1. STATISTICA INFERENZIALE

1.1 Spazi campionari

DEFINIZIONE (Spazio Campionario)

(*Spazio campionario*, indicato con S) : ⇔ (Insieme contenente tutti i possibili risultati di un dato esperimento)

Uno spazio campionario si dice *discreto* quando è finito o numerabile, *continuo* in caso contrario.

ESEMPI

1) Nel lancio di un dado S = {1,2,3,4,5,6}. Spazio campionario discreto.
2) Nel lancio di una moneta S = {T,C}. Spazio campionario discreto.
3) Nel lancio di due monete S = {TT, TC, CT, CC}. Spazio campionario discreto.
4) Consideriamo un esperimento che consiste nel determinare la durata media di un lotto di lampadine. Allora S = $\{t \in \mathbf{R} \setminus t \geq 0\}$. Spazio campionari continuo.
5) Supponiamo di voler studiare il peso di un gruppo di ragazzi aventi un peso che oscilla tra i 52*kg* e 83*kg*. Allora S = $\{p \in \mathbf{R} \setminus 52 \leq p \leq 83\}$ e lo spazio campionario è continuo.

DEFINIZIONE (Evento)

(*Evento*) : ⇔ (Sottoinsieme dello spazio campionario)

In generale un evento è definito da una proposizione aperta.
Un evento formato da un solo elemento si chiama *evento elementare*; se è uguale all'insieme vuoto si dice *evento impossibile*.
Se E è un evento, indichiamo con $\overline{E}$ l'evento contrario, ossia poniamo $\overline{E} = S - E$.

ESEMPI

1) Nel caso del lancio di un dado, se consideriamo la proposizione *e* = "uscita di un numero pari", allora l'evento è E = {2,4,6}e l'evento contrario $\overline{E}$ ={1,3,5}. L'evento corrispondente alla proposizione *e'* = "uscita di un numero maggiore di 6" è l'evento impossibile, mentre se la proposizione è *e"* = "uscita di una faccia recante il numero 1", l'evento è E" = {1}, evento elementare.
2) Nel caso del lancio di due monete, consideriamo la proposizione *e* = "non uscita di due teste", allora l'evento corrispondente è E = {CC,CT,TC}.
3) Lo spazio campionario nel caso di un lancio di due dadi è
   S = {(1,1),(1,2),...,(1,6),(2,1),(2,2),...,(6,6)}.
   Sia *e* = "la somma delle facce è un numero multiplo di 3", allora l'evento corrispondente è
   E = {(1,2),(2,1),(1,5),(5,1),(3,3),(4,2),(2,4),(6,3),(3,6),(5,4),(4,5),(6,6)}.
4) Nel caso in cui si vuole studiare la durata *t* di una pila appartenente a un dato lotto, lo spazio campionario è S = $\{t \in \mathbf{R} \setminus t \geq 0\}$. Considerata *e* = "la durata non supera le 10 ore", l'evento corrispondente è E = $\{t \in R \setminus 0 \leq t \leq 10\}$.

APPLICHIAMO

1) Determinate lo spazio campionario relativo al lancio di tre monete. Trovate quindi gli eventi corrispondenti alle proposizioni e_1 = "uscita di almeno due teste", e_2 = "uscita di non più di due croci" ed e_3 = "non uscita di alcuna testa".
2) Considerate un'urna contenente quattro caramelle, una di limone (L), una al gusto di fragola (F), un'altra al caffè (C) e la quarta al gusto di pesca (P). Determinate lo spazio campionario relativo all'estrazione di due caramelle nel caso in cui l'estrazione avvenga con restituzione. Trovate quindi gli eventi corrispondenti alle proposizioni e_1 = "estrazione di almeno una caramella al caffè", e_2 = "non uscita di alcuna caramella alla frutta", e_3 = "uscita di due caramelle a limone".
3) Ripetete l'esercizio precedente nel caso in cui l'estrazione avvenga senza restituzione o, come si dice, in blocco.
4) Determinate lo spazio campionario relativo al tempo medio trascorso tra due telefonate tra le ore 8.00 e le 13.00. Trovate quindi gli eventi corrispondenti alle proposizioni e_1 = "non c'è alcuna telefonata", e_2 = “una telefonata avviene dopo almeno due minuti” ed e_3 = "una chiamata avviene dopo 3 minuti massimo".

1.2 Probabilità e frequenza

Dato un evento E, possiamo considerare la probabilità P(E) che si verifichi. Essa soddisfa alle seguenti proprietà:
1) $P(E) \geq 0$,
2) $P(S) = 1$,
3) Se E_1 ed E_2 sono due eventi tali che $E_1 \cap E_2 = \varnothing$, allora $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Si dimostra che:

a) $P(\varnothing) = 0$,
b) $P(\overline{E}) = 1 - P(E)$
c) $0 \leq P(E) \leq 1$,
d) $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.

Sappiamo inoltre che nel caso in cui lo spazio campionario è finito e se |S| ed |E| indicano il numero di elementi di S ed E rispettivamente, allora $P(E) = \frac{|E|}{|S|}$.

ESEMPI

1) Consideriamo il caso del lancio di tre monete. Lo spazio campionario è
   S = {TTT,CTT,TCT,TTC,CCT,CTC,TCC,CCC} e $|S| = 8$.
   Se prendiamo in considerazione la proposizione *e* = “uscita di almeno due teste” allora
   E = {CTT,TCT,TTC,TTT} ed $|E| = 4$. Pertanto $P(E) = \frac{4}{8} = \frac{1}{2}$.
2) Consideriamo un’urna contenente cinque palline recanti i numeri 4, 7, 12, 15 e 20. Supponiamo di estrarre due palline in blocco. Lo spazio campionario è
   S = {{4,7},{4,12},{4,15},{4,20},{7,12},{7,15},{7,20},{12,15},{12,20},{15,20}}
   ed $|S| = \binom{5}{2} = 10$.

L'evento corrispondente alla proposizione *e* = "la somma dei numeri che individuano le palline è un numero divisibile per 3" è E = {{4,12},{7,20},{12,15}} e P(E) = $\frac{3}{10}$.

3) Consideriamo un'urna contenente tre palline nere (N) e due rosse (R). Nel caso di estrazione in blocco di due palline (combinazioni semplici) lo spazio campionario è
S = {{N_1,N_2},{N_1,N_3},{N_1,R_1},{N_1,R_2},{N_2,N_3},{N_2,R_1},{N_2,R_2},{N_3,R_1},{N_3,R_2},{R_1,R_2}} ed $|S| = \binom{5}{2} = 10$. Sia *e* la proposizione "estrazione di due palline dello stesso colore". L'evento è quindi
E = {{N_1,N_2},{N_1,N_3},{N_2,N_3},{R_1,R_2}} ed $|E| = \binom{3}{2} + \binom{2}{2} = 3 + 1 = 4$ e P(E) = $\frac{4}{10} = \frac{2}{5}$.

4) Consideriamo l'urna dell'esempio precedente. Nel caso di estrazione con restituzione di due palline e con conteggio delle coppie non considerando l'ordine di uscita (combinazioni con ripetizione), lo spazio campionario è
S = {{N_1,N_1},{N_1,N_2},{N_1,N_3},{N_1,R_1},{N_1,R_2},{N_2,N_2},{N_2,N_3},{N_2,R_1},{N_2,R_2},{N_3,N_3},
{N_3,R_1},{N_3,R_2},{R_1,R_1},{R_1,R_2},{R_2,R_2}} ed $|S| = \binom{5+2-1}{2} = 15$.
Sia *e* la proposizione "estrazione di due palline dello stesso colore". L'evento è quindi
E = {{N_1,N_1},{N_1,N_2},{N_1,N_3},{N_2,N_2},{N_2,N_3},{N_3,N_3},{R_1,R_1},{R_1,R_2},{R_2,R_2}} ed $|E| = \binom{3+2-1}{2} + \binom{2+2-1}{2} = 6 + 3 = 9$ e P(E) = $\frac{9}{15} = \frac{3}{5}$.

5) Consideriamo l'urna dell'esempio 3), ed effettuiamo estrazioni di due palline. Prendiamo una pallina, annotiamone il colore e, senza rimetterla nell'urna, prendiamo una seconda pallina. In tale caso conteggiamo quante coppie possiamo formare nel caso in cui gli elementi di ciascuna coppia sono diversi (disposizioni semplici). Lo spazio campionario è
S = {(N_1,N_2),(N_1,N_3),(N_1,R_1),(N_1,R_2),(N_2,N_1),(N_2,N_3),(N_2,R_1),(N_2,R_2),(N_3,N_1),(N_3,N_2),(N_3,R_1),
(N_3,R_2),(R_1,N_1),(R_1,N_2),(R_1,N_3),(R_1,R_2),(R_2,N_1),(R_2,N_2),(R_2,N_3),(R_2,R_1)} ed $|S| = 5 \cdot 4 = 20$.
L'evento corrispondente alla proposizione "estrazione di due palline dello stesso colore" è E = {(N_1,N_2),(N_1,N_3),(N_2,N_1),(N_2,N_3),(N_3,N_1),(N_3,N_2),(R_1,R_2),(R_2,R_1)} ed
$|E| = 3 \cdot 2 + 2 \cdot 1 = 6 + 2 = 8$ e P(E) = $\frac{8}{20} = \frac{2}{5}$.

6) Nell'urna dell'esempio 3), nel caso di estrazione bernoulliana, ossia di estrazione con ripetizione e con conteggio delle coppie considerando anche l'ordine di uscita (disposizioni con ripetizione), lo spazio campionario è
S = {(N_1,N_1),(N_1,N_2),(N_1,N_3),(N_1,R_1),(N_1,R_2),(N_2,N_1),(N_2,N_2),(N_2,N_3),(N_2,R_1),(N_2,R_2),(N_3,N_1),
(N_3,N_2),(N_3,N_3),(N_3,R_1),(N_3,R_2),(R_1,N_1),(R_1,N_2),(R_1,N_3),(R_1,R_1),(R_1,R_2),(R_2,N_1),(R_2,N_2),(R_2,N_3),
(R_2,R_1),(R_2,R_2)} ed $|S| = 5^2 = 25$. L'evento corrispondente alla proposizione "estrazione di due palline dello stesso colore" è E = {(N_1,N_1),(N_1,N_2),(N_1,N_3),(N_2,N_1),(N_2,N_2),(N_2,N_3),(N_3,N_1),
(N_3,N_2),(N_3,N_3),(R_1,R_1),(R_1,R_2),(R_2,R_1),(R_2,R_2)} ed $|E| = 3^2 + 2^2 = 9 + 4 = 13$ e P(E) = $\frac{13}{25}$.

7) Consideriamo un'urna contenente dodici palline nere (N) e sedici rosse (R) e supponiamo di estrarre in blocco cinque palline. Il numero degli elementi dello spazio campionario è $|S| = \binom{28}{5} = 98.280$. L'evento E corrispondente alla proposizione *e* = "estrazione di almeno tre palline nere" è uguale all'unione degli eventi E_3, E_4 ed E_5, corrispondenti alle proposizioni e_3 =

"estrazione di tre palline nere", e_4 = "estrazione di quattro palline nere" ed e_5 = "estrazione di cinque palline nere", rispettivamente. Pertanto $|E| = |E_3| + |E_4| + |E_5|$.

Poiché $|E_3| = \binom{12}{3} \cdot \binom{16}{2}$, $|E_4| = \binom{12}{4} \cdot \binom{16}{1}$ e $|E5| = \binom{12}{5}$ segue

$|E| = 220 \cdot 120 + 495 \cdot 16 + 792$ e $P(E) = \frac{35112}{98280} \approx 0.357$.

APPLICHIAMO

1) Considerate il caso del lancio di quattro monete. Determinate
   - ◊ lo spazio campionario,
   - ◊ l'evento E corrispondente alla proposizione *e* = "uscita di massimo due croci",
   - ◊ P(E).
2) Considerate un'urna contenente sei palline recanti i numeri 1, 3, 7, 12 e 25. Supponete di estrarre tre palline in blocco. Determinate
   - ◊ lo spazio campionario,
   - ◊ l'evento E corrispondente alla proposizione *e* = "la somma dei numeri che individuano le palline è un numero divisibile per 5",
   - ◊ P(E).
3) Considerate un'urna contenente dieci palline nere (N) e tredici rosse (R). Nel caso di estrazione bernoulliana di tre palline, determinate
   - ◊ il numero degli elementi $|S|$ dello spazio campionario,
   - ◊ il numero degli elementi dell'evento E corrispondente alla proposizione *e* = "estrazione di almeno due palline rosse",
   - ◊ P(E).
4) Considerate un'urna contenente venti palline nere (N) e venticinque rosse (R). Nel caso di estrazione in blocco di sei palline, determinate
   - ◊ il numero degli elementi $|S|$ dello spazio campionario,
   - ◊ il numero degli elementi dell'evento E corrispondente alla proposizione *e* = "estrazione di massimo tre palline nere",
   - ◊ P(E).

Molte volte non è possibile determinare la probabilità *a priori*.

Per esempio non possiamo trovare la probabilità che una squadra di calcio vinca una partita o che fra un mese piova oppure che su 10 lampadine una sia difettosa.
In tali circostanze è necessario sapere cosa è successo in passato, ossia avere a disposizione dei dati statistici, valutando il rapporto fra il numero di volte in cui l'evento si è verificato, chiamato *frequenza assoluta*, sul numero totale di prove effettuate. Tale rapporto prende il nome di *frequenza relativa*.

Tra la frequenza relativa e la probabilità esiste una importante relazione stabilita dal

TEOREMA (legge dei grandi numeri)

Siano rispettivamente *p* ed *f* la probabilità che si verifichi una dato evento e la sua frequenza relativa su *n* prove. Allora $\lim_{n\to\infty} P(|f - p| < e) = 1$, con *e* numero arbitrario positivo.

ESEMPIO

1) Abbiamo già visto che nel caso del lancio di due dadi lo spazio campionario è
   S = {(1,1),(1,2),...,(1,6),(2,1),(2,2),...(2,6),...,(6,6)}.
   Se consideriamo la proposizione *e* = "uscita di due facce aventi come somma un numero maggiore di 8", l'evento corrispondente è
   E = {(3,6),(6,3),(4,5),(5,4),(5,5),(4,6),(6,4),(5,6),(6,5),(6,6)}. Pertanto segue che $P(E) = \frac{\cancel{10}\ 5}{\cancel{36}\ 18}$
2) Supponiamo di lanciare per 573 volte due dadi e di ottenere i seguenti risultati:

somma facce	2	3	4	5	6	7	8	9	10	11	12
frequenza	12	22	34	73	94	120	91	69	32	15	11

   Consideriamo le proposizioni
   e_2 = "uscita di due facce aventi come somma 2",
   e_3 = "uscita di due facce aventi come somma 3",
   e_4 = "uscita di due facce aventi come somma 4",
   .....
   e_{12} = "uscita di due facce aventi come somma 12",

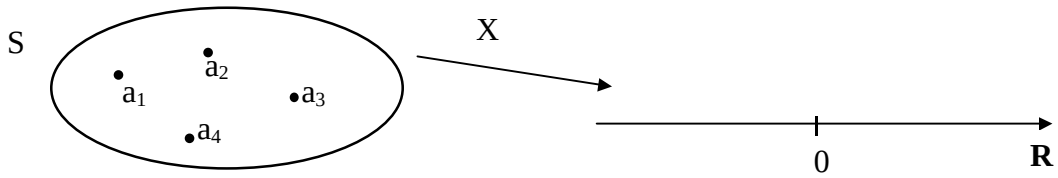
   La seguente tabella mostra la frequenza relativa di ciascun evento corrispondente alle proposizioni precedenti

somma facce	2	3	4	5	6	7	8	9	10	11	12
frequenza relativa	0.021	0.038	0.059	0.127	0.164	0.209	0.159	0.120	0.056	0.026	0.019
probabilità	$\frac{1}{36}=0.028$	$\frac{2}{36}=0.056$	$\frac{3}{36}=0.083$	$\frac{4}{36}=0.111$	$\frac{5}{36}=0.139$	$\frac{6}{36}=0.167$	$\frac{5}{36}=0.139$	$\frac{4}{36}=0.111$	$\frac{3}{36}=0.083$	$\frac{2}{36}=0.056$	$\frac{1}{36}=0.028$

1.3 Variabili casuali

DEFINIZIONE (Variabile casuale)

Sia S uno spazio campionario.
(*Variabile casuale* su S, indicata con X) :⇔ (Funzione avente per dominio S e valori in **R**, in simboli $X : S \to \mathbf{R}$)

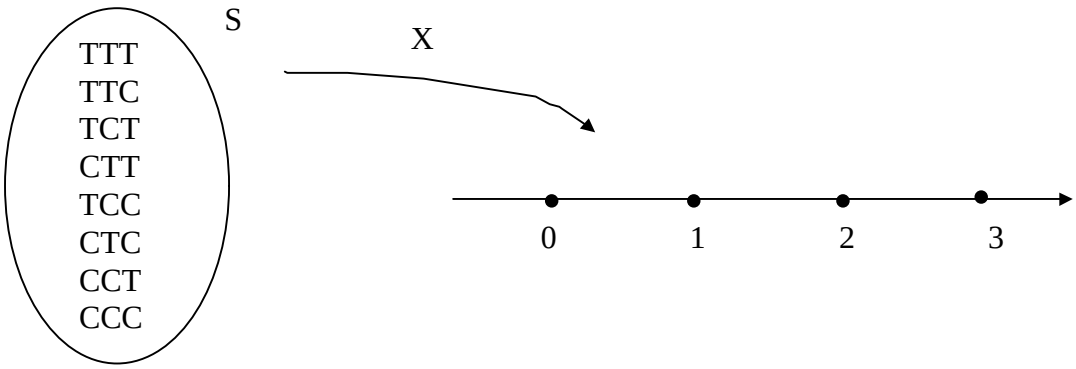


Nel caso in cui S è discreta diciamo che la variabile casuale è *discreta*, altrimenti la chiamiamo *continua*.

ESEMPIO

Nel lancio di tre monete lo spazio campionario è S = {TTT,TTC,TCT,CTT,TCC,CTC,CCT,CCC}.

Consideriamo come variabile casuale la funzione X che ad ogni elemento di S associa il numero 0, 1, 2 o 3 a seconda del numero di T che compaiono. Così abbiamo X(TTT) = 3, X(TCC) = 1 e il codominio è X(S) = {0,1,2,3}.



