera i dati dell'esercizio precedente, supponendo però che il diametro medio sia di 123,20cm e la deviazione standard sia di 0,8cm. [0,9429]

35 Una pasticceria, al fine di decidere se acquistare una partita di merce offerta da un nuovo fornitore, seleziona e pesa un campione casuale di 40 tortine alla confettura. Stabilisce quindi di accettare l'offerta solo nel caso in cui la massa media del campione sia compresa tra 24,50g e 25,00g. Calcola qual è la probabilità di acquistare la fornitura avente una massa media di 24g e deviazione standard di 1g. [0,00071]

STIMA PUNTUALE DEI PARAMETRI

la teoria è a pag. 12

RICORDA

Una stima puntuale di un parametro di una popolazione X restituisce un singolo valore per il parametro incognito, corredato da un errore.

- La **stima di una media** viene fatta mediante la media campionaria e l'errore medio di campionamento è dato da

  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ se è noto σ $\qquad$ $s_{\bar{x}} = \frac{\hat{s}}{\sqrt{n}}$ se non è noto σ

- La **stima di una frequenza** viene fatta mediante la frequenza campionaria e l'errore medio di campionamento è dato da $s_f = \sqrt{\frac{f(1-f)}{n}}$

Comprensione

36 Da un lotto di 1000 elettrodomestici è stato estratto un campione di ampiezza 50 e si è rilevato che il numero medio di ore di funzionamento prima di un guasto è 509. Se lo scarto quadratico medio della popolazione è 62:

a. una stima del numero medio di ore di funzionamento è: ① 509 ② 65 ③ 72

b. l'errore medio di campionamento è: ① 1,24 ② 8,77 ③ 1,12

37 Da una popolazione di 3450 famiglie si estrae un campione di ampiezza $n = 250$ e si trova che 38 di esse hanno meno di due figli.

a. Una stima del numero di famiglie con meno di due figli è: ① 250 ② 380 ③ 524

b. L'errore medio di campionamento è: ① 0,023 ② 6,11 ③ 5,14

Applicazione

Stima puntuale di una media

38 **ESERCIZIO GUIDA**

Un campione di 10 sacchi di patate, scelto in modo casuale da una fornitura di 800 sacchi, sottoposto

alla pesatura, evidenzia una massa media di 5000g. Da rilevazioni precedenti è noto che la massa media è distribuita normalmente con scarto quadratico medio di 100g. Determina:

a. una stima puntuale della massa media di tutta la fornitura;

b. l'errore medio di campionamento.

Sappiamo che $\overline{x} = 5000\text{g}$ e $\sigma = 100\text{g}$

a. La massa media dei sacchi di patate è stimata da $\overline{x} = 5000\text{g}$

b. L'errore medio di campionamento è $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{10}} \approx 31{,}62$.

39 Ripeti l'esercizio precedente, nel caso in cui il campione sia costituito da 150 sacchi di patate. [5000g; 8,16g]

40 Un produttore esamina un campione di 50 batterie, estratto in modo casuale da una confezione di 1000 elementi. Tale campione, sottoposto ad un esame, presenta una durata media di 40 ore. Da controlli precedenti è noto che lo scarto quadratico medio della durata delle batterie è di 6 ore. Determina:

a. una stima puntuale della durata media di tutta la confezione; [40 ore]

b. l'errore medio di campionamento. [0,85 ore]

41 Da una ricerca si osserva che un campione casuale di 20 mense scolastiche, estratto da un elenco di 500 elementi, deve sostenere annualmente una spesa media di € 5000 per i rifornimenti di base. Sapendo, da osservazioni precedenti, che lo scarto quadratico medio è di € 1000, determina:

a. una stima puntuale della spesa media della totalità delle mense considerate; [€ 5000]

b. l'errore medio di campionamento. [€ 223,61]

42 **ESERCIZIO GUIDA**

Un negoziante preleva un campione casuale di 30 bulloni da una scatola che ne contiene 1500. Dall'esame del campione osserva che il diametro medio è 350mm con scarto quadratico medio di 10mm. Determiniamo:

a. un parametro che valuti il diametro medio dei bulloni della scatola;

b. lo scarto quadratico medio della media.

a. Il diametro medio dei bulloni è stimato da $\overline{x} = 350$

b. Non essendo assegnata σ, occorre stimare l'errore medio di campionamento calcolando la varianza corretta del campione, data da

$$\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{30}{29} 100$$

Pertanto, lo scarto quadratico medio è $s_{\overline{x}} = \frac{1}{\sqrt{30}} \sqrt{\frac{3000}{29}} \approx 1{,}8569\text{mm}$

43 Da una fornitura di 1000 vasetti di marmellata si estraggono 100 elementi per costituire un campione casuale in cui si osserva che la massa media di prodotto si attesta su 25g con uno scarto quadratico medio di 5g. Determina:

a. una stima puntuale che valuti la massa media della confezione di vasetti di marmellata; [25g]

b. lo scarto quadratico medio della media. [0,5025]

44 Da un gruppo di 150 scuole si costruisce un campione casuale di ampiezza 20 in cui si rileva che la spesa media di manutenzione ordinaria periodica è di € 3000 con uno scarto quadratico medio di € 150. Determina:

a. una stima puntuale che valuti la spesa media sostenuta da tutto il gruppo di scuole; [€ 3000]

b. lo scarto quadratico medio della media. [€ 34,41]

45 **ESERCIZIO GUIDA**

Da una rilevazione effettuata in una stazione ferroviaria si sono osservati i ritardi accumulati da 100 treni che costituiscono un campione casuale estratto dai 500 che hanno transitato nella stazione. Tali dati sono stati raccolti nella seguente tabella:

ritardo (min.)	[0 - 10)	[10 - 20)	[20 - 30)	[30 - 40)	[40 - 50)	[50 - 60]
n. treni	21	42	17	12	6	2

Determina il ritardo medio di tutta la popolazione dei treni e l'errore medio di campionamento.