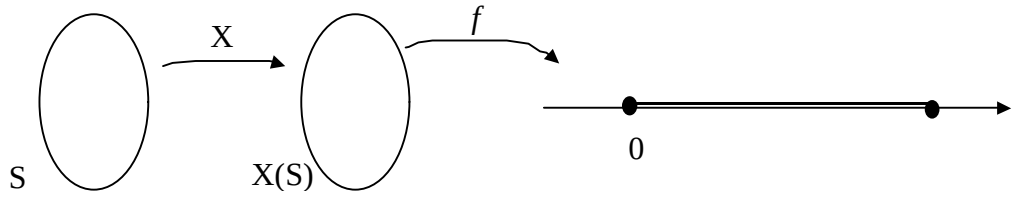
OSSERVAZIONE

L'evento $\{a \in S \setminus X(a) = k\}$ sarà indicato più brevemente con $\{X = k\}$ o con $X = k$.
Pertanto, considerando l'esempio precedente l'insieme $E = \{a \in S \setminus X(a) = 1\} = \{TCC,CTC,CCT\}$ può essere scritto con $\{X = 1\}$ o anche con $X = 1$.
In modo analogo scriveremo per esempio $\{X < k\}$ o $X < k$ invece di $\{a \in S \setminus X(a) < k\}$.

DEFINIZIONE (Funzione di distribuzione)

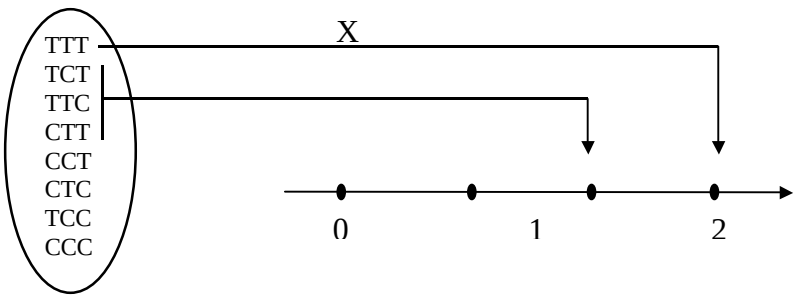
Siano S uno spazio campionario e X una variabile casuale su S.
(*Funzione di distribuzione* della variabile casuale X) : $\Leftrightarrow$ (La funzione $f : X(S) \to [0,1]$ tale che $\forall\, x \in S$, $f(x) = P(X = x)$)

dove $P(X = x)$ indica la probabilità che si verifichi l'evento $\{a \in S \setminus X(a) = x\}$.



ESEMPIO

1) Consideriamo il lancio di tre monete e la variabile casuale che associa a ogni risultato il numero di teste uscite.



Allora abbiamo:

$$P(X=3)=\frac{1}{8};P(X=2)=\frac{3}{8};P(X=1)=\frac{3}{8};P(X=0)=\frac{1}{8}$$

2) Consideriamo un'urna contenente 9 palline di cui 4 bianche (B) e 5 nere (N). Nel caso di estrazione in blocco di 3 palline lo spazio campionario è S = {$b_1b_2b_3$,$b_1b_2b_4$,..,$n_1b_1b_3$,..,$n_2n_4b_4$,..}.

Abbiamo che $|S|=\binom{9}{3}=84$.

Se X è la variabile casuale che associa ad ogni elemento di S il numero di palline bianche, abbiamo

$$P(X=3)=\frac{\binom{4}{3}}{\binom{9}{3}}=\frac{1}{21};P(X=2)=\frac{\binom{4}{2}\cdot\binom{5}{1}}{\binom{9}{3}}=\frac{5}{14};P(X=1)=\frac{\binom{4}{1}\cdot\binom{5}{2}}{\binom{9}{3}}=\frac{10}{21};P(X=0)=\frac{\binom{4}{0}\cdot\binom{5}{3}}{\binom{9}{3}}=\frac{5}{42}$$

OSSERVAZIONE

Nel caso di spazi campionari finiti la funzione di distribuzione è rappresentata mediante una tabella del tipo

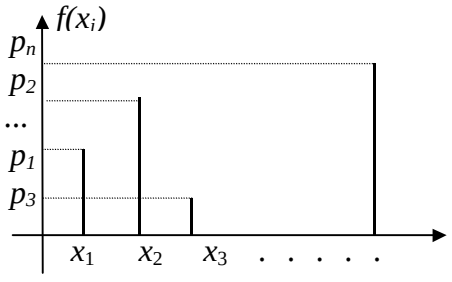
x_i	x_1	x_2	...	x_n
$f(x_i)$	p_1	p_2	...	p_n

chiamata *tabella di distribuzione*.

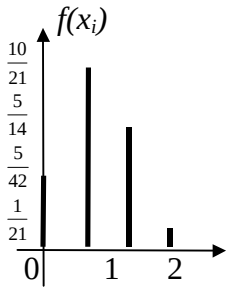
Considerando il precedente esempio 2) abbiamo

x_i	0	1	2	3
$f(x_i)$	$\frac{5}{42}$	$\frac{10}{21}$	$\frac{5}{14}$	$\frac{1}{21}$

Spesso ricorreremo a digrammi del tipo



Considerando il precedente esempio 2) abbiamo



OSSERVAZIONE (Distribuzione binomiale, delle prove ripetute o bernoulliana)

Se un esperimento viene ripetuto nelle medesime condizioni, detta *p* la probabilità che si verifichi un dato evento E e *q* la probabilità dell'evento contrario $\overline{E}$, allora la probabilità che su *n* prove l'evento E si verifichi esattamente *k* volte è $\binom{n}{k}p^k q^{n-k}$.

Possiamo quindi considerare lo spazio campionario S relativo al numero di volte che si verifica, su *n* prove, l'evento E e la variabile casuale X : S $\to$ **R**, tale che X(*a*) = *a*. Ossia

x_i	0	1	2	...	*n*
$f(x_i)$	$\binom{n}{0}p^0q^n$	$\binom{n}{1}p^1q^{n-1}$	$\binom{n}{2}p^2q^{n-2}$	...	$\binom{n}{n}p^nq^0$

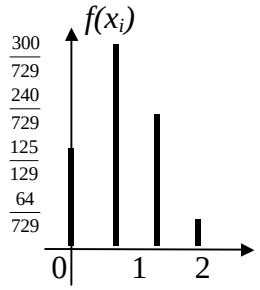
(B)

ESEMPIO

Consideriamo un'urna contenente 9 palline, 4 bianche e 5 nere ed effettuiamo una "estrazione bernoulliana" di 3 palline, ossia per tre volte estraiamo una pallina, ne annotiamo il colore e la rimettiamo nell'urna.
Lo spazio campionario è S = {$b_1b_1b_1$,$b_1b_1b_2$,.....,$b_3b_4n_2$,....,$n_5n_5n_5$} con $|S| = 9^3 = 729$.
Allora abbiamo

$$P(X=0)=\binom{3}{0}\cdot\left(\frac{4}{9}\right)^0\left(\frac{5}{9}\right)^3,\ P(X=1)=\binom{3}{1}\cdot\left(\frac{4}{9}\right)^1\cdot\left(\frac{5}{9}\right)^2,\ P(X=2)=\binom{3}{2}\cdot\left(\frac{4}{9}\right)^2\cdot\left(\frac{5}{9}\right)^1,\ P(X=3)=\binom{3}{3}\cdot\left(\frac{4}{9}\right)^0\cdot\left(\frac{5}{9}\right)^0$$

x_i	0	1	2	3
$f(x_i)$	$\frac{125}{729}$	$\frac{300}{729}$	$\frac{240}{729}$	$\frac{64}{729}$



APPLICHIAMO

1) Costruite la tabella ed il diagramma della funzione di distribuzione relativa al lancio per 10 volte di una moneta. Considerate la variabile casuale X che associa ad ogni elemento di S il numero di volte che esce testa.
2) Costruite la tabella ed il diagramma della funzione di distribuzione relativa all'estrazione senza restituzione di 3 palline da un'urna che ne contiene 2 bianche e 3 nere. Considerate la variabile casuale X che associa ad ogni elemento di S il numero di palline nere estratte.
3) Costruite la tabella ed il diagramma della funzione di distribuzione relativa all'estrazione bernoulliana di 3 palline che ne contiene 2 bianche e 3 nere. Considerate la variabile casuale X che associa ad ogni elemento di S il numero di palline nere estratte.

OSSERVAZIONE

Data una tabella di una funzione di distribuzione

x_i	x_1	x_2	...	x_n
$f(x_i)$	p_1	p_2	...	p_n

(1), facciamo osservare che $\sum_{i=1}^{n} p_i = 1$.

1.4 Valore medio e varianza

DEFINIZIONE (Valore medio, varianza, scarto quadratico medio)

Date una variabile casuale X sullo spazio campionario S e una funzione di distribuzione

f : X(S) $\rightarrow$ [0,1] avente come tabella di distribuzione **(1)**,

(*Valore medio* o *media* di X, indicata con M(X) o con m) : $\Leftrightarrow$ (Il valore numerico $x_1 p_1 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i$)

(*Varianza* di X, indicata con V(X) o con s^2) :$\Leftrightarrow$ (Il valore numerico $\left[x_1 - \mathrm{M(X)}\right]^2 p_1 + \ldots + \left[x_n - \mathrm{M(X)}\right]^2 p_n = \sum_{i=1}^{n} \left[x_i - \mathrm{M(X)}\right]^2 p_i$)

(Scarto quadratico medio, indicato con s) :$\Leftrightarrow$ ($\sqrt{\mathrm{V(X)}}$)

OSSERVAZIONE

Risulta chiaramente dalle precedenti definizioni che il valore medio ha lo scopo di "riassumere" in un unico numero alcune grandezze numeriche. Un numero avente la proprietà di sintetizzare un insieme di dati statistici di carattere quantitativo si chiama *valore di sintesi*. Esistono altri valori di sintesi come la media geometrica, armonica, la mediana, la moda, ecc.
La varianza e ancor meglio lo scarto quadratico medio sono *indici di dispersione* o *indici di variabilità*. Per una più completa analisi di una distribuzione statistica, non è sufficiente conoscere i soli valori di sintesi, tali indici di dispersione indicano quindi come i dati sono distribuiti rispetto al valore di sintesi preso in esame.

ESEMPIO

Considerata la seguente tabella di distribuzione

x_i	3	5	6	10
$f(x_i)$	$\frac{1}{3}$	$\frac{3}{7}$	$\frac{2}{9}$	$\frac{1}{63}$

risulta

$$\mathrm{M(X)} = 3\cdot\frac{1}{3} + 5\cdot\frac{3}{7} + 6\cdot\frac{2}{9} + 10\cdot\frac{1}{63} = \frac{292}{63} \approx 4.63,$$

$$\mathrm{V(X)} = \left(3 - \frac{292}{63}\right)^2 \cdot \frac{1}{3} + \left(5 - \frac{292}{63}\right)^2 \cdot \frac{3}{7} + \left(6 - \frac{292}{63}\right)^2 \cdot \frac{2}{9} + \left(10 - \frac{292}{63}\right)^2 \cdot \frac{1}{63} \approx 1.82$$

e $s \approx 1.35$.

APPLICHIAMO

1) Dopo aver determinato lo spazio campionario S relativo al lancio di 4 monete, considerate la variabile casuale X che associa ad ogni elemento di S il numero di volte che esce croce. Quindi, dopo aver costruito la tabella di distribuzione, determinate M(X), V(X) e s.
2) Considerate un gioco che consiste nell'estrarre una carta napoletana. Con esso si vince L.10.000 se viene estratto una carta di denari, L.1.000 se viene estratta una figura di coppe, si perde l.500 nei casi restanti. Da quali elementi è costituito lo spazio campionario. Sia X la variabile casuale che associa ad ogni elemento di S la vincita o la perdita corrispondente. Dopo aver costituito la tabella di distribuzione, determinate M(X), V(X), e s.
3) All'interno di un'urna ci sono 5 palline numerate dall'1 al 5. Dopo aver determinato lo spazio campionario S relativo all'estrazione in blocco di due palline, considerate la variabile casuale X

che associa ad ogni elemento di S la somma dei numeri. Costruite la tabella di distribuzione e determinate M(X), V(X) e s.

OSSERVAZIONE

Nel caso di una distribuzione binomiale avente la tabella di distribuzione **(B)** si dimostra che
M(X) = $n \cdot p$ e V(X) = $n \cdot p \cdot q = n \cdot p \cdot (1 - p)$.

OSSERVAZIONI

Il valore medio e la varianza soddisfano le seguenti proprietà:

A) M(aX) = a·M(X), $\forall a \in \mathbf{R}$,
B) M(aX + b) = a·M(X) + b, $\forall a, b \in \mathbf{R}$,
C) M(X - M(X)) = 0,

D) V(X) = M(X^2) - [M(X)]^2,
E) V(aX) = a^2·V(X), $\forall a \in \mathbf{R}$,
F) V(aX + b) = a^2·V(X), $\forall a, b \in \mathbf{R}$,
G) Se f e g sono funzioni di distribuzioni delle variabili casuali X ed Y rispettivamente

x_i	x_1	x_2	...	x_n
$f(x_i)$	p_1	p_2	...	p_n

y_i	y_1	y_2	...	y_m
$g(y_i)$	q_1	q_2	...	q_m

sia la funzione di distribuzione h della variabile casuale X + Y avente la seguente tabella di distribuzione

$x_i + y_j$	$x_1 + y_1$	$x_1 + y_2$	...	$x_1 + y_m$	...	$x_n + y_1$	...	$x_n + y_m$
$P(X = x_i, Y = y_j) = h(x_i + y_j)$	$p_1 \cdot q_1$	$p_1 \cdot q_2$	...	$p_1 \cdot q_m$	...	$p_n \cdot q_1$	...	$p_n \cdot q_m$

.

Allora M(X + Y) = M(X) + M(Y) e V(X ± Y) = V(X) + V(Y).

DEFINIZIONE (Variabile casuale standardizzata)

Data una variabile casuale X di valore medio m e scarto quadratico medio s

(*Variabile casuale standardizzata* associata ad X) :⇔ (La variabile casuale Z = $\frac{X - m}{s}$)

TEOREMA (proprietà di Z)

Se Z è la variabile casuale standardizzata allora M(Z) = 0 e V(Z) = 1.

Dimostrazione

Infatti M(Z) = M($\frac{X - m}{s}$) = M($\frac{X}{s} - \frac{m}{s}$). Per la proprietà B) del valore medio segue che

$$M(\frac{X}{s}-\frac{m}{s})=\frac{1}{s}M(X)-\frac{m}{s}=\frac{m}{s}-\frac{m}{s}=0\,.$$

Inoltre $V(Z) = V(\frac{X-m}{s}) = V(\frac{X}{s}-\frac{m}{s})$. Per la proprietà F) della varianza risulta

$$V(\frac{X}{s}-\frac{m}{s})=\frac{1}{s^2}V(X)=\frac{1}{s^2}\cdot s^2=1\,.$$

c.v.d.

1.5 Funzione di ripartizione

DEFINIZIONE (Funzione di ripartizione)

Siano S uno spazio campionario e X una variabile casuale su S.

(*Funzione di ripartizione* della variabile casuale X) :⇔ (La funzione F : **R** → [0,1] tale che $\forall x \in S,\ F(x) = P(X \le x)$)

ESEMPIO

Consideriamo il caso del lancio di due monete e della variabile casuale che associa ad ogni elemento di S il numero di teste. La tabella della distribuzione è

x_i	0	1	2
$f(x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Pertanto $F(0) = P(X \le 0) = \frac{1}{4}$, $F(1) = P(X \le 1) = \frac{1}{4}+\frac{1}{2}$, $F(2) = P(X \le 2) = \frac{1}{4}+\frac{1}{2}+\frac{1}{4}$.

Possiamo quindi completare la tabella precedente con una riga dedicata alla funzione di ripartizione.

x_i	0	1	2
$f(x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$F(x_i)$	$\frac{1}{4}$	$\frac{3}{4}$	1

OSSERVAZIONE

Poiché la funzione di ripartizione è definita su **R** ha senso considerare oltre ad $F(x_i)$, con $x_i \in X(S)$, anche $F(x)$, con x generico numero reale.

Sia X(S) = $\{x_1, x_2, \ldots, x_n\}$, con x_1 ed x_n sono minimo e massimo valore di X(S). Allora

◊ se $x < x_1$, $F(x) = 0$,
◊ se $x_i \le x < x_{i+1}$, $F(x) = p_1 + p_2 + \ldots + p_i$,
◊ se $x \ge x_n$, $F(x) = 1$.