Assumiamo come valore del ritardo di ogni classe di treni il valore centrale e calcoliamo il valore medio dei ritardi:

$$\overline{x} = \frac{5 \cdot 21 + 15 \cdot 42 + \ldots\ldots + 55 \cdot 2}{100} = 19,6$$

Possiamo adesso costruire la seguente tabella:

ritardo (min.)	n. treni	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$
[0 - 10)	21	−14,6	213,16
[10 - 20)	42	−4,6	21,16
[20 - 30)	17	5,4	29,16
[30 - 40)	12	15,4	237,16
[40 - 50)	6	25,4	645,16
[50 - 60]	2	35,4	1253,16

da cui ricaviamo che la varianza corretta è

$$\widehat{s}^2 = \frac{213,16 \cdot 21 + 21,16 \cdot 42 + \ldots\ldots + 1253,16 \cdot 2}{99} = \frac{15084}{99}$$

L'errore medio di campionamento è quindi: $s_{\overline{x}} = \frac{1}{\sqrt{100}} \sqrt{\frac{15084}{99}} \approx 1,2343.$

46 Da una rilevazione effettuata in una scuola si è osservata l'altezza di 80 studenti che sono un campione casuale estratto dai 1300 alunni dell'Istituto. Tali dati sono stati raccolti nella seguente tabella:

altezza (cm)	155 - 160	160 - 165	165 - 170	170 - 175	175 - 180	180 - 185	185 - 190
n. studenti	3	9	14	18	21	11	4

Determina l'altezza media di tutta la popolazione degli studenti e l'errore medio di campionamento. [173,375; 0,833]

47 Da una rilevazione effettuata in un cinema multisala si è osservata l'età di 100 spettatori che costituisco-

no un campione casuale estratto dai 1500 presenti alle proiezioni. Tali dati sono stati raccolti nella seguente tabella:

età (anni)	15 - 20	20 - 25	25 -30	30 - 35	35 - 40	40 - 45	45 - 50
n. spettatori	18	24	22	16	8	7	5

Determina l'età media di tutta la popolazione degli spettatori e l'errore medio di campionamento nel caso di campionamento bernoulliano. [28,15; 0,849]

Stima puntuale di una frequenza

48 **ESERCIZIO GUIDA**

In una popolazione scolastica di 10000 studenti si preleva un campione casuale di 300 elementi e si osserva che di questi 50 non praticano alcuna attività sportiva extrascolastica. Determiniamo una stima puntuale sulla percentuale di tutti gli alunni che non praticano alcuno sport in orario extrascolastico e l'errore medio di campionamento.

La frequenza relativa del campione è $f = \frac{k}{n} = \frac{50}{300}$ con scarto quadratico medio

$$s_f = \sqrt{\frac{f(1-f)}{n}} \approx 0{,}02152$$

49 Da una popolazione di 5000 studenti si preleva un campione casuale di 100 elementi e si osserva che 75 di essi leggono almeno un quotidiano al giorno. Determina una stima puntuale sulla percentuale di tutti gli alunni che leggono almeno un quotidiano e l'errore medio di campionamento. [0,75; 0,0433]

50 In riferimento all'esercizio precedente calcola il numero medio di studenti della popolazione che leggono almeno un quotidiano e l'errore medio di campionamento. [3750; 217]

51 In un'urna sono presenti palline di colore bianco, di colore rosso, di colore giallo e di colore nero. Se ne estraggono 400, con reimmissione, di cui 120 nere e si osserva che in tale campione non vi è alcuna pallina bianca. Determina una stima puntuale sulla percentuale di palline nere contenute nell'urna e l'errore medio di campionamento. [0,30; 0,0229]

52 Da un lotto di 32000 biciclette si preleva un campione casuale di 500 elementi e si osserva che 325 di esse sono mountain bike. Determina una stima puntuale sulla percentuale di mountain bike presenti nel campione e l'errore medio di campionamento. [0,65; 0,0213]

53 In riferimento all'esercizio precedente calcola il numero medio di mountain bike del lotto e l'errore medio di campionamento. [20800; 682]

54 Da una popolazione di 4000 auto si preleva un campione casuale di 400 elementi e si osserva che 160 di esse hanno i fari fendinebbia. Determina una stima puntuale sulla percentuale di auto con fendinebbia presenti nel campione e l'errore medio di campionamento. [0,40; 0,0245]

55 In riferimento all'esercizio precedente calcola il numero medio di auto con fendinebbia della popolazione e l'errore medio di campionamento. [1600; 98]

56 Da un lotto di 1500 zaini, posti in vendita in un grande magazzino, si preleva un campione casuale di 100 elementi e si osserva che 30 di essi sono di marca *"A"*. Determina una stima puntuale sulla percentuale di zaini di marca *"A"* presenti nel campione e l'errore medio di campionamento. [0,30; 0,0458]

RICORDA

- Posto $Z = \dfrac{\overline{X}_n - \mu}{\dfrac{\sigma}{\sqrt{n}}}$
  - $p\left(\mu - z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}} < \overline{X}_n < \mu + z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right) = 1-\alpha$
  - $p\left(\overline{X}_n - z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right) = 1-\alpha$
  - se il valore di σ non è noto ed il campione è grande, basta sostituire σ con la varianza corretta del campione
  - se il valore di σ non è noto ed il campione è piccolo si usa la variabile $t = \dfrac{\overline{X}_n - \mu}{\dfrac{S}{\sqrt{n-1}}}$ ed è

    $p\left(\overline{X}_n - {}_\nu t_{1-\alpha}\dfrac{S}{\sqrt{n-1}} < \mu < \overline{X}_n + {}_\nu t_{1-\alpha}\dfrac{S}{\sqrt{n-1}}\right) = 1-\alpha \quad \text{con} \quad \nu = n-1$

- $p\left(f - z_{1-\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}} < p < f + z_{1-\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}}\right) = 1-\alpha$
- Con una valutazione di prudenza si ha che $p\left(f - z_{1-\frac{\alpha}{2}}\dfrac{1}{2\sqrt{n}} < p < f + z_{1-\frac{\alpha}{2}}\dfrac{1}{2\sqrt{n}}\right) = 1-\alpha$

Applicazione

Stima per intervallo di una media

57 **ESERCIZIO GUIDA**

In un'azienda una macchina confeziona sacchetti di farina di massa media di 152g con uno scarto quadratico medio di 8g. Scelto un campione casuale di 10 confezioni di farina, calcoliamo tra quali valori deve essere compresa la massa media del campione ad un livello fiduciario del:

a. 95%; **b.** 99%; **c.** 99,73%; **d.** 99,90%.

Siano $\mu = 152\text{g}$ e $\sigma = 8\text{g}$. Lo scarto quadratico medio è $\sigma_{\overline{x}} = \dfrac{8}{\sqrt{10}} \approx 2{,}5298$

Si ha dunque

a. $p(152 - 1{,}96 \cdot 2{,}5298 < \overline{x} < 152 + 1{,}96 \cdot 2{,}5298) = 0{,}95$

Al livello fiduciario del 95%, l'intervallo delle medie dei campioni è $(147{,}04;\ 156{,}96)$ da cui risulta che se la macchina che produce tali confezioni è sottoposta a controllo nel 95% dei campioni di 10 sacchetti, la massa media delle confezioni appartiene all'intervallo, mentre nel 5% la massa non appartiene all'intervallo.

b. $p(152 - 2{,}58 \cdot 2{,}5298 < \overline{X}_{10} < 152 + 2{,}58 \cdot 2{,}5298) = 0{,}99$

Al livello fiduciario del 99%, l'intervallo delle medie dei campioni è $(145{,}47;\ 158{,}53)$

c. $p(152 - 3 \cdot 2{,}5298 < \overline{X}_{10} < 152 + 3 \cdot 2{,}5298) = 0{,}9973$

Al livello fiduciario del 99,73%, l'intervallo delle medie dei campioni è (144,41; 159,59)

d. $p(152 - 3{,}29 \cdot 2{,}5298 < \overline{X}_{10} < 152 + 3{,}29 \cdot 2{,}5298) = 0{,}9990$

Al livello fiduciario del 99,90%, l'intervallo delle medie dei campioni è (143,68; 160,32)

58 In una ditta una macchina produce aste di legno di altezza media di 15cm con uno scarto quadratico medio di 3cm. Scelto un campione casuale di 100 aste, calcola tra quali valori deve essere compresa l'altezza media del campione ad un livello fiduciario del

a. 95%; [(14,41; 15,59)]

b. 99%; [(14,23; 15,77)]

c. 99,73%; [(14,10; 15,90)]

d. 99,90%. [(14,01; 15,99)]

59 Una ditta produce bulloni il cui diametro medio è di 28,5mm con uno scarto quadratico medio di 1,5mm. Si estrae un campione di n elementi. Calcola tra quali valori deve essere compreso il diametro medio del campione ad un livello fiduciario del 99%, se

a. $n = 10$; [(27,28; 29,72)]

b. $n = 20$; [(27,63; 29,36)]

c. $n = 100$. [(28,11; 28,89)]

60 Da una fornitura di 1 500 viti si estrae un campione casuale di 100 elementi. La lunghezza media rilevata nel campione è di 502mm con scarto quadratico medio di 20mm. Calcola tra quali valori deve essere compresa la lunghezza media delle viti della popolazione ad un livello fiduciario del:

a. 95%; [(498,08; 505,92)]

b. 99%; [(496,84; 507,16)]

c. 99,73%; [(496; 508)]

d. 99,90%. [(495,42; 508,58)]

61 Da una popolazione di 300 studenti se ne estrae un campione casuale di 50. Nel campione si rileva che il tempo medio dedicato al sonno è di 7,5 ore con scarto quadratico medio di 0,5 ore. Calcola tra quali valori deve essere compresa la media della popolazione ad un livello fiduciario del:

a. 95%; [(7,36; 7,64)]

b. 99%; [(7,32; 7,68)]

c. 99,90%; [(7,27; 7,73)]

d. determina inoltre a quale livello fiduciario l'intervallo (7,29; 7,71) contiene la media considerata. [99,73%]

62 **ESERCIZIO GUIDA**

In un campione di 20 studenti l'altezza media è di 178cm con uno scarto quadratico medio di 4cm. Supponendo che la popolazione sia distribuita normalmente, calcola l'intervallo di confidenza della media della popolazione ad un livello fiduciario del:

a. 95%;

b. 99%.