Sia $a \le b$, allora possiamo scrivere $P(\{X \le b\}) = P(\{X \le a\} \cup \{a < X \le b\})$. Per la proprietà 3) della probabilità segue $P(X \le b) = P(X \le a) + P(a < X \le b)$, da cui $P(a < X \le b) = P(X \le b) - P(X \le a)$, ossia

$$P(a < X \le b) = F(b) - F(a) \quad \textbf{(2)}$$

Inoltre, poiché $P(\{X \le a\} \cup \{X > a\}) = 1$, abbiamo $P(\{X \le a\}) + P(\{X > a\}) = 1$,
ossia $P(\{X > a\}) = 1 - P(\{X \le a\})$ e quindi

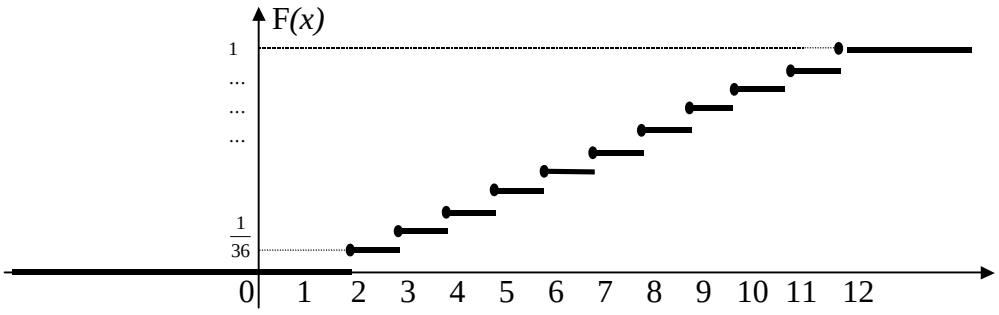
$$P(\{X > a\}) = 1 - F(a). \quad \textbf{(3)}$$

ESEMPI

1) Consideriamo il caso del lancio di due dadi e la variabile casuale che associa ad ogni risultato (a,b) la somma $a + b$. Abbiamo

x_i	2	3	4	5	6	7	8	9	10	11	12
$f(x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$F(x_i)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

Il grafico della funzione di ripartizione è il seguente



2) All'interno di un'urna ci sono 3 palline numerate dall'1 al 3. Considerando l'estrazione bernoulliana di due palline, lo spazio campionario è

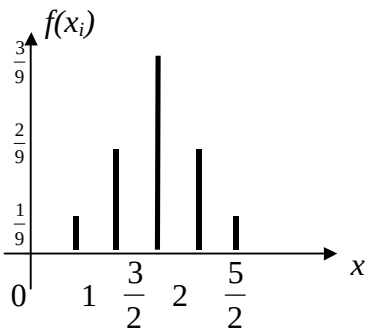
$$S = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$$

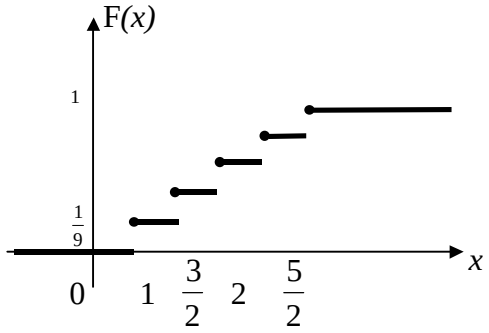
Sia X la variabile casuale tale che $X : S \to \mathbf{R}$, $\forall (a,b) \in S : X((a,b)) = \frac{a+b}{2}$.

Abbiamo quindi

x_i	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
$f(x_i)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$
$F(x_i)$	$\frac{1}{9}$	$\frac{3}{9}$	$\frac{6}{9}$	$\frac{8}{9}$	$\frac{9}{9}$

Le seguenti figure rappresentano rispettivamente la funzione di distribuzione e di ripartizione





APPLICHIAMO

1) Dopo aver determinato lo spazio campionario S relativo al lancio di un dado e di una moneta, considerate la variabile casuale X tale che $X(i,m) = \begin{cases} \frac{i+1}{2} & \text{se } m \text{ è testa} \\ \frac{i-1}{2} & \text{se } m \text{ è croce} \end{cases}$, dove i è il valore numerico del dado ed m può essere Testa o Croce. Costruite la tabella di distribuzione e i diagrammi delle funzioni di distribuzione e di ripartizione. Trovate quindi M(X), V(X) e s.
2) Determinate lo spazio campionario S relativo all'estrazione in blocco di due palline da un'urna contenente 5 palline numerate dall'1 al 5. Considerate la variabile casuale $X : S \to \mathbf{R}$ tale che $X(a) = \frac{60}{a}$, con a somma dei numeri riportati sulle palline.
   Costruite la tabella di distribuzione e i diagrammi delle funzioni di distribuzione e di ripartizione. Trovate infine M(X), V(X) e s.
3) Dopo aver determinato lo spazio campionario S relativo al numero di pezzi difettosi che possono trovarsi tra 5 pezzi, considerate la variabile casuale $X : S \to \mathbf{R}$ tale che $X(a) = a$. Sapendo che la probabilità p di estrarre un pezzo difettoso è uguale a 0.1, costruite la tabella di distribuzione e di ripartizione. Trovate infine M(X), V(X) e s.
4) Dopo aver determinato lo spazio campionario S relativo al numero di pazienti che guariscono da una malattia dopo aver ingerito un farmaco, considerate la variabile casuale $X : S \to \mathbf{R}$ tale che $X(a) = a$, sapendo che i pazienti sono 10. Se la probabilità che un paziente guarisca è uguale a 0.6, costruite la tabella di distribuzione e di ripartizione. Trovate infine M(X), V(X) e s.

1.6 Spazi campionari continui

Nel caso di spazi campionari e variabili casuali continui è di particolare interesse determinare la probabilità che la variabile casuale assuma valori appartenenti a un dato intervallo piuttosto che la probabilità che essa sia uguale a un determinato numero.

ESEMPIO

Supponiamo che i pesi di alcuni ragazzi siano compresi tra 59*kg* e 85*kg*. Pertanto lo spazio campionario è $S = \{p \in \mathbf{R} \setminus 59 \leq p \leq 85\}$ ed è continuo, come pure la variabile casuale $X : S \to \mathbf{R}$ tale che $X(a) = a$. Per indicare la probabilità che il peso di un ragazzo sia compreso tra 50*kg* e 52*kg* oppure uguale a 52*kg*, scriviamo $P(50 < X \leq 52)$ che, per la **(1)**, è uguale ad $F(52) - F(50)$, essendo $F(x)$ la funzione di ripartizione.

In generale, nel caso di funzioni di distribuzione continue, da

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

posto $x_1 = x$ e $x_2 = x + h$, segue

$$P(x < X \leq x + h) = F(x + h) - F(x)$$

Se $F(x)$ è derivabile possiamo considerare il $\lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = F'(x)$.

Possiamo pertanto introdurre la seguente

DEFINIZIONE (Funzione di densità di probabilità)

Siano date una variabile casuale continua $X : S \to \mathbf{R}$ e la funzione di ripartizione $F(x)$ di X, derivabile in **R** con derivata continua.

(*Funzione di densità di probabilità* della variabile casuale X) :⇔ (La funzione $g(x) = F'(x)$)

Nel caso di variabili casuali continue la funzione di densità di probabilità prende il posto della funzione di distribuzione trattate con le variabili casuali discrete.
Pertanto per determinare il valore medio, la varianza e lo scarto quadratico medio si considera la funzione di densità di probabilità.

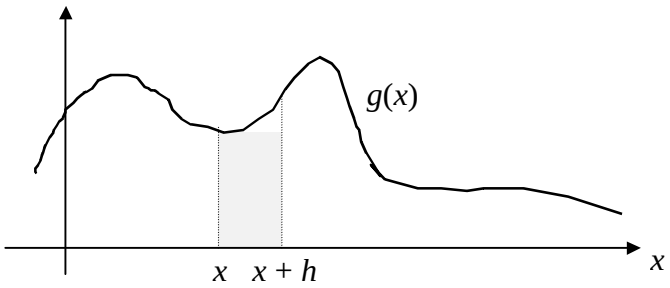
OSSERVAZIONI

◊ Poiché $\lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = g(x)$, per *h* "prossimo a zero" possiamo scrivere

$$F(x + h) - F(x) \approx g(x) \cdot h.$$

Diamo allora una giustificazione grafica delle funzioni di ripartizione $F(x)$ e di densità di probabilità.

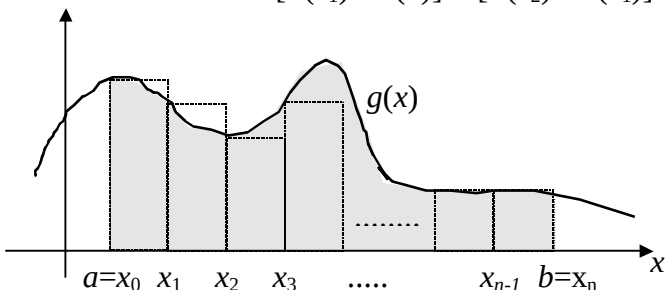
Supponiamo che il grafico della funzione di densità di probabilità $g(x)$ sia il seguente



Poiché $g(x) \cdot h$ è uguale all'area del rettangolo evidenziato nella precedente figura, possiamo quindi affermare che, per h prossimo a 0, l'espressione $F(x + h) - F(x)$ approssima l'area del rettangolo avente per base h e altezza uguale a $g(x)$.

Nel caso in cui abbiamo $P(a < X \leq b)$ possiamo scrivere $P(a < X \leq b) = F(b) - F(a) =$

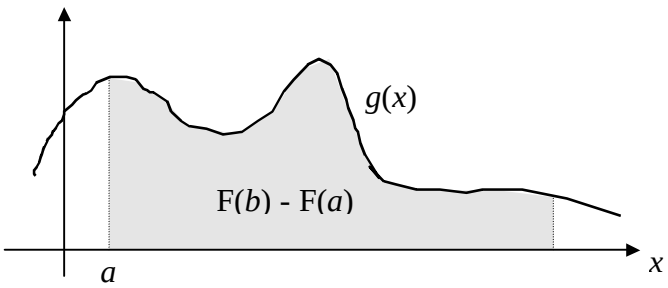
$$[F(x_1) - F(a)] + [F(x_2) - F(x_1)] + ... + [F(b) - F(x_{n-1})]$$



Per h prossimo a 0, l'ultima espressione può essere approssimata da
$h \cdot g(a) + h \cdot g(x_1) + h \cdot g(x_2) + ... + h \cdot g(x_{n-1}) = \sum_{i=0}^{n-1} h \cdot g(x_i)$.

Quando h tende a diventare sempre più piccolo il numero n degli intervalli tende a crescere e il lato superiore di ciascun rettangolo tende a "confondersi" con il grafico della funzione $g(x)$.
Pertanto la somma delle aree dei rettangoli, per h prossimo a 0, tende ad essere uguale all'area della figura delimitata dal grafico della funzione, dall'asse delle ascisse e dalle rette $x = a$ e $x = b$.

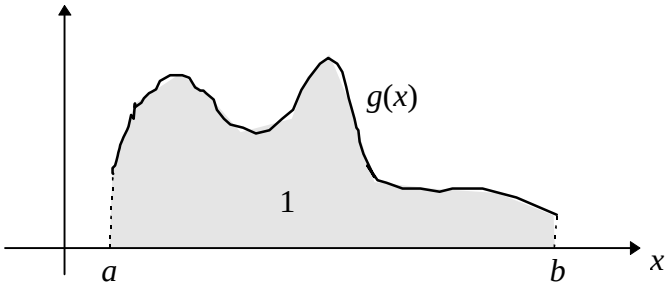
Concludiamo quindi che $F(b) - F(a)$, che è uguale a $P(a < X \leq b)$, corrisponde all'area evidenziata.



Inoltre

◊ Possiamo scrivere $P(X = a) = \lim_{h \to 0} P(a < X \leq a + h) = \lim_{h \to 0} F(a + h) - F(a)$, e poiché il limite è uguale a zero, perché per ipotesi la funzione di ripartizione $F(x)$ è continua, segue che $P(X = a) = 0$.
Pertanto $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$.

◊ Se la variabile casuale può assumere valori compresi tra *a* e *b*, poiché P($a \le X \le b$) = 1, risulta F(*b*) - F(*a*) = 1. Pertanto l'area compresa tra il grafico di *g*(*x*), l'asse delle ascisse e le rette *x* = *a* e *x* = *b* è uguale a 1.

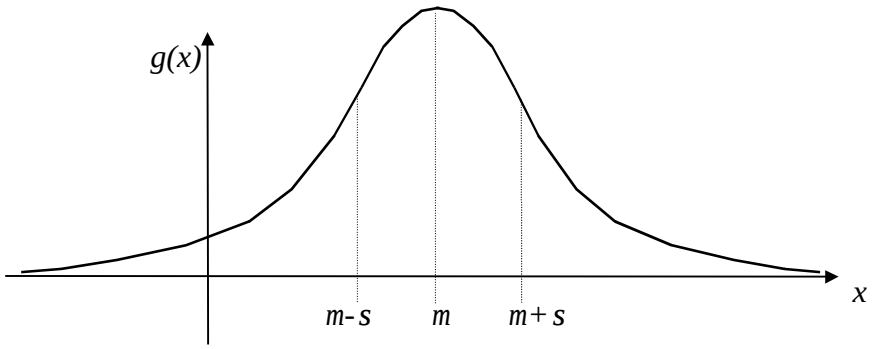


1.7 Distribuzione Normale

La funzione di densità di probabilità nel caso di distribuzione normale è

$$g(x) = \frac{1}{s\sqrt{2p}} e^{-\frac{(x-m)^2}{2s^2}}$$

I parametri m e s sono rispettivamente il valore medio e lo scarto quadratico medio
Essa è indicata con N(m,s) e il grafico è

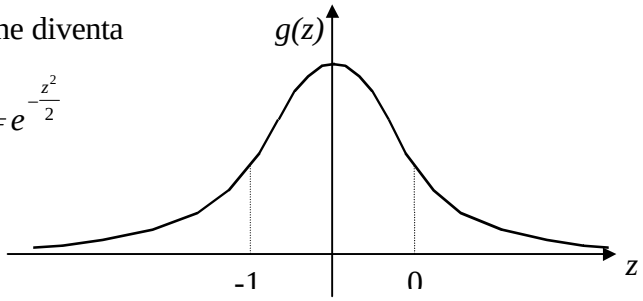


Tale funzione, simmetrica rispetto alla retta di equazione *x* = *m*, possiede il massimo nel punto M($m, \frac{1}{s\sqrt{2p}}$) ed i flessi nei punti $F_1(m-s, \frac{1}{s\sqrt{2pe}})$ ed $F_2((m+s, \frac{1}{s\sqrt{2pe}})$.

Effettuando il cambio di variabile $z = \frac{x-m}{s}$ la funzione diventa

$$g(z) = \frac{1}{\sqrt{2p}} e^{-\frac{z^2}{2}}$$

ed ha il seguente grafico



Grazie al precedente cambio di variabili l'espressione della funzione è più semplice e compatta, inoltre *g*(*z*) è simmetrica rispetto all'asse delle ordinate e, per il teorema sulla variabile casuale standardizzata, il valore medio è uguale a 0 e lo scarto quadratico medio uguale a 1. Inoltre l'area della figura delimitata dalla curva, dagli assi coordinati e dalla retta di equazione *x* = 1 è uguale a 0.3415 circa, da cui segue che l'area della figura delimitata dalla curva, dall'asse delle ascisse e

◊ dalle rette di equazioni *x* = -1 e *x* = 1 è uguale a 0.683,

◊ dalle rette di equazioni $x = -3$ e $x = 3$ è uguale a 0.997.

In tale caso la distribuzione prende il nome di *distribuzione normale standardizzata*, indicata con N(0,1).

1.8 Campioni e medie campionarie

Supponiamo di voler considerare il peso medio di un gruppo di *N* persone procedendo a estrazioni di tipo bernoulliano di *n* di essi, formando così dei *campioni* di dimensione *n*. In tale caso lo **spazio campionario** S è l'**insieme costituito da tutti i possibili campioni** di dimensione *n*. Poiché le estrazioni sono bernoulliane abbiamo $|S| = N^n$.

Sia X_1 la variabile casuale che associa ad un campione il peso della prima persona estratta, X_2 la variabile casuale che associa ad un campione il peso della seconda persona estratta, e così via sino a X_n che associa ad un campione il peso della *n*-esima persona estratta. Poiché le estrazioni sono bernuolliane abbiamo *n* variabili casuali X_1, X_2, ..., X_n indipendenti fra loro e aventi la stessa distribuzione di probabilità e quindi stesso valore medio m e stessa varianza s^2.
La variabile casuale $X = \frac{X_1 + X_2 + ... + X_n}{n}$ rappresenta il peso medio delle *n* persone estratte. Ovviamente per un dato campione assume un certo valore, se procediamo a una nuova estrazione cambiando il campione, X assumerà un valore diverso dal precedente.

Per tale variabile casuale abbiamo

$M(X) = M(\frac{X_1 + X_2 + ... + X_n}{n}) = \frac{1}{n} M(X_1 + X_2 + ... + X_n) =$ (per la proprietà A) del valore medio)

$= \frac{1}{n} [M(X_1) + M(X_2) + ... + M(X_n)]$ = (per la proprietà G) del valore medio)

$= \frac{1}{n} n\, m = m.$

Pertanto

$$M(X) = m.$$

Inoltre $V(X) = V(\frac{X_1 + X_2 + ... + X_n}{n}) = \frac{1}{n^2} V(X_1 + X_2 + ... + X_n)$ per la proprietà D) della varianza,

$= \frac{1}{n^2} [V(X_1) + V(X_2) + ... + V(X_n)]$ per la proprietà G) della varianza,

$= \frac{1}{n^2} n\, s^2 = \frac{s^2}{n}$. Pertanto $V(X) = \frac{s^2}{n}$.

Quindi per la variabile casuale X, chiamata *media campionaria*, risulta

$$M(X) = m \text{ e } V(X) = \frac{s^2}{n}.$$

La notevole importanza posseduta dalla distribuzione normale è messa in evidenza dal seguente

TEOREMA (del limite centrale)

Siano X_1, X_2, ..., X_n, n variabili casuali indipendenti fra loro e aventi la stessa distribuzione di probabilità e quindi stesso valore medio m e stessa varianza s^2.
Allora ,

posto $X = \frac{X_1 + X_2 + \ldots + X_n}{n}$,

la variabile casuale standardizzata $\frac{X - m}{\frac{s}{\sqrt{n}}}$, per n tendente all'infinito, tende a distribuirsi come la variabile casuale avente come funzione di densità di probabilità N(0,1).

Cerchiamo di mostrare il significato del precedente teorema ricorrendo al seguente

ESEMPIO

Consideriamo cinque persone individuate dalle lettere A, B, C, D ed E, di altezza 1.72*m*, 1.54*m*, 1.83*m*, 1.65*m* e 1.80*m* rispettivamente. La media della precedente distribuzione è m = 1.708 e la varianza s^2 = 0.011016.

◊ Effettuando estrazioni bernoulliana di **2** persone, abbiamo 25 campioni (n = 2). Dette X_1 la variabile casuale che associa ad un campione l'altezza della prima persona estratta e X_2 la variabile casuale che associa ad un campione l'altezza della seconda persona estratta.
Ad esempio X_1(BE) = 1.54 e X_2(BE) = 1.80.