Sappiamo che $\overline{x} = 178$cm e che $s = 4$cm. Trattandosi di un piccolo campione e non conoscendo σ dobbiamo servirci della variabile t; essendo

$$\frac{s}{\sqrt{n-1}} = \frac{4}{\sqrt{19}} \approx 0{,}918$$

si ha che:

a. $p(178 - {}_{19}t_{0,95} \cdot 0{,}918 < \mu < 178 + {}_{19}t_{0,95} \cdot 0{,}918) = 0{,}95$.

Al livello fiduciario del 95%, l'intervallo della media della popolazione è (176,08; 179,92).

b. Analogamente

$p(178 - {}_{19}t_{0,99} \cdot 0{,}918 < \mu < 178 + {}_{19}t_{0,99} \cdot 0{,}918) = 0{,}99$.

Al livello fiduciario del 99%, l'intervallo della media della popolazione è (175,37; 180,63).

63 In un campione di 20 bulloni, il diametro medio è di 152mm con scarto quadratico medio di 2,5mm. Supponendo che la popolazione sia distribuita normalmente, calcola tra quali valori deve essere compresa la media della popolazione ad un livello fiduciario dell'80%. [(151,24; 152,76)]

64 In un campione di 20 iscritti ad una biblioteca comunale, il numero medio annuale di libri presi in prestito è di 10 con scarto quadratico medio di 2,5. Supponendo che la popolazione sia distribuita normalmente, calcola tra quali valori deve essere compresa la media della popolazione ad un livello fiduciario del:

a. 90%; **b.** 95%; **c.** 99%. [**a.** (9,01; 10,99); **b.** (8,80; 11,20); **c.** (8,36; 11,64)]

65 In un campione di 15 studenti, il costo medio riservato all'acquisto di quotidiani è di € 1,5 con scarto quadratico medio di € 0,05. Supponendo che la popolazione sia distribuita normalmente, calcola tra quali valori deve essere compresa la media della popolazione ad un livello fiduciario del:

a. 80%; **b.** 95%. [**a.** (1,48; 1,52); **b.** (1,47; 1,53)]

66 In un campione di 10 botti di vino presenti in una azienda vinicola, il numero medio di litri di vino è di 94 con scarto quadratico medio di 8. Supponendo che la popolazione sia distribuita normalmente, calcola tra quali valori deve essere compresa la media della popolazione ad un livello fiduciario del:

a. 90%; **b.** 99,9%. [**a.** (89,12; 98,88); **b.** (81,25; 106,75)]

Stima per intervallo di una frequenza

67 **ESERCIZIO GUIDA**

Da una popolazione avente distribuzione di tipo binomiale si estrae un campione casuale di 150 elementi di cui 50 godono di una assegnata proprietà.
Calcola la stima per intervallo del parametro p della popolazione ad un livello di fiducia del 95%.

Nelle ipotesi dell'esercizio risulta $p\left(f - 1{,}96\sqrt{\frac{f(1-f)}{n}} < p < f + 1{,}96\sqrt{\frac{f(1-f)}{n}}\right) = 0{,}95$

da cui $p(0{,}33 - 1{,}96 \cdot 0{,}038 < p < 0{,}33 + 1{,}96 \cdot 0{,}038) = 0{,}95$

ed, a un livello fiduciario del 95%, l'intervallo risulta (0,258; 0,409).

68 In riferimento ai dati dell'esercizio precedente, stima l'intervallo richiesto, ad un livello di fiducia del

a. 99%; **b.** 99,73%; **c.** 99,90%. [**a.** (0,232; 0,428); **b.** (0,216; 0,444); **c.** (0,205; 0,455)]

69 In una popolazione di 500 bambini, affidati alle cure di un asilo nido, si determina che il 62,5% di un campione casuale di 80 elementi apprezza un dato omogeneizzato. Stima l'intervallo di confidenza, su tutta la popolazione, ad un livello del

a. 95%; **b.** 99%. [**a.** (259; 366); **b.** (243; 382)]

70 **ESERCIZIO GUIDA**

Se la popolazione degli esercizi 67 e 68 è costituita da 5000 elementi, calcola l'intervallo di confidenza relativo all'intera popolazione ad un livello fiduciario del

a. 95%; **b.** 99%; **c.** 99,73%; **d.** 99,90%.

a. Nelle ipotesi di grande campione, con distribuzione binomiale, si ha

$$p\left(N f - 1{,}96N\sqrt{\frac{f(1-f)}{n}} < K < N f + 1{,}96\, N \sqrt{\frac{f(1-f)}{n}}\right) = 0{,}95$$

71 In un referendum il 54,5% della popolazione ha votato SI. In un campione casuale di 300 elementi è emerso che il 60% ha confermato il SI. Verifica se tale campione cade nell'intervallo di confidenza della popolazione, ad un livello fiduciario del

a. 95%; **b.** 99%; **c.** 99,73%. [**a.** si; **b.** si; **c.** si]

72 In una popolazione di 1000 macchine prodotte, si determina che, in un campione casuale di 100 elementi, 40 sono gravemente danneggiate. Stima l'intervallo di confidenza, su tutta la popolazione, ad un livello del

a. 95%; **b.** 99%; **c.** 99,90%. [**a.** (304; 496); **b.** (274; 526); **c.** (239; 561)]

73 In una popolazione di 3000 studenti si determina che in un campione casuale di 300 elementi, 120 studiano da più di 5 anni. Stima l'intervallo di confidenza, su tutta la popolazione, ad un livello del

a. 95%; **b.** 99% [**a.** (1034; 1366); **b.** (981; 1419)]

74 In un campione casuale di 150 automobili, 80 sono dotate di autoradio. Stima l'intervallo di confidenza ad un livello del

a. 95%; **b.** 99%. [**a.** (0,453; 0,613); **b.** (0,428; 0,638)]

75 In riferimento all'esercizio precedente e supponendo una popolazione di 10000 automobili, determina l'intervallo di confidenza, su tutta la popolazione, ad un livello di fiducia del

a. 95%; **b.** 99%. [**a.** (4535; 6132); **b.** (4284; 6383)]

76 In una popolazione di 8000 votanti si determina che in un campione casuale di 100 elementi, 45 scelgono di astenersi dalla votazione. Stima l'intervallo di confidenza, su tutta la popolazione, ad un livello del

a. 95%; **b.** 99%. [**a.** (2820; 4380); **b.** (2573; 4627)]

LA VERIFICA DELLE IPOTESI

la teoria è a pag. 25

RICORDA

- Le regole di decisione per un test bilaterale sono
  - $|z| < z_{1-\frac{\alpha}{2}}$ si accetta H_0
  - $|z| \geq z_{1-\frac{\alpha}{2}}$ si rifiuta H_0

  (per piccoli campioni z è sostituita dalla variabile t e $z_{1-\frac{\alpha}{2}}$ da ${}_{\nu}t_{1-\alpha}$)
- Le regole di decisione per un test unilaterale sono
  - test unilaterale destro $z < z_{1-\alpha}$ non si rifiuta H_0
    $z \geq z_{1-\alpha}$ si rifiuta H_0
  - test unilaterale sinistro $z > -z_{1-\alpha}$ non si rifiuta H_0
    $z \leq -z_{1-\alpha}$ si rifiuta H_0

Applicazione

77 **ESERCIZIO GUIDA**

Una fabbrica produce bulloni di diametro medio di 1cm e scarto quadratico medio $\sigma = 0{,}05$cm. Estratto un campione di ampiezza $n = 50$, il valore medio campionario è $\overline{x} = 1{,}1$cm. Determina, ad un livello di significatività dell'1%, se la produzione è sotto controllo.

Si ha $H_0 : \mu_0 = 1$ e $H_1 : \mu_0 \neq 1$

Calcoliamo la variabile standardizzata

$$|z| = \left| \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| \begin{cases} < 2{,}58 \text{ non si rifiuta } H_0 \\ \geq 2{,}58 \text{ si rifiuta } H_0 \end{cases} \qquad z = \frac{1{,}1 - 1}{\frac{0{,}05}{\sqrt{50}}} = 14{,}14$$

Tenendo presente la regola di decisione: se $|z| < z_{1-\frac{\alpha}{2}}$ non si rifiuta H_0

se $|z| \geq z_{1-\frac{\alpha}{2}}$ si rifiuta H_0

poiché $z_{1-\frac{\alpha}{2}} = 2{,}58$ e nel nostro caso si ha che $14{,}14 > 2{,}58$, dobbiamo rifiutare H_0; questo significa che, ad un livello di significatività dell'1%, la produzione non è più sotto controllo. L'errore che può essere commesso, in tale caso, ha una probabilità dell'1%.

È possibile infatti che il campione considerato abbia una probabilità dell'1% di attuarsi. In tale caso, l'errore eventuale è di prima specie, a causa del rifiuto di una ipotesi vera.

78 Una azienda produce lampadine aventi una durata media di 1430 ore con scarto quadratico medio di 100 ore. Da un campione casualmente estratto, costituito da 80 elementi si osserva che la durata media si attesta sulle 1490 ore. Verifica, ad un livello di significatività dello 0,01 l'ipotesi che la produzione sia sotto controllo, cioè che $H_0 : \mu_0 = 1430$ ore contro l'ipotesi alternativa $H_1 : \mu \neq 1430$ ore.