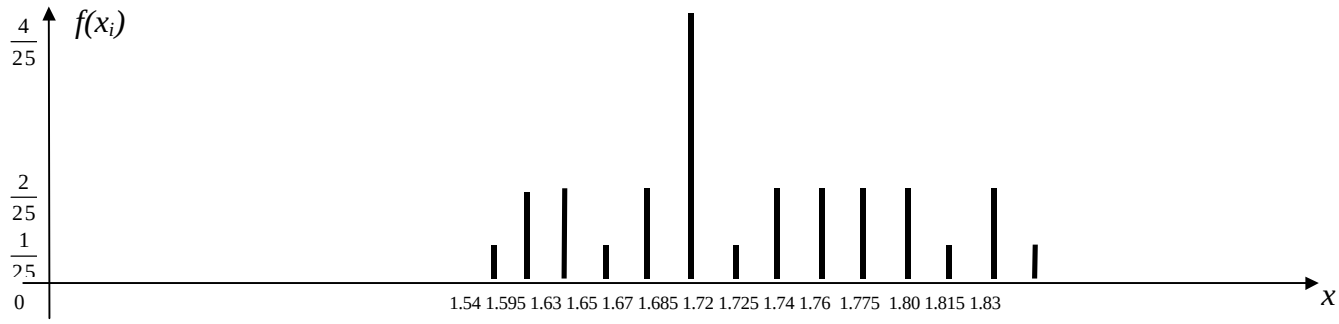
Posto $X = \frac{X_1 + X_2}{2}$, abbiamo

	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
X	1.72	1.63	1.775	1.685	1.76	1.63	1.54	1.685	1.595	1.67	1.775	1.685	1.83	1.74	1.815	1.685	1.595	1.74	1.65	1.725	1.76	1.67	1.815	1.725	1.80

e la tabella di distribuzione è

X	1.54	1.595	1.63	1.65	1.67	1.685	1.72	1.725	1.74	1.76	1.775	1.80	1.815	1.83
$f(x_i)$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

e il diagramma è



Per tale variabile casuale risulta

$$M(X) = 1.54 \cdot \frac{1}{25} + 1.595 \cdot \frac{2}{25} + \ldots + 1.83 \cdot \frac{1}{25} = 1.708 = m \quad \text{e}$$

$$V(X) = (1.54 - 1.708)^2 \cdot \frac{1}{25} + (1.595 - 1.708)^2 \cdot \frac{2}{25} + \ldots + (1.83 - 1.708)^2 \cdot \frac{1}{25} = 0.005508 = \frac{s^2}{2}$$

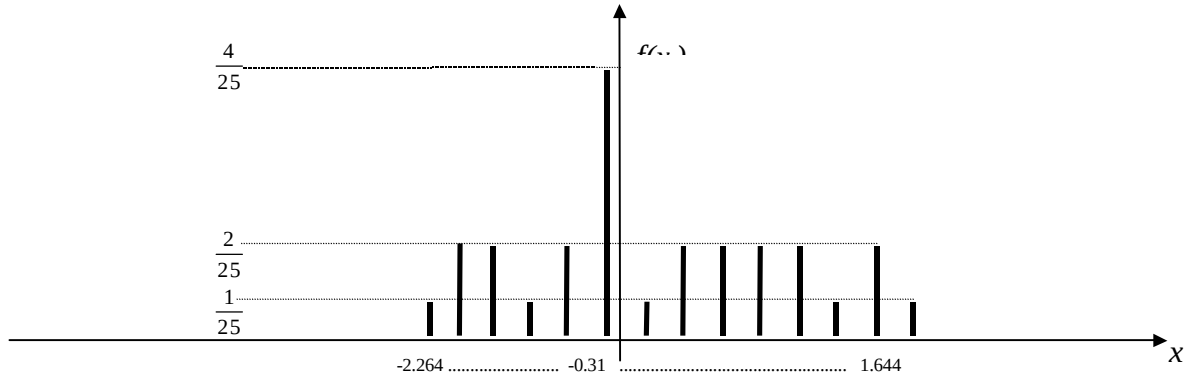
Pertanto la variabile casuale standardizzata è

$$\frac{X - 1.708}{\frac{0.10496}{\sqrt{2}}} = \frac{X - 1.708}{0.07422}$$

ed ha la seguente tabella di distribuzione (i valori di X sono approssimati)

X	-2.264	-1.523	-1.051	-0.781	-0.512	-0.31	0.162	0.229	0.431	0.701	0.903	1.24	1.442	1.644
$f(x_i)$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

e il seguente diagramma



◊ Effettuando estrazioni bernoulliana di **3** persone, abbiamo 125 campioni (n = 3). Dette X_1 la variabile casuale che associa ad un campione l'altezza della prima persona estratta, X_2 la variabile casuale che associa ad un campione l'altezza della seconda persona estratta e X_3 la variabile casuale che associa ad un campione l'altezza della terza persona estratta,

posto $X = \frac{X_1 + X_2 + X_3}{3}$, abbiamo la seguente tabella di distribuzione

X	1.54	$1.57\overline{3}$	1.60	$1.61\overline{3}$	$1.62\overline{6}$	$1.63\overline{6}$	1.65	1.66	$1.66\overline{3}$	$1.67\overline{3}$	$1.68\overline{6}$	$1.69\overline{6}$	1.70	1.71	1.72
$f(x_i)$	$\frac{1}{125}$	$\frac{3}{125}$	$\frac{3}{125}$	$\frac{3}{125}$	$\frac{3}{125}$	$\frac{9}{125}$	$\frac{1}{125}$	$\frac{3}{125}$	$\frac{2}{125}$	$\frac{9}{125}$	$\frac{6}{125}$	$\frac{9}{125}$	$\frac{3}{125}$	$\frac{3}{125}$	$\frac{1}{125}$

X	$1.72\bar{3}$	$1.7\bar{3}$	$1.74\bar{6}$	1.75	$1.75\bar{6}$	1.76	1.77	$1.77\bar{3}$	$1.78\bar{3}$	$1.79\bar{3}$	1.80	1.81	1.82	1.83
$f(x_i)$	$\frac{12}{125}$	$\frac{9}{125}$	$\frac{3}{125}$	$\frac{3}{125}$	$\frac{3}{125}$	$\frac{6}{125}$	$\frac{3}{125}$	$\frac{3}{125}$	$\frac{6}{125}$	$\frac{3}{125}$	$\frac{1}{125}$	$\frac{3}{125}$	$\frac{3}{125}$	$\frac{1}{125}$

Per tale variabile casuale risulta

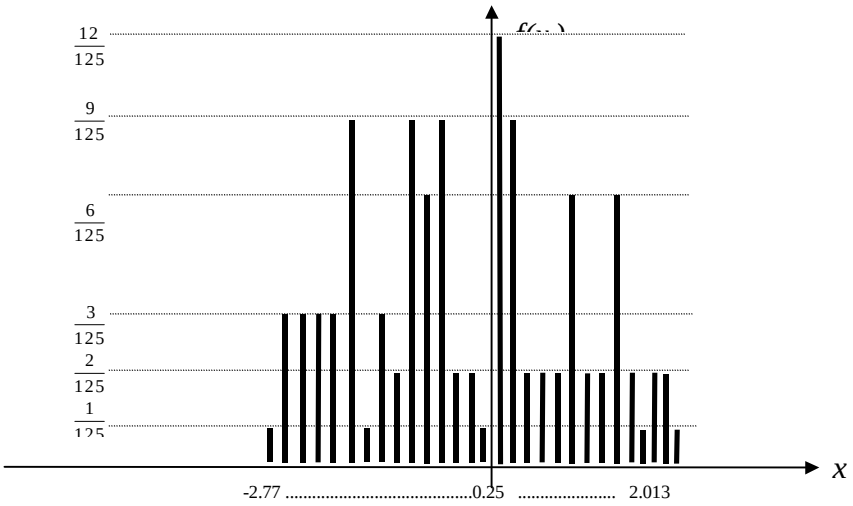
$$M(X) = 1.54 \cdot \frac{1}{125} + 1.57\bar{3} \cdot \frac{3}{125} + \ldots + 1.83 \cdot \frac{1}{125} = 1.708 = m \text{ e}$$

$$V(X) = (1.54 - 1.708)^2 \cdot \frac{1}{125} + \left(1.57\bar{3} - 1.708\right)^2 \cdot \frac{3}{25} + \ldots + (1.83 - 1.708)^2 \cdot \frac{1}{125} = 0.003672 = \frac{s^2}{3}$$

Pertanto la variabile casuale standardizzata è

$$\frac{X - 1.708}{\frac{0.10496}{\sqrt{3}}} = \frac{X - 1.708}{0.0606}$$

ed ha il seguente diagramma



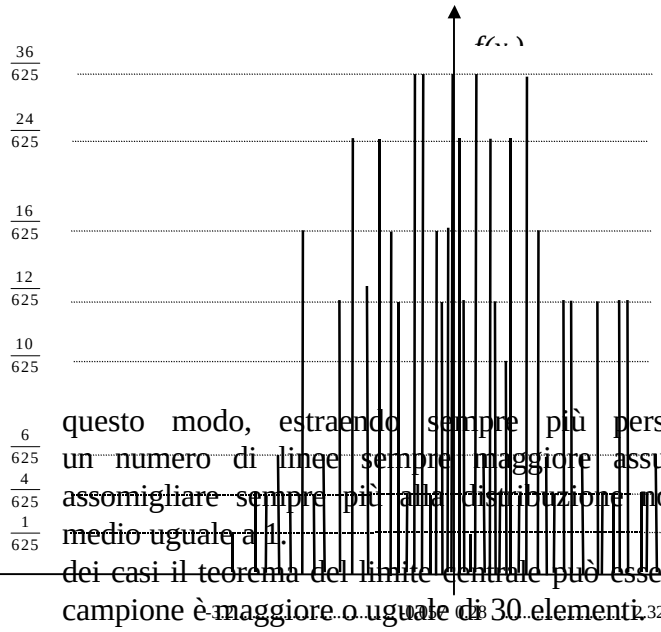
◊ Effettuando estrazioni bernoulliana di **4** persone, abbiamo 625 campioni (n = 4). Dette X_1 la variabile casuale che associa ad un campione l'altezza della prima persona estratta, X_2 la variabile casuale che associa ad un campione l'altezza della seconda persona estratta, X_3 la variabile casuale che associa ad un campione l'altezza della terza persona estratta e X_4 la variabile casuale che associa ad un campione l'altezza della quarta persona estratta, posto

$$X = \frac{X_1 + X_2 + X_3 + X_4}{4},$$

si avrà una certa tabella di distribuzione, la media
M(X) =1.708 = m e

la varianza V(X) = 0.002754 = $\frac{s^2}{4}$

La variabile casuale standardizzata è $\frac{X - 1.708}{\frac{0.10496}{\sqrt{4}}} = \frac{X - 1.708}{0.05248}$ ed ha il seguente diagramma



Procedendo in questo modo, estraendo sempre più persone, vedremo come il diagramma conterrà un numero di linee sempre maggiore assumendo una disposizione tendente ad assomigliare sempre più alla distribuzione normale avente media 0 e scarto quadratico medio uguale a 1.
Nella maggior parte dei casi il teorema del limite centrale può essere utilizzato già quando la dimensione di un campione è maggiore o uguale di 30 elementi.

Vediamo ora delle possibili applicazioni utili del precedente teorema.

ESEMPI

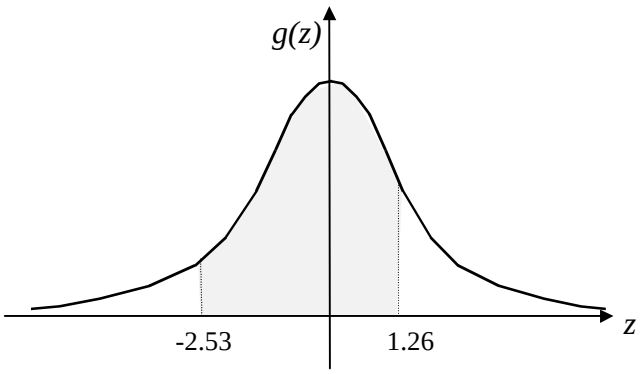
Il peso medio di un gruppo di persone è di 68*kg* con varianza di 500.

1) Determinate la probabilità che il peso medio di un campione di 50 persone prese a caso con estrazione bernoulliana sia compreso tra 60*kg* e 72*kg*.

   Pur non sapendo la distribuzione dei pesi, applichiamo il teorema del limite centrale. Pertanto la distribuzione della media campionaria standardizzata $Z = \frac{X - 68}{\sqrt{\frac{500}{50}}}$ può essere approssimata alla normale con media 0 e scarto quadratico medio 1. Allora

$$P(60 < X < 72) = P\left(\frac{60 - 68}{\sqrt{\frac{500}{50}}} < Z < \frac{72 - 68}{\sqrt{\frac{500}{50}}} \right) =$$

$$= P(-2.53 < Z < 1.26) = F(1.26) - F(-2.53).$$



Consultando la "tabella di probabilità sotto la curva normale standardizzata", dove sono riportate le aree comprese tra il grafico di $g(x)$, l'asse delle ascisse e le rette $z = 0$ e $z = a$, posto $F_0(z) = P(0 < Z < z)$, troviamo
$F_0(1.26) = 0.3962$
$F_0(2.53) = 0.4943$.

Inoltre, per la simmetria della curva normale, abbiamo

- $F_0(-z) = F_0(z)$,
- $F(z) = 0.5 + F_0(z)$ se z è positivo,
- $F(z) = 0.5 - F_0(-z)$ se z è negativo.

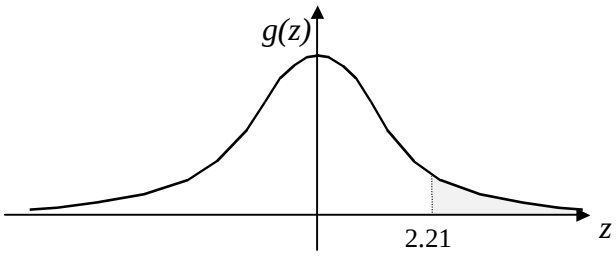
Pertanto segue

$F(1.26) - F(-2.53) = [0.5 + F_0(1.26)] - [0.5 - F_0(2.53)] = F_0(1.26) + F_0(2.53) =$
$= 0.3962 + 0.4943 = 0.8905$ che rappresenta la probabilità che cerchiamo.

2) Determinate la probabilità che il peso medio di un campione di 50 persone prese a caso con estrazione bernoulliana sia maggiore di 75*kg*.

In tale caso vogliamo conoscere

$$P(X > 75) = P\left(Z > \frac{75-68}{\sqrt{\frac{500}{50}}}\right) = P(Z > 2.21) = 1 - F(2.21).$$



In tale caso abbiamo
$F_0(2.21) = 0.4864$.

Poiché $F(2.21) = 0.5 + F_0(2.21)$ segue
$P(X > 75) = 1- [0.5 + F_0(2.21) =$
$= 0.5 - F_0(2.21) = 0.5 - 0.4864 = 0.0136$.

3) Determinate la probabilità che il peso medio di un campione di 210 persone prese a caso con estrazione bernoulliana non superi i 65*kg*.

Vogliamo conoscere $P(X \le 65) = P\left(Z \le \frac{65-68}{\sqrt{\frac{500}{210}}}\right) = P(Z \le -1.94) = F(-1.94) =$

$= 0.5 - F_0(1.94) = 0.5 - 0.4738 = 0.0262$.

APPLICHIAMO

1) Il peso medio di un gruppo di persone è di 68*kg* con scarto quadratico medio di 30.
   - ◊ Determinate la probabilità che il peso medio di un campione di 80 persone prese a caso con estrazione bernoulliana sia compreso tra 45*kg* e 67*kg*.
   - ◊ Determinate la probabilità che il peso medio di un campione di 110 persone prese a caso con estrazione bernoulliana non superi i 65*kg*.
   - ◊ Determinate la probabilità che il peso medio di un campione di 80 persone prese a caso con estrazione bernoulliana sia maggiore di 80*kg*.
2) L'altezza media di un gruppo di persone è di 1.80*m* con scarto quadratico medio di 0.75.
   - ◊ Determinate la probabilità che il peso medio di un campione di 120 persone prese a caso con estrazione bernoulliana sia compreso tra 1.70*m* e 1.75*m*.
   - ◊ Determinate la probabilità che il peso medio di un campione di 70 persone prese a caso con estrazione bernoulliana superi 1.85*m*.
   - ◊ Determinate la probabilità che il peso medio di un campione di 85 persone prese a caso con estrazione bernoulliana non sia inferiore di 1.75*m*.
3) Il voto medio di un gruppo di studenti è di 6.30 con scarto quadratico medio di 2.
   - ◊ Determinate la probabilità che il voto medio di un campione di 32 studenti presi a caso con estrazione bernoulliana sia compreso tra 6 e 6.2.
   - ◊ Determinate la probabilità che il voto medio di un campione di 45 studenti presi a caso con estrazione bernoulliana non superi 5.8.
   - ◊ Determinate la probabilità che il voto medio di un campione di 30 persone prese a caso con estrazione bernoulliana sia inferiore di 6 oppure superiore di 8.

OSSERVAZIONE

Supponiamo di ripetere un dato esperimento *n* volte nelle medesime condizioni. Consideriamo quindi la distribuzione binomiale o bernoulliana. Solitamente in questi casi si ricerca il numero di volte che un certo evento si ripete, ossia della sua frequenza assoluta.
Ciascun campione dello spazio campionario è composto dalle uscite degli *n* esperimenti.

Definiamo X_1 quella variabile casuale che associa ad un campione
- 1 se nella prima prova si è verificato l'evento,
- 0 in caso contrario,

X_2 quella variabile casuale che associa ad un campione
- 1 se nella seconda prova si è verificato l'evento,
- 0 in caso contrario,

e così via sino a X_n, che associa ad un campione
- 1 se nell'*n*-esima prova si è verificato l'evento,
- 0 in caso contrario.

Poiché le prove sono effettuate nelle medesime condizioni, le *n* variabili casuali X_1, X_2, ..., X_n risultano indipendenti fra loro ed hanno la stessa distribuzione di probabilità, stesso valore medio, che in questo caso coincide con la probabilità *m* dell'evento, e stessa varianza.

La somma $X_1 + X_2 + \ldots + X_n$, applicata a un dato campione, rappresenta il numero di volte che in esso si è verificato l'evento, ossia la frequenza assoluta, che indichiamo con *n*X, essendo X la *frequenza relativa campionaria*.

Si dimostra che la variabile casuale standardizzata $\frac{n\mathrm{X}-nm}{\sqrt{nm(1-m)}}$, per n tendente all'infinito, tende a distribuirsi come la variabile casuale avente come funzione di densità di probabilità N(0,1).

Stesse conclusioni valgono anche nel caso in cui la somma $\mathrm{X}_1+\mathrm{X}_2+\ldots+\mathrm{X}_n$, applicata a un dato campione, rappresenta il numero di volte che compare un dato attributo. In tale caso m rappresenta la frequenza relativa dell'attributo ed

X_1 è la variabile casuale che associa ad un campione
- 1 se il primo elemento del campione ha l'attributo,
- 0 in caso contrario,

X_2 quella variabile casuale che associa ad un campione
- 1 se il secondo elemento del campione ha l'attributo,
- 0 in caso contrario,

e così via sino a X_n, che associa ad un campione
- 1 se l'n-esimo elemento del campione ha l'attributo,
- 0 in caso contrario.

Vediamo come utilizzare le precedenti considerazioni con i seguenti

ESEMPI

Per 90 volte estraiamo due palline da un'urna che ne contiene 3 bianche e 5 nere.

1) Determinate la probabilità che per 6 volte estraiamo due palline bianche.

   La probabilità m di estrarre due palline bianche è $m=\frac{\binom{3}{2}}{\binom{8}{2}}=\frac{3}{28}$, mentre la probabilità contraria $1-m=1-\frac{\binom{3}{2}}{\binom{8}{2}}=\frac{25}{28}$.

   Dovendo trovare la probabilità ricorrendo alla distribuzione binomiale troviamo

$$\binom{90}{6}\left(\frac{3}{28}\right)^6\left(\frac{25}{28}\right)^{84}=0.069133 \quad \textbf{(1)}$$

   Possiamo approssimare il risultato procedendo in modo diverso, utilizzando il teorema del limite centrale e l'osservazione precedente.

   Consideriamo campioni ciascuno formato dai risultati ottenuti da 90 estrazioni.

   Sia X_1 la variabile casuale che associa ad un campione
   - 1 se nella sua prima estrazione abbiamo preso due palline bianche,
   - 0 in caso contrario,

   X_2 che associa ad un campione
   - 1 se nella sua seconda estrazione abbiamo preso due palline bianche,

- 0 in caso contrario,

e così via sino ad X_{90}, che associa ad un campione

- 1 se nella novantesima estrazione nel campione considerato abbiamo estratto due palline bianche,
- 0 in caso contrario.

Allora $90\cdot X = X_1 + X_2 + \ldots + X_{90}$ rappresenta il numero di volte in cui sono state estratte due palline bianche nel campione preso in esame, ossia rappresenta la frequenza assoluta riguardante l'estrazione di due palline bianche.

Determiniamo allora P(90·X = 6), approssimando tale probabilità con P(5.5 < 90·X < 6.5).
Abbiamo che P(5.5 < 90·X < 6.5) =

$$P\left(\frac{5.5-nm}{\sqrt{nm(1-h)}} < \frac{n\mathrm{X}-nm}{\sqrt{nm(1-h)}} < \frac{6.5-nm}{\sqrt{nm(1-h)}}\right) = P\left(\frac{5.5-90\cdot\frac{3}{28}}{\sqrt{90\cdot\frac{3}{28}\cdot\frac{25}{28}}} < \frac{90\cdot X-90\frac{3}{28}}{\sqrt{90\cdot\frac{3}{28}\cdot\frac{25}{28}}} < \frac{6.5-90\cdot\frac{3}{28}}{\sqrt{90\cdot\frac{3}{28}\cdot\frac{25}{28}}}\right) =$$

$$= P(-1.41 < Z < -1.07) = F(-1.07) - F(-1.41) =$$

$$[0.5 - F_0(1.07)] - [0.5 - F_0(1.41)] = F_0(1.41) - F_0(1.07).$$

Infine risulta

$$F_0(1.41) - F_0(1.07) = 0.4207 - 0.3577 = 0.063$$

che può ritenersi una approssimazione accettabile di **(1)**

2) Determinate la probabilità che il numero di volte che estraiamo due palline nere sia maggiore o uguale di 30.

La probabilità m di estrarre due palline nere è $m = \frac{\binom{5}{2}}{\binom{8}{2}} = \frac{5}{14}$, mentre la probabilità contraria è

$1 - m = 1 - \frac{\binom{5}{2}}{\binom{8}{2}} = \frac{9}{14}$.

Dovendo trovare la probabilità richiesta ricorrendo alla distribuzione binomiale calcoliamo la somma

$$\binom{90}{30}\left(\frac{5}{14}\right)^{30}\left(\frac{9}{14}\right)^{60} + \binom{90}{31}\left(\frac{5}{14}\right)^{31}\left(\frac{9}{14}\right)^{59} + \ldots + \binom{90}{90}\left(\frac{5}{14}\right)^{90}\left(\frac{9}{14}\right)^{0} \quad \textbf{(2)}$$

calcolo laborioso e piuttosto lungo, che dà come risultato 0.717024.

Possiamo approssimare il risultato utilizzando il teorema del limite centrale.

Determiniamo P(90·X ≥ 30).

Per l'esempio precedente, il passaggio dal discreto al continuo consente di scrivere
$P(90\cdot X \geq 30) = P(90\cdot X = 30) + P(90\cdot X = 31) + ... =$
$P(29.5 < 90\cdot X < 30.5) + P(30.5 < 90\cdot X < 31.5) + ... =$
$P(90\cdot X > 29.5)$, dopo aver ricordato che nel caso di variabili casuali continue $P(a \leq X \leq b) = P(a < X < b)$.

Abbiamo dunque

$$P(90\cdot X > 29.5) = P\left(\frac{90\cdot X - 90\frac{5}{14}}{\sqrt{90\cdot\frac{5}{14}\cdot\frac{9}{14}}} > \frac{29.5 - 90\cdot\frac{5}{14}}{\sqrt{90\cdot\frac{5}{14}\cdot\frac{9}{14}}} \right) =$$

$= P(Z > -0.58) = 1 - F(-0.58) = 1- [0.5 - F_0(0.58)] = 0.5 + F_0(0.58)$.

Infine risulta

$$0.5 + F_0(0.58) = 0.5 + 0.2190 = 0.719$$

che può ritenersi una approssimazione accettabile di **(2)**.

3) Determinate la probabilità che il numero di volte che estraiamo due palline di diverso colore sia minore di 42.

La probabilità m di estrarre due palline di colore diverso è $m = \frac{\binom{3}{1}\cdot\binom{5}{1}}{\binom{8}{2}} = \frac{15}{28}$, la probabilità contraria $1 - m = 1 - \frac{\binom{3}{2}}{\binom{8}{2}} = \frac{13}{28}$.

Ricorrendo alla distribuzione binomiale, per determinare la probabilità richiesta calcoliamo la somma

$$\binom{90}{0}\left(\frac{15}{28}\right)^0\left(\frac{13}{28}\right)^{90} + \binom{90}{1}\left(\frac{15}{28}\right)^1\left(\frac{13}{28}\right)^{89} + ... + \binom{90}{41}\left(\frac{15}{28}\right)^{41}\left(\frac{13}{28}\right)^{49} \quad \textbf{(3)}$$

che dà come risultato 0.07811.

Possiamo approssimare il risultato utilizzando il teorema del limite centrale.
Vogliamo determinare $P(90\cdot X < 42)$.

Passando dal discreto al continuo, scriviamo

$P(90\cdot X < 42) = P(90\cdot X = 41) + P(90\cdot X = 40) + ... =$
$P(40.5 < 90\cdot X < 41.5) + P(39.5 < 90\cdot X < 40.5) + ... =$

$$P(90\cdot X < 41.5) = P\left(\frac{90\cdot X - 90\frac{15}{28}}{\sqrt{90\cdot\frac{15}{28}\cdot\frac{13}{28}}} < \frac{41.5 - 90\cdot\frac{15}{28}}{\sqrt{90\cdot\frac{15}{28}\cdot\frac{13}{28}}} \right) =$$

$= P(Z < -1.42) = F(-1.42) = [0.5 - F_0(1.42)] = 0.5 - 0.4222 = 0.0778.$

che può ritenersi una approssimazione accettabile di **(3)**.

APPLICHIAMO

Risolvete i seguenti esercizi sia mediante la distribuzione bernoulliana sia approssimando mediante la distribuzione normale.

1) Per 900 volte estraiamo due palline da un'urna che ne contenente 10 bianche e 7 nere.
   - ◊ Determinate la probabilità che per 75 volte estraiamo due palline bianche.
   - ◊ Determinate la probabilità che il numero di volte che estraiamo due palline nere sia maggiore di 351.
   - ◊ Determinate la probabilità che il numero di volte che estraiamo due palline di diverso colore sia minore o uguale di 370.
2) Per 100 volte lanciamo due dadi.
   - ◊ Determinate la probabilità che per 15 volte esca 7.
   - ◊ Determinate la probabilità che il numero di volte che esca un numero divisibile per tre sia maggiore di 65.
   - ◊ Determinate la probabilità che il numero di volte che esca un numero minore di quattro sia minore o uguale di 15.
3) Per 500 volte lanciamo due monete.
   - ◊ Determinate la probabilità che per 50 volte escano una Croce e una Testa.
   - ◊ Determinate la probabilità che il numero di volte che escano due Teste sia maggiore o uguale di 200.
   - ◊ Determinate la probabilità che il numero di volte che escano due Teste sia minore di 300.
4) Sapendo che l'indice di disoccupazione di un paese è del 15%,
   - ◊ Determinate la probabilità che in un campione di 80 persone, prese con estrazione bernoulliana, 7 siano disoccupate.
   - ◊ Determinate la probabilità che in un campione di 40 persone, prese con estrazione bernoulliana, i disoccupati siano in numero minore di 5.
   - ◊ Determinate la probabilità che su 300 persone, prese con estrazione bernoulliana, i disoccupati siano più di 20 ma meno di 40.
5) Sapendo che l'indice di difettosità di un lotto di pezzi prodotti da una fabbrica sia uguale a 0.1,
   - ◊ Determinate la probabilità che in un campione di 100 pezzi, presi con estrazione bernoulliana, 12 siano difettosi.
   - ◊ Determinate la probabilità che in un campione di 130 pezzi, presi con estrazione bernoulliana, i pezzi difettosi siano maggiori di 10.
   - ◊ Determinate la probabilità che in un campione di 200 pezzi, presi con estrazione bernoulliana, i pezzi difettosi siano più di 10 ma meno di 50.

1.9 Stimatori

Consideriamo un insieme formato da *N* elementi, chiamata popolazione, e di volerne studiare alcuni caratteri sia quantitativi, come altezza, peso, reddito, che qualitativi, come preferenze musicali, politiche, aspirazioni professionali ecc.

Questi caratteri sono di solito sintetizzati da alcuni *parametri* che di solito sono la *media* indicata con m e lo *scarto quadratico medio*, indicato con s.
In generale non è possibile rilevare i dati dell'intera popolazione, pertanto si preferisce ricorrere all'estrazione di un suo sottoinsieme, detto *campione*, formato da n elementi ($n < N$). Supponiamo che le estrazioni dalla popolazione per formare i campioni sono di tipo bernoulliano.

L'i*nferenza statistica* ha lo scopo di dedurre delle ipotesi riguardanti l'intera popolazione dallo studio del solo campione.

Facciamo notare come la scelta degli elementi di un campione deve avvenire mediante estrazione casuale rispettando, però, la distribuzione di alcuni caratteri della popolazione ritenuti importanti, come la classe di età, la posizione geografica, la religione, la categoria sociale, ecc.

DEFINIZIONI (stima, stimatore)

Dati una popolazione, un suo *parametro* q e un campione della popolazione

(*Stima*) :⇔ (Valore numerico ottenuto dal campione che fornisce delle indicazioni sul parametro q della popolazione)

(*Stimatore*) :⇔ (Formula che dipende dagli elementi del campione e che permette di determinare la stima di q)

Inoltre uno stimatore è

(*Corretto*) :⇔(Il valore medio delle stime di tutti i possibili campioni di dimensione n che si possono formare dalla popolazione, è uguale al parametro q della popolazione)

(*Coerente*) :⇔(Quando n diventa sempre più grande, la stima di ciascun campione si avvicina a q o, che è lo stesso, la sua varianza tende a 0)

(*Efficiente*) :⇔(La varianza delle stime di tutti i possibili campioni di dimensione n che si possono formare dalla popolazione, è minore di qualsiasi altra varianza ottenibile da altri stimatori di q)

Date $X_1, X_2, ..., X_n$, n variabili casuali indipendenti fra loro, poniamo $X = \frac{X_1 + X_2 + ... + X_n}{n}$.

- Nel caso di estrazione bernoulliana, se consideriamo la media campionaria X come stimatore della media m, ossia $q = m$, abbiamo
  - ◊ è corretta, perché abbiamo già visto che $M(X) = m$,
  - ◊ è coerente perché $\lim_{n\to\infty} V(X) = \lim_{n\to\infty} \frac{s^2}{n} = 0$,
  - ◊ inoltre si dimostra che è anche efficiente.
- Nel caso in cui X rappresenta invece la *frequenza relativa* campionaria e m la frequenza relativa della popolazione, ossia se $q = m$, abbiamo

◊ è corretta, perché $M(X) = \frac{1}{n}[M(X_1) + M(X_2) + ... + M(X_n)] = \frac{1}{n} n \cdot m = m$,

◊ è coerente. Infatti $V(X) = \frac{1}{n^2}[V(X_1) + V(X_2) + ... + V(X_n)] = \frac{1}{n^2} n \cdot m \cdot (1\text{-}m) = \frac{m \cdot (1-m)}{n}$, ossia

$$V(X) = \frac{m \cdot (1-m)}{n} \textbf{ (F)}$$

e quindi $\lim_{n\to\infty} V(X) = \lim_{n\to\infty} \frac{m(1-m)}{n} = 0$.

◊ inoltre si dimostra che è efficiente.

♦ Possiamo considerare un altro stimatore, la *varianza campionaria*, che dovrebbe dare una stima della varianza s^2 della popolazione. Essa è definita con la seguente formula

$$S^2 = \frac{(X_1 - X)^2 + ... + (X_n - X)^2}{n}$$

Poiché si dimostra che $M(S^2) = \frac{(n-1)}{n} s^2$, abbiamo che la varianza campionaria non è corretta.

Per questo si prende come stimatore X della varianza s^2 della popolazione lo stimatore

$$S_C^2 = \frac{n}{(n-1)} \cdot S^2$$

Infatti $M(S_C^2) = M\left(\frac{n}{n-1} S^2\right) = \frac{n}{n-1} M(S^2) = \frac{\cancel{n}}{\cancel{n-1}} \frac{\cancel{n-1}}{\cancel{n}} s^2 = s^2$.

che chiamiamo varianza campionaria corretta ed

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

lo *scarto quadratico medio campionario corretto*.

ESEMPIO

Supponiamo che una ricerca sul reddito medio condotta su un campione A di 300 famiglie siano stati determinati i seguenti valori X(A) = £ 2.560.000 ed $S_C^2(A) = £126.000$. Poiché non abbiamo ulteriori dati possiamo dire che il reddito medio dell'intera popolazione presa in esame è di £2.560.000.

In questo caso abbiamo effettuato una *stima puntuale* del parametro "reddito medio".

Nel seguente paragrafo tratteremo la cosiddetta *stima per intervallo* che più viene utilizzata perché maggiormente ricca di informazioni.

1.10 Stime per intervallo

Nei prossimi esempi l'obiettivo è quello di determinare un intervallo a cui "quasi sicuramente" appartiene il parametro della popolazione che solitamente è incognito.

Di solito si va alla ricerca dell'intervallo, chiamato *intervallo di confidenza*, che, con una probabilità uguale a 0.95 ovvero a 0.99, il parametro q vi appartenga. Questa probabilità p si chiama *grado di fiducia* mentre il valore 1 - p *margine di errore*.

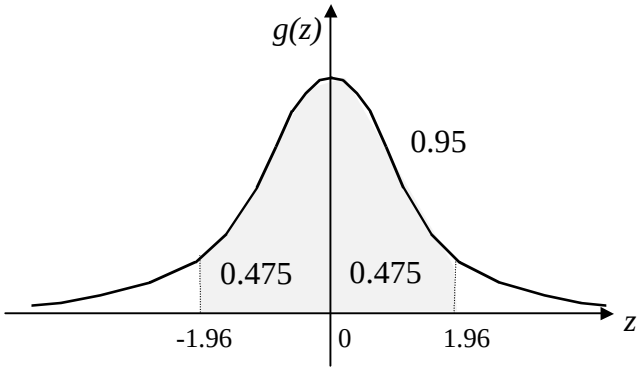
Quindi negli esercizi in cui, per esempio, si chiede di stimare la media m della popolazione con grado di fiducia del 95% avendo a disposizione un campione A, nell'ipotesi che esso abbia valore medio M(A) e scarto quadratico medio corretto $S_C^2(A)$, dobbiamo ricercare quell'intervallo contenente m, con una probabilità di 0.95.

A tal fine ricordiamo che quando i campioni sono "abbastanza numerosi" ($n \geq 30$), per il teorema del limite centrale,

◊ per la media X del campione A, possiamo utilizzare la distribuzione normale standardizzata per la variabile casuale

$$\frac{X(A) - m}{\frac{S_C(A)}{\sqrt{n}}}$$

Quindi in generale vogliamo determinare z_1 e z_2 in modo che $P\left(z_1 \leq \frac{X(A) - m}{\frac{S_C(A)}{\sqrt{n}}} \leq z_2 \right) = 0.95$



Per la simmetria della distribuzione normale risulta che l'area corrispondente a 0.95 equivale a un'area pari a 0.475 per parte.
Dalla tabella dell'area di probabilità sotto la curva normale standardizzata otteniamo che

$$F_0(z) = 0.475 \text{ per } z = 1.96.$$

Pertanto la precedente probabilità diventa $P\left(-1.96 \leq \frac{X(A) - m}{\frac{S_C(A)}{\sqrt{n}}} \leq 1.96 \right) = 0.95$

ossia

$$P\left(X(A) - 1.96\frac{S_C(A)}{\sqrt{n}} \le m \le X(A) + 1.96\frac{S_C(A)}{\sqrt{n}} \right) = 0.95$$

◊ anche per la frequenza campionaria X del campione A, possiamo utilizzare la distribuzione normale standardizzata per la variabile casuale

$$\frac{X(A) - m}{\sqrt{\frac{X(A)[1 - X(A)]}{n}}}$$

essendo, per **(F)**, lo scarto quadratico medio della frequenza.
Procedendo come sopra perveniamo a

$$P\left(X(A) - 1.96\sqrt{\frac{X(A)[1 - X(A)]}{n}} \le m \le X(A) + 1.96\sqrt{\frac{X(A)[1 - X(A)]}{n}} \right) = 0.95$$

Infine facciamo osservare che, nel caso in cui il grado di fiducia richiesto sia uguale a 0.99, il valore di 1.96 deve essere sostituito con 2.58. Questi valori di z che separano una zona da un'altra si chiamano *valori critici*.