[si rifiuta H_0]

79 Si rileva che la resistenza media, di una fune, alla rottura è di 1500N con scarto quadratico medio di 80N. L'azienda produttrice osserva che in un campione casuale di 100 elementi, la resistenza media è di 1540N. Verifica, ad un livello di significatività dello 0,05, l'ipotesi che la produzione sia sotto controllo, cioè che $H_0 : \mu_0 = 1500$N contro l'ipotesi alternativa $H_1 : \mu \neq 1500$N. [si rifiuta H_0]

80 La durata di un apparecchio televisivo di marca *"A"* è di 20 anni con scarto quadratico medio di 5 anni. Da un campione casuale di 60 elementi si osserva che la durata media si attesta sui 18 anni. Verifica, ad un livello di significatività dello 0,01, l'ipotesi che la produzione sia sotto controllo, cioè che $H_0 : \mu_0 = 20$ anni contro l'ipotesi alternativa $H_1 : \mu \neq 20$ anni. Si perviene ad analogo risultato se $\mu_0 = 19$ anni? [si rifiuta H_0; non si rifiuta H_0]

81 Un quotidiano ha riportato la notizia che il salario medio mensile di un dirigente di alto livello è di € 6600. Un campione casuale di 150 persone che esercitano quella professione ha un salario medio di € 6730 con scarto quadratico medio di € 900. Al livello di significatività del 5%, c'è un incremento? E al livello di significatività dell'1%? Motiva la risposta. [accetto H_0; accetto H_0]

82 Il tempo medio impiegato dagli studenti per raggiungere la scuola è di 55 minuti. Scelto un campione di 10 alunni si rileva che il tempo medio impiegato si attesta sui 50 minuti con scarto quadratico medio di 10 minuti. Si può affermare, ad un livello di significatività dello 0,05, che il tempo medio è 55 minuti? (Suggerimento: si tratta di piccoli campioni; essendo $\alpha = 0{,}05$, calcola ${}_9t_{0,95} = ...$) [non si rifiuta H_0]

83 In una azienda il numero medio di ore di lavoro straordinario svolte settimanalmente da un dipendente è 5,5. In un campione casualmente estratto, costituito da 20 lavoratori, si osserva che il numero medio di ore di straordinario è 6 con scarto quadratico medio di 0,5 ore. Verifica, ad un livello di significatività

dello 0,01, l'ipotesi che il fenomeno sia sotto controllo, cioè che $H_0 : \mu_0 = 5{,}5$ ore contro l'ipotesi alternativa $H_1 : \mu \neq 5{,}5$ ore. [si rifiuta H_0]

84 **ESERCIZIO GUIDA**

Una casa editrice afferma che la propria rivista *"A"* è letta dal 22% dei giovani. In un campione di 200 giovani si rileva che il 25% legge la rivista *"A"*. Verifichiamo, ad un livello di significatività dello 0,05, se la diffusione della rivista è quella asserita dalla casa editrice.

Consideriamo l'ipotesi nulla $H_0 : p_0 = 0{,}22$

Essendo $|z| = \left| \dfrac{f - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} \right| = 1{,}024 < 1{,}96$ non si rifiuta l'ipotesi H_0.

85 Una casa farmaceutica afferma che il farmaco *A* è efficace almeno nel 75% dei casi trattati. In un campione di 200 degenti si rileva che il 90% ha tratto beneficio dal farmaco *A*. Verifica, al livello di significatività dello 0,01, se l'efficacia del farmaco è quella asserita dalla casa farmaceutica. [si rifiuta H_0]

86 La percentuale di studenti che hanno un lettore CD è del 65%. Scelto un campione casuale di 625 studenti si rileva che il 62% lo possiede. Verifica, ad un livello di significatività dello 0,05, se l'ipotesi nulla $H_0 : p_0 = 0{,}65$ è accettabile essendo l'ipotesi alternativa $H_0 : p > 0{,}65$ [non si rifiuta H_0]

87 Si lanciano 260 volte due dadi e si rileva che la somma 5 o 10 si è verificata 40 volte su 260. Verificare al livello del 5% che i due dadi non siano truccati sia con un test a una coda che con test a due code. [rifiuto H_0; accetto H_0]

88 Un'azienda farmaceutica dichiara che dai controlli effettuati il 42% dei pazienti che usa un certo medicinale ha avuto benefici. Si estrae un campione casuale di 2 500 pazienti e si verifica che 1 200 di essi hanno avuto benefici dopo il trattamento con il farmaco. Sottoponi a verifica l'ipotesi dell'azienda farmaceutica ad un livello di significatività dello 0,05. [si accetta H_0]

89 Un'azienda di giocattoli dichiara che almeno il 95% dei suoi prodotti rispetta le norme CEE di fabbricazione. Esaminando un campione di 200 giocattoli se ne trovano 18 che non rientrano nei parametri di qualità stabiliti. Sottoponi e verifica la dichiarazione dell'azienda a un livello di significatività pari a 0,01 e 0,05. [in entrambi i casi si rifiuta]

Per la verifica delle competenze

1 Da un lotto di 1000 lavastoviglie è stato estratto un campione per verificare il numero di lavaggi che mediamente si possono fare senza che l'elettrodomestico si guasti. I risultati hanno dato un numero medio di 509 lavaggi. Da esperimenti fatti precedentemente su altri lotti si sa che lo scarto quadratico medio della popolazione è di 62 lavaggi.
Dai una stima puntuale del numero medio di lavaggi di tutte le lavastoviglie del lotto e valuta l'errore di campionamento nel caso in cui il campione è di:
a. 15 lavastoviglie **b.** 50 lavastoviglie. [509; 15,89; 8,55]

2 Da un lotto di 1000 viti se ne estraggono 80 e si trova che hanno una lunghezza media di 3,52 cm, con scarto quadratico medio di 0,02cm. Da un altro lotto di 2000 viti se ne estraggono 150 e si trova che la

lunghezza media è di 3,49cm con uno scarto quadratico medio di 0,015cm. Dai una stima puntuale della differenza media delle lunghezze delle viti dei due lotti e dell'errore di campionamento.