ESEMPI

1) L'altezza media di 40 bambini di una scuola elementare è di 1.18m. Sapendo che lo scarto quadratico medio della distribuzione delle altezze di quella scuola è di 0.40m, stimate per intervallo l'altezza media dei bambini con un grado di fiducia uguale a 0.95.

   Poiché $n \ge 30$ allora possiamo ricorrere al teorema del limite centrale.
   In questo problema è noto lo scarto quadratico medio dell'intera popolazione, ossia, indicando con A il campione dei 40 bambini, abbiamo

$$X(A) = 1.18m \text{ e } s = 0.40m$$

   pertanto possiamo scrivere

$$P\left(1.18 - 1.96\frac{0.40}{\sqrt{40}} \le m \le 1.18 + 1.96\frac{0.40}{\sqrt{40}} \right) = 0.95$$

   da cui otteniamo che l'intervallo di confidenza al grado di fiducia uguale a 0.95 è

$$1.056 \le m \le 1.304 .$$

2) Il peso medio di 50 ragazzi di una scuola secondaria superiore è di 62kg ed uno scarto quadratico medio di 9kg. Stimate per intervallo l'altezza media dei bambini con un grado di fiducia uguale a 0.95.

   Poiché $n \ge 30$ allora possiamo ricorrere al teorema del limite centrale.
   In questo problema non è noto lo scarto quadratico medio dell'intera popolazione, pertanto, indicando con A il campione dei 50 ragazzi, abbiamo

$$X(A) = 62kg \text{ e } S(A) = 9kg$$

Lo scarto quadratico medio corretto è $S_C(A) = \sqrt{\frac{n}{n-1}} \cdot S(A) = \sqrt{\frac{50}{49}} \cdot 9 = 9.09 kg$.

Abbiamo quindi $P\left(62 - 1.96\frac{9.09}{\sqrt{50}} \le m \le 62 + 1.96\frac{9.09}{\sqrt{50}}\right) = 0.95$

da cui otteniamo l'intervallo di confidenza, al grado di fiducia uguale a 0.95

$$59.48 \le m \le 64.52 .$$

3) In un collegio elettorale si presentano alle elezioni due candidati. Da un sondaggio effettuato su un campione di 1200 persone è risultato che il primo candidato ha ottenuto una preferenza del 52.5%. Stimate per intervallo la percentuale dei voti che avrà il primo candidato con un grado di fiducia uguale a 0.95.

   Poiché $n \ge 30$ allora possiamo ricorrere al teorema del limite centrale.
   In tale caso si vuole conoscere la stima di una percentuale e non di una media.
   Indicando con A il campione composto da 1200 persone, abbiamo X(A) = 0.525. Da cui segue

$$S(A) = \sqrt{\frac{X(A) \cdot [1 - X(A)]}{n}} = \sqrt{\frac{0.525 \cdot 0.475}{1200}} = 0.01442.$$

   Abbiamo quindi

$$P(0.525 - 1.96 \cdot 0.01442 \le m \le 0.525 + 1.96 \cdot 0.01442) = 0.95$$

   da cui otteniamo l'intervallo di confidenza, al grado di fiducia uguale a 0.95

$$0.497 \le m \le 0.553 .$$

   Pertanto con una probabilità uguale a 0.95 la percentuale di preferenza del primo candidato riguardante l'intera popolazione è compresa tra il 49.7% e il 55.3%.

APPLICHIAMO

1) L'altezza media di 300 adulti di un dato paese è di 1.72*m*. Sapendo che lo scarto quadratico medio della distribuzione delle altezze del paese è di 0.62*m*, stimate per intervallo l'altezza media dei bambini con un grado di fiducia uguale a 0.99.
2) In un collegio elettorale si presentano alle elezioni due candidati. Da un sondaggio effettuato su un campione di 1500 persone è risultato che il primo candidato ha ottenuto una preferenza del 52.5%. Stimate per intervallo la percentuale dei voti che avrà il primo candidato con un grado di fiducia uguale a 0.99.
3) Da un mazzo di carte, di cui non si conosce la composizione, si estraggono in modo bernoulliano 73 carte ottenendo 12 figure e 61 non figure. Stimate per intervallo la percentuale di figure presenti nel mazzo con un grado di fiducia uguale a 0.95 prima, successivamente uguale a 0.99.
4) Da una scatola si estraggono in modo bernoulliano 39 pezzi rilevandone 5 difettosi. Stimate per intervallo la percentuale di pezzi difettosi presenti nella scatola con un grado di fiducia uguale a 0.95 prima, successivamente uguale a 0.99.
5) Una macchina produce dolcetti da vendere all'ingrosso. Sappiamo che lo scarto quadratico medio dell'intera "popolazione" è uguale a 3*g*. Stimate per intervallo il peso medio dei dolcetti con un grado di fiducia uguale a 0.95, sapendo che il peso medio di un campione di 48 dolcetti è risultato pari a 150*g*.

6) La lunghezza media di un campione di 50 viti prodotte da una ditta è risultato uguale a 15*cm* con scarto quadratico medio uguale a 0.03*cm*. Stimate per intervallo la lunghezza media dell'intera produzione con un grado di fiducia uguale a 0.99.

1.11 Test delle ipotesi della media e della frequenza

In genere ogni indagine di tipo statistico serve per verificare un'ipotesi che può essere rafforzata o meno dai risultati. Pertanto qualsiasi esperimento comporta la formulazione di un'ipotesi di lavoro che si sottopone a verifica, chiamata *ipotesi nulla* e indicata con H_0. L'ipotesi alternativa viene indicata con H_1.

Se q è un parametro della popolazione, l'ipotesi nulla afferma che la stima trovata da un campione non differisce "sostanzialmente" da q, mentre l'ipotesi alternativa consiste nel ritenere che la differenza è tanto grande da rifiutare H_0.

Il margine di errore viene chiamato anche *livello di significatività*. Esso indica quindi la probabilità di sbagliare ed è uguale a 1 - *p*, essendo *p* il grado di fiducia.

ESEMPI

1) Una ditta produce delle viti che dovrebbero avere una lunghezza di 8*cm*. Da un campione di 75 viti è stata registrata una lunghezza media di 7.8*cm* con scarto quadratico medio di 0.9*cm*. Verificare l'ipotesi, al livello di significatività uguale a 0.05, che la differenza riscontrata non sia significativa.

   I dati sono

   $$X(A) = 7.8,\ S(A) = 0.9,\ n = 75,$$

   essendo A il campione.

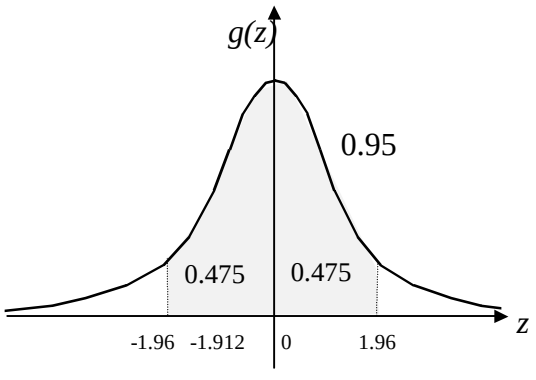
   Formuliamo le seguenti ipotesi

   H_0: $m = 8$, con livello di significatività uguale a 0.05,
   H_1: $m \neq 8$, con livello di significatività uguale a 0.05.

   Poiché $n \geq 30$ allora possiamo ricorrere al teorema del limite centrale.

   Lo scarto quadratico medio corretto è $S_C(A) = \sqrt{\frac{n}{n-1}} \cdot S(A) = \sqrt{\frac{75}{74}} \cdot 0.9 = 0.9061$, quindi la variabile standardizzata è $z = \frac{7.8 - 8}{\frac{0.9061}{\sqrt{75}}} \approx -1.912$

   Poiché ci viene richiesto un livello di significatività uguale a 0.05 allora z deve essere tale che $P(-1.96 \leq z \leq 1.96) = 0.95$. Pertanto l'ipotesi nulla è accettata se $-1.96 \leq z \leq 1.96$, ossia se $|z| \leq 1.96$.
   Dal momento che abbiamo trovato z = -1.912, accettiamo H_0, con livello di significatività uguale a 0.05.



L'intervallo [-1.96,1.96] è detto *zona di accettazione dell'ipotesi nulla* mentre la parte restante della retta reale si chiama *zona di rifiuto*.
Questo esempio è un caso di *test a due code*, ossia di test nel quale la zona di rifiuto è formata da due parti simmetriche rispetto all'asse delle ordinate.

2) Una ditta produce delle lampadine che dovrebbero avere una durata media almeno di 750 ore. Per un campione di 100 pezzi risulta una durata media di 747 ore con scarto quadratico medio pari a 12 ore. A livello di significatività uguale a 0.05 dite se la produzione è sotto controllo.

I dati sono

$$X(A) = 747,\ S(A) = 12,\ n = 100,$$

essendo A il campione.

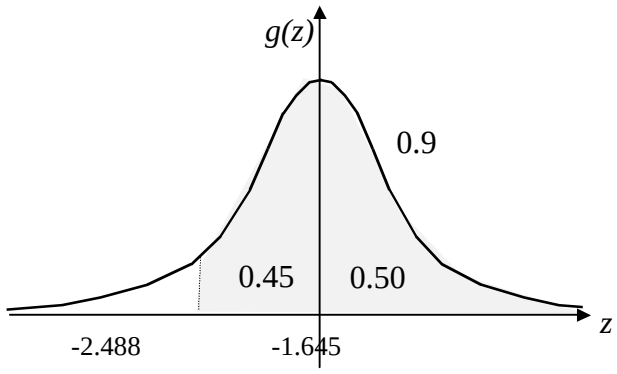
Formuliamo le seguenti ipotesi

H_0: $m = 750$, con livello di significatività uguale a 0.05,
H_1: $m < 750$, con livello di significatività uguale a 0.05.

Poiché $n \geq 30$ allora possiamo ricorrere al teorema del limite centrale.

Lo scarto quadratico medio corretto è $S_C(A) = \sqrt{\frac{n}{n-1}} \cdot S(A) = \sqrt{\frac{100}{99}} \cdot 12 = 12.06$, quindi la variabile standardizzata è

$$z = \frac{747 - 750}{\frac{12.06}{\sqrt{100}}} \approx -2.488$$



L'ipotesi nulla viene accettata se $z \geq -1.645$. Infatti nella tabella troviamo come valore critico 1.645, ossia risulta $F_0(1.645) \approx 0.45$. Quindi l'ipotesi H_0 è rifiutata, con livello di significatività uguale a 0.05.

L'intervallo [-1.645,+∞ [è la *zona di accettazione dell'ipotesi nulla* mentre la parte restante della retta reale è *zona di rifiuto*.
Questo esempio è un caso di *test a una coda*, ossia di test nel quale la zona di rifiuto è formata da una sola parte. In tali casi si chiede se il parametro è maggiore o minore di un dato valore. Nel caso in cui il livello di significatività invece che uguale a 0.05 (*probabile significatività*) è di 0.01 (*alta significatività)* abbiamo che il valore critico è, a seconda dei casi, ± 2.33

3) Supponiamo di effettuare 300 lanci di un dado e di aver osservato che il 6 è uscito 34 volte. Verifichiamo l'ipotesi secondo cui il dado, almeno per l'uscita del 6, non è truccato, con un livello di significatività uguale a 0.01.

I dati sono

$$X(A) = \frac{34}{300} = 0.113,\ S(A) = \sqrt{\frac{X(A) \cdot [1 - X(A)]}{n}} = \sqrt{\frac{0.113 \cdot 0.887}{300}} = 0.0183,\ n = 100,$$

essendo A il campione costituito dai 300 lanci.

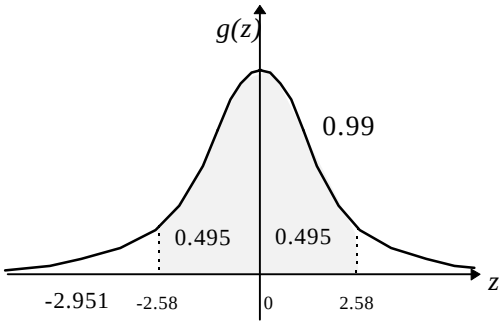
Formuliamo allora le seguenti ipotesi

H_0: $m = \frac{1}{6} = 0.167$, con livello di significatività uguale a 0.01,

H_1: $m \neq \frac{1}{6}$, con livello di significatività uguale a 0.01.

Poiché $n \geq 30$ allora possiamo ricorrere al teorema del limite centrale.

La variabile standardizzata è $z = \frac{0.113 - 0.167}{0.0183} \approx -2.951$



Poiché ci viene richiesto un livello di significatività uguale a 0.01 allora z deve essere tale che $P(-2.58 \leq z \leq 2.58) = 0.99$. Pertanto l'ipotesi nulla è accettata se $-2.58 \leq z \leq 2.58$, ossia se $|z| \leq 2.58$.

Dal momento che abbiamo trovato z = -2.951, rifiutiamo H_0, con livello di significatività uguale a 0.01.

La zona di accettazione è [-2.58,2.58]. Anche in questo esempio è un caso di *test a due code*.

Nella formulazione delle decisioni si possono commettere due tipi di errore.

Errore di prima specie	quando l'ipotesi H_0 è vera, ma è stata rifiutata. Questo accade perché si è riscontrata una differenza significativa tra il parametro della popolazione e la stima del campione.
Errore di seconda specie	quando l'ipotesi H_0 è falsa, ma è stata accettata. Questo accade perché non è stata riscontrata una differenza significativa tra il parametro della popolazione e la stima del campione.

Si dovrebbe cercare di ridurre il rischio di commettere entrambi i tipi di errore, ma riducendo la probabilità di compiere un tipo di errore aumenta la probabilità di commettere l'altro. A tal fine potremmo ottenere una riduzione di entrambi solo aumentando la dimensione del campione.

Il livello di significatività p indica la probabilità di commettere l'errore di prima specie, ossia la probabilità di rifiutare l'ipotesi H_0 quando dovrebbe essere accettata.

Nel primo esempio l'errore che si potrebbe commettere è di seconda specie.
Nel secondo e terzo esempio invece l'errore che si potrebbe commettere è di prima specie.

APPLICHIAMO

1) La durata media di un campione di 270 pile prodotte da una ditta è di 970 ore con una deviazione standard di 91 ore. Sottoponete a test l'ipotesi m = 982 ore contro l'ipotesi alternativa $m \neq 982$ ore, usando un livello di significatività dello 0.05 prima, dello 0.01 poi.
2) Considerando l'esercizio precedente, sottoponete a test l'ipotesi m = 982 ore contro l'ipotesi alternativa $m < 982$ ore, usando un livello di significatività dello 0.05 prima, dello 0.01 poi.
3) Una ditta costruttrice di cavi garantisce una possibilità di carico di 2000*kg*. Proviamo 84 cavi ottenendo un carico di rottura medio di 1990*kg* e uno scarto quadratico medio di 120*kg*.

Verificare l'ipotesi, al livello di significatività uguale a 0.01, che la differenza riscontrata non sia significativa.

4) Una casa farmaceutica asserisce che un suo prodotto è in grado di guarire una data malattia nel 83% dei casi. In un campione di 70 pazienti che soffrono di questa malattia la medicina ne ha guariti 54. Determinate, con un livello di significatività uguale a 0.05, se la casa farmaceutica può asserire la precedente affermazione legittimamente.
5) Una individuo asserisce di possedere dei poteri paranormali. In un esperimento consistente nell'indovinare l'uscita nel lancio di una moneta egli indovina 52 volte su 80. Determinate, con un livello di significatività uguale a 0.01, se l'individuo ha reali poteri oppure ha semplicemente indovinato.

1.12 Test delle ipotesi della differenza fra medie e tra frequenze

♦ Supponiamo di avere due popolazioni aventi media m_1 e m_2 e scarto quadratico medio s_1 e s_2. Per ciascun campione di dimensione n_1 preso dalla prima popolazione consideriamo la media, ottenendo la media campionaria X_1. Per essa risulta che $M(X_1) = m_1$ e $V(X_1) = \frac{s_1^2}{n_1}$.

Analogamente, per ciascun campione di dimensione n_2 preso dalla seconda popolazione consideriamo la media, ottenendo la media campionaria X_2. Per essa risulta che $M(X_2) = m_2$ e $V(X_2) = \frac{s_2^2}{n_2}$.

Possiamo allora considerare la distribuzione campionaria delle differenze delle medie

$$X_1 - X_2,$$

determinando $M(X_1 - X_2)$ e $V(X_1 - X_2)$.

Abbiamo

$$M(X_1 - X_2) = M(X_1) - M(X_2),$$

per le proprietà A) e G) del valore medio e

$$V(X_1 - X_2) = V(X_1) + V(X_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2},$$

per la proprietà G) della varianza.

Pertanto la variabile standardizzata è

$$Z = \frac{(X_1 - X_2) - (m_1 - m_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

La distribuzione può essere approssimata da quella normale in cui n_1 ed n_2 sono entrambi maggiori o uguali di 30.

Nella verifica delle ipotesi l'ipotesi H_0 consiste nel ritenere che non c'è differenza tra le medie delle due popolazioni m_1 e m_2. Di conseguenza la variabile standardizzata diventa

$$Z = \frac{X_1 - X_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

ESEMPIO

Supponiamo che uno stesso compito di matematica sia stato dato ai ragazzi appartenenti alle classi quinte di due istituti scolastici diversi. Sono stati estratti da questi due campioni A_1 ed A_2. I dati ottenuti sono i seguenti:
$n_1 = 35$, $X(A_1) = 6.2$ ed $S_c(A_1) = 0.9$,
$n_2 = 39$, $X(A_2) = 6.6$ ed $S_c(A_2) = 1.4$.
Esiste una differenza significativa fra il profitto delle quinte delle due scuole, con un livello di significatività uguale a 0.05?

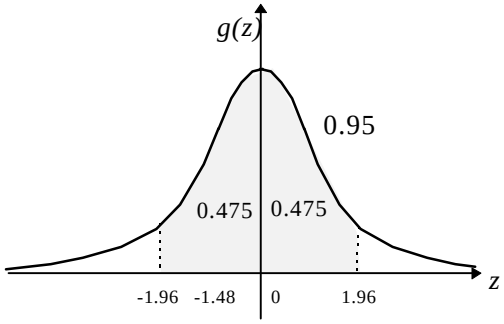
Detti m_1 e m_2 i voti medi di matematica delle classi quinte dei due istituti, formuliamo le seguenti ipotesi, formuliamo le seguenti ipotesi

H_0: $m_1 = m_2$, con livello di significatività uguale a 0.05,
H_1: $m_1 \neq m_2$, con livello di significatività uguale a 0.05.

Poiché la dimensione dei campioni è maggiore di 30, possiamo ricorrere al teorema del limite centrale.

$$z = \frac{6.2 - 6.6}{\sqrt{\frac{0.9^2}{35} + \frac{1.4^2}{39}}} \approx -1.48$$

Poiché ci viene richiesto un livello di significatività uguale a 0.05 allora z deve essere tale che $P(-1.96 \leq z \leq 1.96) = 0.95$. Pertanto l'ipotesi nulla è accettata se $-1.96 \leq z \leq 1.96$, ossia se $|z| \leq 1.96$.



Dal momento che abbiamo trovato z = -1.48, accettiamo H_0, con livello di significatività uguale a 0.05.

- Supponiamo di avere due popolazioni aventi frequenze relative riferite un dato attributo o evento m_1 e m_2. Per ciascun campione di dimensione n_1 preso dalla prima popolazione consideriamo la frequenza, ottenendo così la frequenza campionaria X_1. Per essa risulta

$$M(X_1) = m_1 \text{ e } V(X_1) = \frac{m_1(1 - m_1)}{n_1}.$$

  Analogamente, per ciascun campione di dimensione n_2 preso dalla seconda popolazione consideriamo la frequenza, ottenendo la frequenza relativa campionaria X_2. Per essa risulta

$$M(X_2) = m_2 \text{ e } V(X_2) = \frac{m_2(1 - m_2)}{n_2}.$$

  Possiamo allora considerare la distribuzione campionaria delle differenze delle frequenze relative

$$X_1 - X_2,$$

  determinando $M(X_1 - X_2)$ e $V(X_1 - X_2)$.

  Abbiamo

$$M(X_1 - X_2) = M(X_1) - M(X_2),$$

  per le proprietà A) e G) del valore medio e

$$V(X_1 - X_2) = V(X_1) + V(X_2) = \frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2},$$

per la proprietà G) della varianza.

Pertanto la variabile standardizzata è

$$Z = \frac{(X_1 - X_2) - (m_1 - m_2)}{\sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}}$$

La distribuzione può essere approssimata da quella normale nel caso in cui n_1 ed n_2 sono entrambi maggiori o uguali di 30.

Nella verifica delle ipotesi l'ipotesi H_0 consiste nel ritenere che non c'è differenza tra le medie delle due popolazioni m_1 e m_2. Poniamo $m_1 = m_2 = m$ e consideriamo la media ponderata delle frequenze m_1 e m_2 al fine di stimare la frequenza della popolazione,

$$m = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}$$

Di conseguenza la variabile standardizzata diventa

$$Z = \frac{(X_1 - X_2)}{\sqrt{m(1-m)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

ESEMPIO

Viene testata la capacità di un medicinale di curare un tipo di allergia utilizzando due campioni A e B, formati rispettivamente da 76 e 59 persone affetti da allergia cronica. Il medicinale viene somministrato alle persone del campione A ma non a quelle di B (questo campione viene chiamato *campione di controllo*). Si riscontra che nei due campioni A e B guariscono dall'allergia 60 e 40 persone rispettivamente. Sottoponiamo a test l'ipotesi secondo cui il medicinale cura l'allergia, con livello di significatività uguale a 0.01.

Dette m_1 e m_2 rispettivamente le frequenze relative di guarigioni della popolazioni a cui è stato somministrato il medicinale e a quella di controllo, formuliamo le seguenti ipotesi

H_0: $m_1 = m_2$, con livello di significatività uguale a 0.01, ossia le differenze riscontrate sono causate da fluttuazioni casuali.
H_1: $m_1 > m_2$, con livello di significatività uguale a 0.01, ossia il medicinale influenza positivamente la guarigione.

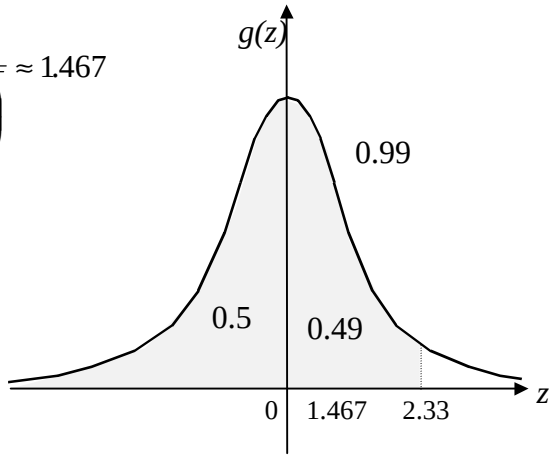
Poiché la dimensione dei campioni è maggiore di 30, possiamo ricorrere al teorema del limite centrale.

Stimiamo come frequenza della popolazione la media ponderata delle due frequenze relative

$$m = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2} = \frac{76\left(\frac{60}{76}\right) + 59\left(\frac{40}{59}\right)}{76 + 59} = \frac{20}{27}$$

La variabile casuale standardizzata è

$$z = \frac{\frac{60}{76} - \frac{40}{59}}{\sqrt{\frac{20}{27} \cdot \frac{7}{27}\left(\frac{1}{76} + \frac{1}{59}\right)}} \approx 1.467$$



Poiché ci viene richiesto un livello di significatività uguale a 0.01 allora z deve essere tale che $P(z \le 2.33) = 0.99$. Pertanto l'ipotesi nulla è accettata se $z \le 2.33$.
Dal momento che abbiamo trovato z = 1.467, accettiamo H_0, con livello di significatività uguale a 0.01. Pertanto la medicina non cura in modo significativo l'allergia.

APPLICHIAMO

1) Supponiamo che uno stesso compito di matematica sia stato dato ai ragazzi appartenenti alle classi quinte di due istituti scolastici diversi. Sono stati estratti da questi due campioni A_1 ed A_2. I dati ottenuti sono i seguenti:
   n_1 = 42, $X(A_1)$ = 5.9 ed $S_c(A_1)$ = 2.1,
   n_2 = 41, $X(A_2)$ = 6.1 ed $S_c(A_1)$ = 0.9.
   Esiste una differenza significativa fra il profitto delle quinte delle due scuole, con un livello di significatività uguale a 0.01?
2) Due campioni di 220 pezzi costruiti da una ditta e di 128 pezzi costruiti da una seconda ditta hanno mostrato rispettivamente 20 e 11 pezzi difettosi.
   Esiste una differenza significativa fra l'affidabilità delle due ditte, con un livello di significatività uguale a 0.05?
3) Con un campione di 320 pile prodotte da una ditta abbiamo ottenuto una durata media pari a 1300 ore con scarto quadratico medio di 120 ore. Con un campione di 290 pile prodotte da una seconda ditta abbiamo ottenuto una durata media pari a 1250 ore con scarto quadratico medio di 130 ore. Esiste una differenza fra le durate delle pile al livello di significatività dello 0.05? E dello 0.01?
4) Due campioni di 100 pile costruiti da una ditta e di 120 pile costruiti da una seconda ditta hanno mostrato rispettivamente 5 e 6 pile non funzionanti.
   Esiste una differenza significativa fra l'affidabilità delle due ditte di pile, con un livello di significatività uguale a 0.05?

1.13 Test delle ipotesi nel confronto tra distribuzioni

Sappiamo che se nX è la *frequenza assoluta campionaria*, ossia la variabile casuale che ad ogni campione di dimensione n associa il numero di volte che in n prove si è verificato un dato evento,

allora la variabile casuale standardizzata $Z = \frac{nX - nm}{\sqrt{nm(1-m)}}$, per n tendente all'infinito, tende a distribuirsi come la variabile casuale avente come funzione di densità di probabilità N(0,1). A seconda dei casi, m della formula precedente è la frequenza relativa o la probabilità che si verifichi l'evento.

Quadrando l'uguaglianza precedente possiamo scrivere $Z^2 = \frac{(nX - nm)^2}{nm(1-m)}$.

TEOREMA

Ponendo $p = m$, $q = 1 - m$, $X_1 = X$ e $X_2 = 1 - X$ abbiamo che

$$\frac{(nX - np)^2}{npq} = \frac{(nX_1 - np)^2}{np} + \frac{(nX_2 - nq)^2}{nq}$$

Dimostrazione

Risulta

$$\frac{(nX_1 - np)^2}{np} + \frac{(nX_2 - nq)^2}{nq} = \frac{q(nX_1 - np)^2 + p(nX_2 - nq)^2}{npq} = \frac{q(nX - np)^2 + p[n(1-X) - n(1-p)]^2}{npq} =$$

$$\frac{q(nX - np)^2 + p(n - nX - n + np)^2}{npq} = \frac{q(nX - np)^2 + p(np - nX)^2}{npq} = \frac{q(nX - np)^2 + p(nX - np)^2}{npq} =$$

$$\frac{(q + p)(nX - np)^2}{npq} = \frac{(nX - np)^2}{npq}$$

c.v.d.

Nell'espressione

$$Z^2 = \frac{(nX_1 - np)^2}{np} + \frac{(nX_2 - nq)^2}{nq}$$

X_1 è la variabile casuale associata alla frequenza relativa dell'evento di frequenza relativa o probabilità p, mentre X_2 è la variabile casuale associata alla frequenza relativa dell'evento contrario, di frequenza relativa o probabilità $q = 1 - p$.
Z^2 fornisce una misura della differenza fra le frequenze osservate nel campione e quelle probabili.

Supponiamo ora di avere k eventi E_1, E_2, ..., E_k, con le rispettive probabilità o frequenze relative p_1, p_2, ..., p_k.
Da un campione A di dimensione n otteniamo, per gli eventi precedenti, le frequenze relative $X_1(A)$, $X_2(A)$, ..., $X_k(A)$.

L'espressione sopra riportata di Z^2 ne suggerisce una generalizzazione al caso in cui gli eventi considerati sono più di due. Pertanto anche le seguente espressione

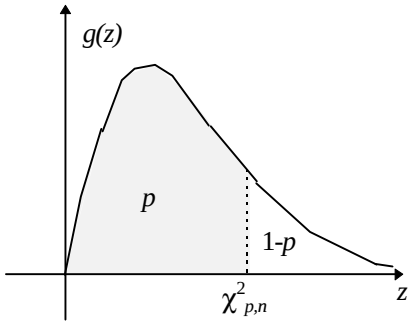
$$c^2 = \frac{\left(n\mathrm{X}_1 - np_1\right)^2}{np_1} + \frac{\left(n\mathrm{X}_2 - np_2\right)^2}{np_2} + \cdots + \frac{\left(n\mathrm{X}_k - np_k\right)^2}{np_k} = \sum_{i=1}^{k} \frac{\left(n\mathrm{X}_i - np_i\right)^2}{np_i}$$

fornisce una misura della differenza fra le frequenze osservate nel campione e quelle probabili dei *k* eventi.

Nel caso in cui $c^2 = 0$, non esiste alcuna differenza fra le frequenze osservate e quelle probabili, mentre in generale $c^2 > 0$ e quanto maggiore è il suo valore tanto più grande è la discordanza tra frequenze osservate e quelle probabili.

Si dimostra che la distribuzione di c^2 tende, per *n* tendente all'infinito (in pratica basta che le np_i siano maggiori di 5) , alla distribuzione *chi-quadrato* la cui funzione di densità è

$$g(z) = \begin{cases} \dfrac{1}{G \cdot 2^{n/2}} z^{(n/2)-1} e^{-z/2} & \text{se } z > 0 \\ 0 & \text{se } z \le 0 \end{cases}$$ ed il diagramma



G è una costante dipendente da *n*.

Esiste una tabella di distribuzione in cui sono riportati valori numerici in corrispondenza dei diversi livelli di significatività 1 - *p* e dei *gradi di libertà* *n*.

Se troviamo che $c^2 < c^2_{p,n}$, concludiamo che non esiste differenza significativa tra le frequenze osservate e quelle probabili.

ESEMPIO

Ad un campione di 2000 persone è stato chiesto di scegliere la marca di pasta preferita tra quattro marche. Indicando con E_1, E_2, E_3 ed E_4 le quattro marche, l'indagine ha dato i seguenti risultati.

marca	E_1	E_2	E_3	E_4
n° preferenze osservate: $n\mathrm{X}_i$	480	510	496	514

Tabelle come la precedente sono chiamate *tabelle di classificazione ad una via* perché le *k* frequenze osservate occupano una sola riga.
Possiamo affermare, con un livello di significatività uguale a 0.05, che la maggioranza delle persone preferisce la marca E_4?

Formuliamo l'ipotesi

H_0: con livello di significatività uguale a 0.05 le differenze riscontrate sono causate da fluttuazioni casuali, pertanto non c'è differenza significativa nelle preferenze tra le quattro marche,

Se non esistesse alcuna differenza fra le quattro marche si dovrebbero avere le seguenti frequenze probabili

marca	E_1	E_2	E_3	E_4
n° preferenze teoriche: np_i	500	500	500	500

Abbiamo pertanto la seguente tabella,

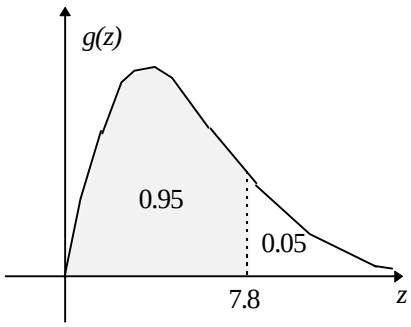
marca	E_1	E_2	E_3	E_4
n° preferenze osservate: $n\mathrm{X}_i$	480	510	496	514
n° preferenze teoriche: np_i	500	500	500	500
$\frac{\left(n\mathrm{X}_i - np_i\right)^2}{np_i}$	0.8	0.2	0.032	0.392

Otteniamo quindi $c^2 = \sum_{i=1}^{4} \frac{\left(n\mathrm{X}_i - np_i\right)}{np_i} = 1.424$.

In questo caso, poiché ci è chiesto un livello di significatività dello 0.05, p = 0.95.

Nell'esempio il numero di gradi di libertà è uguale a 3, ossia n = 3. Questo perché se conosciamo 3 frequenze probabili l'ultima frequenza può essere determinata dalle prime tre e dalla dimensione campionaria, uguale a 2000.

Nella tabella leggiamo $c^2_{0.95,3} = 7.81$, avendo quindi una situazione grafica simile a quella riportata a fianco.



Accettiamo l'ipotesi H_0 se $c^2 < c^2_{0.95,3}$.

Poiché 1.424 < 7.81 concludiamo che non c'è una preferenza per una marca, accettando l'ipotesi H_0.

Possiamo generalizzare considerando tabelle di classificazione a più vie nelle quali le frequenze osservate occupano *h* righe e *k* colonne. In questi casi le tabelle sono chiamate *di contingenza*.

Per fissare le idee consideriamo una tabella di contingenza 2×3.

	C_1	C_2	C_3
R_1	$n\mathrm{X}_{11}$	$n\mathrm{X}_{12}$	$n\mathrm{X}_{13}$
R_2	$n\mathrm{X}_{21}$	$n\mathrm{X}_{22}$	$n\mathrm{X}_{23}$

le somme delle frequenze di ciascuna riga e di ciascuna colonna si chiamano *frequenze marginali*. Nel nostro caso esse sono

$$n_{\mathrm{R1}} = n\mathrm{X}_{11} + n\mathrm{X}_{12} + n\mathrm{X}_{13}; \qquad n_{\mathrm{R2}} = n\mathrm{X}_{21} + n\mathrm{X}_{22} + n\mathrm{X}_{23};$$
$$n_{\mathrm{C1}} = n\mathrm{X}_{11} + n\mathrm{X}_{21}; \qquad n_{\mathrm{C2}} = n\mathrm{X}_{12} + n\mathrm{X}_{22}; \qquad n_{\mathrm{C3}} = n\mathrm{X}_{13} + n\mathrm{X}_{23}.$$

In genere in una tabella di contingenza sono inserite le frequenze marginali. Ovviamente la somma delle frequenze è uguale ad n che rappresenta la dimensione del campione.

	C_1	C_2	C_3	
R_1	nX_{11}	nX_{12}	nX_{13}	n_{R1}
R_2	nX_{21}	nX_{22}	nX_{23}	n_{R2}
	n_{C1}	n_{C2}	n_{C3}	n

Anche in questi casi alla tabella di sopra si associa una tabella composta dalle frequenze probabili.

	C_1	C_2	C_3	
R_1	np_{11}	np_{12}	np_{13}	n_{R1}
R_2	np_{21}	np_{22}	np_{23}	n_{R2}
	n_{C1}	n_{C2}	n_{C3}	n

I numeri $\frac{n_{R1}}{n}$ ed $\frac{n_{C1}}{n}$ rappresentano rispettivamente la frequenza relativa dell'attributo R1 e C1. Per il teorema della probabilità composta $\frac{n_{R1}}{n} \cdot \frac{n_{C1}}{n}$ è la frequenza relativa ai due attributi congiunti R1 e C1. Pertanto la frequenza assoluta teorica dei due attributi congiunti R1 e C1 è $n \cdot \frac{n_{R1}}{n} \cdot \frac{n_{C1}}{n} = \frac{n_{R1} \cdot n_{C1}}{n}$, pertanto abbiamo che $np_{11} = \frac{n_{R1} \cdot n_{C1}}{n}$. Analogamente per le restanti frequenze teoriche, abbiamo che

$$np_{ij} = \frac{n_{Ri} \cdot n_{Cj}}{n}.$$

Anche nel caso di tabelle di contingenza, per studiare la differenza fra le frequenze osservate nel campione e quelle probabili si ricorre all'espressione

$$c^2 = \frac{\left(nX_{11} - np_{11}\right)^2}{np_{11}} + \frac{\left(nX_{12} - np_{12}\right)^2}{np_{12}} + \cdots + \frac{\left(nX_{1k} - np_{1k}\right)^2}{np_{1k}} + \cdots + \frac{\left(nX_{h1} - np_{h1}\right)^2}{np_{h1}} + \frac{\left(nX_{h2} - np_{h2}\right)^2}{np_{h2}} + \cdots + \frac{\left(nX_{hk} - np_{hk}\right)^2}{np_{hk}}$$

ossia

$$c^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{\left(nX_{ij} - np_{ij}\right)^2}{np_{ij}}$$

Passiamo ora ad esaminare come determinare il numero di gradi di libertà n nel caso di tabelle di contingenza $h \times k$.