[0,03cm; 0,00245]

3 Un campione di 400 sbarrette di acciaio ha dato una misura media della lunghezza di 25,2cm e si sa che la varianza dell'intera popolazione è $\sigma^2 = 4\text{cm}^2$. Nell'ipotesi in cui l'estrazione del campione sia di tipo bernoulliano, calcola l'intervallo di fiducia della lunghezza media delle sbarrette nei seguenti casi:

a. livello di fiducia dell'80%; [(25,072; 25,328)]

b. livello di fiducia del 95%; [(25,004; 25,396)]

c. livello di fiducia del 99%. [(24,942; 25,458)]

Per quale livello di fiducia risulta $24{,}995 < \mu < 25{,}405$? [95,98%]

(Suggerimento: per rispondere all'ultima domanda devi usare l'interpolazione lineare)

4 Utilizzando la tavola in cui sono calcolati i valori critici della distribuzione di Student trova:

${}_8t_{0,98}$ ${}_{12}t_{0,90}$ ${}_{17}t_{0,99}$ ${}_6t_{0,80}$

5 Da una produzione di uova pasquali, che si suppone si distribuisca normalmente, è stato estratto bernoullianamente un campione di 16 elementi e si è trovato un peso medio di 345g con scarto quadratico medio di 5g. Determina l'intervallo di fiducia della media della produzione a un livello di fiducia:

a. del 95%; **b.** dell'80%. [**a.** (342,25; 347,75); **b.** (343,27; 346,73)]

Risultati di alcuni esercizi.

1 a. V, **b.** F, **c.** V, **d.** F **2 a.** **3 a.** V, **b.** F, **c.** V **4 a.** 2,7%; **b.** 2,5%; **c.** 23,8%; **d.** 0,5%

16 d. **17 b.** **18 a.** **19 b.**

20 a. **21 a.**, **c.** **36 a.** ①, **b.** ② **37 a.** ③, **b.** ①

Test finale di autovalutazione

1 In relazione ai costi che i 4000 dipendenti di un'azienda sostengono mensilmente per recarsi sul posto di lavoro, si osserva che in un campione casuale di 100 elementi la spesa media è di € 200 con uno scarto quadratico medio di € 35. Determina una stima puntuale della spesa media sostenuta dai dipendenti e il relativo scarto quadratico medio.

15 punti

2 Da un lotto di 5000 confezioni di deodorante poste in vendita in un grande magazzino, si preleva un campione casuale di 300 elementi e si osserva che 225 di essi sono di una certa marca. Determina una stima puntuale della percentuale di deodoranti di quella marca presenti nel campione e l'errore medio di campionamento.

15 punti

3 In una popolazione di 8000 votanti alle elezioni di un piccolo Comune viene intervistato un campione di 100 persone e 45 di esse dichiarano che voteranno per il candidato A. Stima l'intervallo di confidenza sull'intera popolazione ad un livello di fiducia del 95%.

20 punti

4 Su un campione casuale di 20 negozi di abbigliamento, la media mensile di jeans venduti è risultata di 180 capi con una deviazione standard pari a 48. Costruisci un intervallo di confidenza di livello 90% per la vera media mensile di jeans venduti da un negozio.

20 punti

5 Un lotto di 5000 pezzi prodotti da un'azienda deve essere spedito al cliente; nel contratto è stabilita la possibilità di rifiuto del lotto se contiene un numero di pezzi difettosi maggiori o uguali al 5%. Il cliente estrae un campione di 100 pezzi e ne trova 9 difettosi. Stabilisci se, a livello di significatività del 5%, il lotto può essere rifiutato.

20 punti

Esercizio	1	2	3	4	5	Totale
Punteggio						

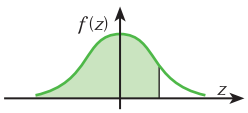
Voto: $\frac{\text{totale}}{10} + 1 =$

Soluzioni

1 € 200; € 3,50 **2** 0,75; 0,025 **3** 2820; 4380

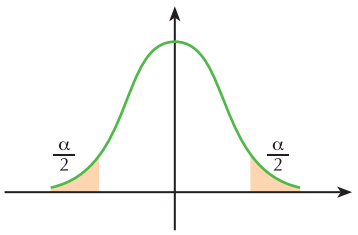
4 161; 199 **5** no

Gaussiana standardizzata - Valori di $F(z)$



z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,500000	0,503989	0,507978	0,511967	0,515953	0,519939	0,523922	0,527903	0,531881	0,535856
0,1	0,539828	0,543795	0,547758	0,551717	0,555670	0,559618	0,563559	0,567495	0,571424	0,575345
0,2	0,579260	0,583166	0,587064	0,590954	0,594835	0,598706	0,602568	0,606420	0,610261	0,614092
0,3	0,617911	0,621719	0,625516	0,629300	0,633072	0,636831	0,640576	0,644309	0,648027	0,651732
0,4	0,655422	0,659097	0,662757	0,666402	0,670031	0,673645	0,677242	0,680822	0,684386	0,687933
0,5	0,691462	0,694974	0,698468	0,701944	0,705402	0,708840	0,712260	0,715661	0,719043	0,722405
0,6	0,725747	0,729069	0,732371	0,735653	0,738914	0,742154	0,745373	0,748571	0,751748	0,754903
0,7	0,758036	0,761148	0,764238	0,767305	0,770350	0,773373	0,776373	0,779350	0,782305	0,785236
0,8	0,788145	0,791030	0,793892	0,796731	0,799546	0,802338	0,805106	0,807850	0,810570	0,813267
0,9	0,815940	0,818589	0,821214	0,823814	0,826391	0,828944	0,831472	0,833977	0,836457	0,838913
1	0,841345	0,843752	0,846136	0,848495	0,850830	0,853141	0,855428	0,857690	0,859929	0,862143
1,1	0,864334	0,866500	0,868643	0,870762	0,872857	0,874928	0,876976	0,878999	0,881000	0,882977
1,2	0,884930	0,886860	0,888767	0,890651	0,892512	0,894350	0,896165	0,897958	0,899727	0,901475
1,3	0,903199	0,904902	0,906582	0,908241	0,909877	0,911492	0,913085	0,914656	0,916207	0,917736
1,4	0,919243	0,920730	0,922196	0,923641	0,925066	0,926471	0,927855	0,929219	0,930563	0,931888
1,5	0,933193	0,934478	0,935744	0,936992	0,938220	0,939429	0,940620	0,941792	0,942947	0,944083
1,6	0,945201	0,946301	0,947384	0,948449	0,949497	0,950529	0,951543	0,952540	0,953521	0,954486
1,7	0,955435	0,956367	0,957284	0,958185	0,959071	0,959941	0,960796	0,961636	0,962462	0,963273
1,8	0,964070	0,964852	0,965621	0,966375	0,967116	0,967843	0,968557	0,969258	0,969946	0,970621
1,9	0,971284	0,971933	0,972571	0,973197	0,973810	0,974412	0,975002	0,975581	0,976148	0,976705
2	0,977250	0,977784	0,978308	0,978822	0,979325	0,979818	0,980301	0,980774	0,981237	0,981691
2,1	0,982136	0,982571	0,982997	0,983414	0,983823	0,984222	0,984614	0,984997	0,985371	0,985738
2,2	0,986097	0,986447	0,986791	0,987126	0,987455	0,987776	0,988089	0,988396	0,988696	0,988989
2,3	0,989276	0,989556	0,989830	0,990097	0,990358	0,990613	0,990863	0,991106	0,991344	0,991576
2,4	0,991802	0,992024	0,992240	0,992451	0,992656	0,992857	0,993053	0,993244	0,993431	0,993613
2,5	0,993790	0,993963	0,994132	0,994297	0,994457	0,994614	0,994766	0,994915	0,995060	0,995201
2,6	0,995339	0,995473	0,995603	0,995731	0,995855	0,995975	0,996093	0,996207	0,996319	0,996427
2,7	0,996533	0,996636	0,996736	0,996833	0,996928	0,997020	0,997110	0,997197	0,997282	0,997365
2,8	0,997445	0,997523	0,997599	0,997673	0,997744	0,997814	0,997882	0,997948	0,998012	0,998074
2,9	0,998134	0,998193	0,998250	0,998305	0,998359	0,998411	0,998462	0,998511	0,998559	0,998605
3	0,998650	0,998694	0,998736	0,998777	0,998817	0,998856	0,998893	0,998930	0,998965	0,998999
3,1	0,999032	0,999064	0,999096	0,999126	0,999155	0,999184	0,999211	0,999238	0,999264	0,999289
3,2	0,999313	0,999336	0,999359	0,999381	0,999402	0,999423	0,999443	0,999462	0,999481	0,999499
3,3	0,999517	0,999533	0,999550	0,999566	0,999581	0,999596	0,999610	0,999624	0,999638	0,999650
3,4	0,999663	0,999675	0,999687	0,999698	0,999709	0,999720	0,999730	0,999740	0,999749	0,999758
3,5	0,999767	0,999776	0,999784	0,999792	0,999800	0,999807	0,999815	0,999821	0,999828	0,999835
3,6	0,999841	0,999847	0,999853	0,999858	0,999864	0,999869	0,999874	0,999879	0,999883	0,999888
3,7	0,999892	0,999896	0,999900	0,999904	0,999908	0,999912	0,999915	0,999918	0,999922	0,999925
3,8	0,999928	0,999930	0,999933	0,999936	0,999938	0,999941	0,999943	0,999946	0,999948	0,999950
3,9	0,999952	0,999954	0,999956	0,999958	0,999959	0,999961	0,999963	0,999964	0,999966	0,999967

Distribuzione t di Student



$$P(-_{\nu}t_{1-\frac{\alpha}{2}} < t <_{\nu} t_{1-\frac{\alpha}{2}}) = 1-\alpha$$

Gradi di libertà ν \ $1-\alpha$	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,98	0,99	0,999
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291