	C_1	C_2	$\cdots$	C_k	
R_1	$*$	$*$	$\cdots$	$*$	n_{R1}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
R_h	$*$	$*$	$\cdots$	$*$	n_{Rh}
	n_{C1}	n_{C2}	$\cdots$	n_{Ck}	n

Se inseriamo nella tabella precedente (h - 1) (k - 1) numeri i restanti sono univocamente determinati dal momento che le frequenze marginali ne vincolano il valore numerico. Quindi possiamo scrivere

$$n = (k-1)(h-1)$$

ESEMPIO

Nella tabella seguente sono riportate le distribuzioni dei voti (in centesimi) di 850 ragazzi, suddivise per sesso, secondo un'indagine.

sesso \ voto	[60,70[	[70,75[	[75,80[	[80,90[	[90,100]
maschi	45	87	126	110	54
femmine	51	67	130	120	60

Verifichiamo l'ipotesi H_0 : non c'è differenza significativa dello 0.05 tra le distribuzioni relative ai voti dei maschi e quelli delle femmine.

Completiamo la precedente tabella con le frequenze assolute marginali

sesso \ voto	[60,70[	[70,75[	[75,80[	[80,90[	[90,100]	
maschi	45	87	126	110	54	422
femmine	51	67	130	120	60	428
	96	154	256	230	114	850

e mostriamo la tabella delle frequenze teoriche

sesso \ voto	[60,70[	[70,75[	[75,80[	[80,90[	[90,100]	
maschi	47.66	76.46	127.1	114.19	56.60	422
femmine	48.34	77.54	128.90	115.81	57.40	428
	96	154	256	230	114	850

In tale esempio abbiamo

$$c^2 = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{\left(n\mathrm{X}_{ij} - np_{ij}\right)^2}{np_{ij}} = \frac{(45-47.66)^2}{47.66} + \frac{(87-76.46)^2}{76.46} + \cdots + \frac{(60-57.40)^2}{57.40} = 3.7419$$

Sapendo che i gradi di libertà sono n = (2-1)·(5-1) = 4, troviamo $c^2_{0.95,4} = 9.49$ e, dal momento che $c^2 < c^2_{0.95,4}$, accettiamo l'ipotesi H_0.

Se la richiesta fosse stata la seguente:

verificare l'ipotesi H_0 : non c'è differenza significativa dello 0.1 tra le distribuzioni relative ai voti dei maschi e quelli delle femmine,

avremmo ugualmente accettato l'ipotesi nulle in quanto $c^2_{0.90,4} = 7.78$.

APPLICHIAMO

1) Due campioni di 220 pezzi costruiti da una ditta e di 128 pezzi costruiti da una seconda ditta hanno mostrato rispettivamente 20 e 11 pezzi difettosi.
   Esiste una differenza significativa fra l'affidabilità delle due ditte, con un livello di significatività uguale a 0.05? *Utilizzate il test del chi-quadrato*.
2) I pazienti di due distinti campioni sono stati sottoposti a indagine statistica per studiare l'effetto di un vaccino anti-influenzale. Dei 340 pazienti del primo campione che sono stati vaccinati solo 3 si sono influenzati, mentre dei 240 pazienti del secondo campione che non sono stati vaccinati, 5 si sono influenzati.
   Esiste una differenza significativa fra i risultati ottenuti dai due campioni, con un livello di significatività uguale a 0.05?
3) Siano A un campione formato da 90 ragazzi e B da 85 ragazzi. Da un'indagine risulta che dei due campioni leggono quotidianamente un giornale rispettivamente 20 e 29 ragazzi. Stabilite se c'è differenza significativa tra A e B rispetto alla lettura dei giornali, a livello di significatività dello 0.05 e dello 0.01.
4) Un preside ha effettuato su due gruppi di ragazzi della sua scuola un'indagine per verificare se l'insegnamento della matematica fatto alle prime due ore o alla quarta e quinta ora influenzasse il loro apprendimento. Dopo alcuni mesi di prova è stata somministrata una verifica oggettiva e si sono osservati i seguenti risultati

	insufficiente	sufficiente	buono
1°-2° ora	15	23	13
4°-5° ora	10	32	10

   Supponendo che i ragazzi sono stati estratti casualmente e con un metodo di campionamento corretto, valutate c'è differenza significativa tra le due distribuzioni a livello di significatività dello 0.05.
5) Nella seguente tabella sono indicati i numeri degli studenti promossi e bocciati in tre differenti istituti tecnici.

	Istituto A	Istituto B	Istituto C
promossi	432	354	567
bocciati	35	25	40

   Verificate se c'è differenza significativa tra le percentuali di bocciati nei tre istituti con livello di significatività dello 0.05.
6) È stato condotto un sondaggio d'opinione su un campione di 120 ragazzi per stabilire se preferissero o meno un complesso rock. Si sono osservati i dati di seguito riportati

	preferenza	non preferenza
ragazzi	43	21
ragazze	46	10

   Verificate se c'è differenza significativa tra le preferenze dei ragazzi e quelle delle ragazze con livello di significatività dello 0.01.

1.14 Test delle ipotesi nel caso di piccoli campioni

Solitamente nel caso in cui la dimensione campionaria non supera le 30 unità si ricorre alla distribuzione *t di Student* (dallo pseudonimo dello statistico inglese William Gosset).
Tale distribuzione per *n* grande tende ad essere uguale a quella normale.
La funzione di densità è

$$f(t) = \mathrm{H}\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

H è una costante dipendente da *n* e il diagramma è molto simile a quello della funzione di densità della distribuzione normale.

Anche in tali casi esiste una tabella di distribuzione in cui sono riportati valori numerici in corrispondenza dei diversi livelli di significatività 1 - *p* e dei *gradi di libertà n*, ossia del numero dei valori che possono essere diversi senza che varino i valori dei parametri da sottoporre a stima.

- **Media**

  Se vogliamo sottoporre a test l'ipotesi H_0 che una popolazione ha media uguale a *m*, effettuiamo la consueta standardizzazione

$$T = \frac{X - m}{\frac{S_C}{\sqrt{n}}}$$

  e procediamo come nella distribuzione normale utilizzando però la tabella dell'area di probabilità sotto la curva della funzione di densità di probabilità della distribuzione *t di Student*, osservando che, se *n* è la dimensione del campione, allora il numero di gradi di libertà *n* è pari a *n* - 1.

  ESEMPIO

  Una ditta produce dei pezzi metallici che dovrebbero avere una lunghezza di 10*cm*. Da un campione di 12 viti è stata registrata una lunghezza media di 11*cm* con scarto quadratico medio corretto di 0.6*cm*. Verificare l'ipotesi, al livello di significatività uguale a 0.05, che la differenza riscontrata non sia significativa.

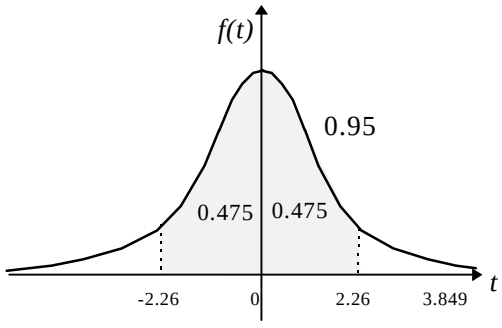
  I dati sono

$$X(A) = 11,\ S_C(A) = 0.6,\ n = 12,$$

  essendo A il campione.

  Formuliamo le seguenti ipotesi

  H_0: $m = 10$, con livello di significatività uguale a 0.05,
  H_1: $m \neq 10$, con livello di significatività uguale a 0.05.



  Ci troviamo davanti ad un test a due code.
  Il valore della variabile T corrispondente al campione dato è $t = \frac{11-10}{\frac{0.9}{\sqrt{12}}} = 3.849$, essendo i gradi di libertà pari a $n = 12 - 1 = 11$.
  Dalla tabella troviamo che all'intervallo [-2.26 e 2.26] corrisponde un'area pari a 0.95, pertanto accettiamo l'ipotesi nulla se il valore di *t* è ivi compreso.
  Poiché però *t* = 3.849 rigettiamo l'ipotesi H_0.

- **Differenza fra medie**

  Consideriamo due campioni aventi dimensione, media e varianza corretta, n_1, X_1 ed S_{C1}^2, n_2, X_2 ed S_{C2}^2 rispettivamente.
  Ripetendo le osservazioni fatte nel caso della differenza delle medie di due campioni di distribuzioni normali, troviamo la variabile standardizzata

$$T = \frac{X_1 - X_2}{\sqrt{\frac{S_{C1}^2}{n_1} + \frac{S_{C2}^2}{n_2}}}$$

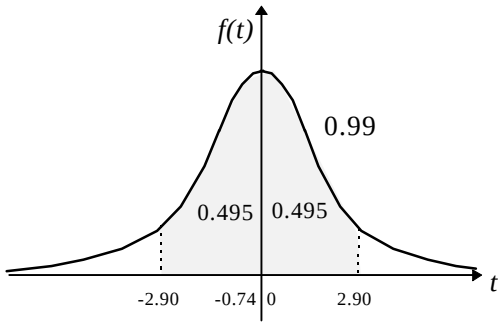
Procediamo quindi come nella distribuzione normale utilizzando la tabella dell'area di probabilità sotto la curva della funzione di densità di probabilità della distribuzione*t di Student*. In tale caso il numero di gradi di libertà n è pari a $n_1 + n_2 - 2$.

ESEMPIO

Supponiamo che uno stesso compito di matematica sia stato dato ai ragazzi appartenenti alle classi quinte di due istituti scolastici diversi. I dati ottenuti sono i seguenti:
$n_1 = 11$, $X(A_1) = 6.2$ ed $S_c(A_1) = 0.9$,
$n_2 = 9$, $X(A_2) = 6.6$ ed $S_c(A_2) = 1.4$.
Esiste una differenza significativa fra il profitto delle quinte delle due scuole, con un livello di significatività uguale a 0.01?

Detti m_1 e m_2 i voti medi di matematica delle classi quinte dei due istituti, formuliamo le seguenti ipotesi

H_0: $m_1 = m_2$, con livello di significatività uguale a 0.01, ossia non esiste differenza significativa fra i due gruppi
H_1: $m_1 \neq m_2$, con livello di significatività uguale a 0.01, ossia esiste differenza significativa tra i due gruppi.



Abbiamo

$$t = \frac{6.2 - 6.6}{\sqrt{\frac{0.9^2}{11} + \frac{1.4^2}{9}}} \approx -0.74$$

Poiché ci viene richiesto un livello di significatività uguale a 0.01 e dal momento che il numero di gradi di libertà è uguale a $n = 11 + 9 - 2 = 18$, allora t deve essere tale che $P(-2.90 \leq z \leq 2.90) = 0.99$. Pertanto l'ipotesi nulla è accettata se $-2.90 \leq t \leq 2.90$, ossia se $|t| \leq 2.90$.

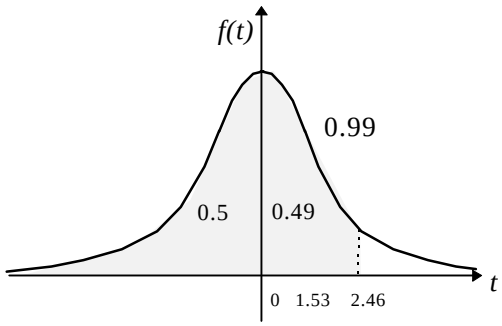
Dal momento che abbiamo trovato $z = -0.74$, accettiamo H_0, con livello di significatività uguale a 0.01.

ESEMPIO

Un contadino vuole indagare sull'efficacia di un fertilizzante sulla produzione di insalata. Sceglie quindi due appezzamenti di terreno, uno di $21m^2$ trattato con il fertilizzante, l'altro di $18m^2$ non trattato (gruppo di controllo). Dopo un certo periodo c'è stato un raccolto medio pari a $6kg$ e scarto quadratico medio corretto di $0.7kg$ e $5.7kg$ e scarto quadratico medio corretto di $0.9kg$ rispettivamente. Potete concludere che a livello di significatività dello 0.01 il fertilizzante fa aumentare la produzione di insalata?

I dati sono:
$n_1 = 21$, $X(A_1) = 6$ ed $S_c(A_1) = 0.7$,
$n_2 = 18$, $X(A_2) = 5.6$ ed $S_c(A_2) = 0.9$.

Detti m_1 e m_2 i raccolti medi di insalata di due terreni rispettivamente trattati e non trattati con il fertilizzante, formuliamo le seguenti ipotesi



H_0: $m_1 = m_2$, con livello di significatività uguale a 0.01, ossia la differenza non è significativa,
H_1: $m_1 > m_2$, con livello di significatività uguale a 0.01, ossia il fertilizzante è efficace.

Abbiamo

$$t = \frac{6 - 5.6}{\sqrt{\frac{0.7^2}{21} + \frac{0.9^2}{18}}} \approx 1.53$$

Poiché ci viene richiesto un livello di significatività uguale a 0.01 ed il numero di gradi di libertà è uguale a $n = 21 + 18 - 2 = 37$, allora t deve essere tale che $P(z \le 2.46) = 0.99$. Pertanto l'ipotesi nulla è accettata perché $t < 2.46$. Pertanto possiamo affermare che il fertilizzante non è efficace con un livello di significatività dello 0.01.

APPLICHIAMO

1) Una ditta costruttrice di cavi garantisce una possibilità di carico di 2000*kg*. Proviamo 14 cavi ottenendo un carico di rottura medio di 1990*kg* e uno scarto quadratico medio di 120*kg*. Verificare l'ipotesi, al livello di significatività uguale a 0.05 ed a 0.01, che la differenza riscontrata non sia significativa.
2) Una casa farmaceutica asserisce che un suo prodotto è in grado di guarire una data malattia nel 80% dei casi. In un campione di 10 pazienti che soffrono di questa malattia la medicina ne ha guariti 6. Determinate, con un livello di significatività uguale a 0.05 ed a 0.01, se la casa farmaceutica può asserire la precedente affermazione legittimamente.
3) Due campioni di 20 pezzi costruiti da una ditta e di 12 costruiti da una seconda ditta hanno mostrato rispettivamente 5 e 3 pezzi difettosi.
   Esiste una differenza significativa fra l'affidabilità delle due ditte, con un livello di significatività uguale a 0.01?
4) Con un campione di 20 pile prodotte da una ditta abbiamo ottenuto una durata media pari a 1300 ore con scarto quadratico medio di 120 ore. Con un campione di 16 pile prodotte da una seconda ditta abbiamo ottenuto una durata media pari a 1250 ore con scarto quadratico medio di 130 ore. Esiste una differenza fra le durate delle pile al livello di significatività dello 0.05? E dello 0.01?
5) Due campioni di 10 pile costruiti da una ditta e di 12 pile costruiti da una seconda ditta hanno mostrato rispettivamente 2 e 3 pile non funzionanti.
   Esiste una differenza significativa fra l'affidabilità delle due ditte di pile, con un livello di significatività uguale a 0.05?

1.15 Dimensione dei campioni

ESERCIZI

VERIFICA 1

Gruppo A

1) Considerate un'urna contenente 5 palline recanti i numeri 2, 5, 7, 10 e 12. Supponete di estrarre tre palline in blocco.
   a) Determinate lo spazio campionario S.
   b) Individuate l'evento E corrispondente alla proposizione *e*: "la somma *s* dei numeri scritti sulle palline è divisibile per 3.
   c) Determinate P(E).
   d) Considerate la variabile casuale X che associa ad ogni elemento di S:
      1 se $s \leq 20$; 2 se $21 \leq s \leq 22$, 3 se $s \geq 23$.
      Costruite la tabella di distribuzione ed il diagramma della funzione di distribuzione.
   e) Completate la precedente tabella inserendo la funzione di ripartizione, rappresentandola graficamente.
   f) Calcolate M(X), V(X) e s.

2) Il voto medio degli esami di maturità di un gruppo di studenti è di 50.34 con scarto quadratico medio di 5.
   a) Determinate la probabilità che il voto medio di un campione di 41 studenti presi a caso con estrazione bernoulliana sia compreso tra 48 e 52.
   b) Determinate la probabilità che il voto medio di un campione di 60 studenti presi a caso con estrazione bernoulliana non superi 47.
   c) Determinate la probabilità che il voto medio di un campione di 38 persone prese a caso con estrazione bernoulliana sia inferiore di 38 oppure superiore di 57.

Gruppo B

1) Considerate un'urna contenente 5 palline recanti i numeri 1, 2, 8, 20 e 22. Supponete di estrarre tre palline in blocco.
   a) Determinate lo spazio campionario S.
   b) Individuate l'evento E corrispondente alla proposizione *e*: "la somma *s* dei numeri scritti sulle palline è pari.
   c) Determinate P(E).

d) Considerate la variabile casuale X che associa ad ogni elemento di S:
   1 se $s \leq 25$; 2 se $26 \leq s \leq 40$, 3 se $s \geq 41$.
   Costruite la tabella di distribuzione ed il diagramma della funzione di distribuzione.
e) Completate la precedente tabella inserendo la funzione di ripartizione, rappresentandola graficamente.
f) Calcolate M(X), V(X) e s.

2) Il reddito medio di un gruppo di famiglie è di L.2.450.000 con scarto quadratico medio di L.360.000.
   a) Determinate la probabilità che il reddito medio di un campione di 41 famiglie prese a caso con estrazione bernoulliana sia compreso tra L.2.350.000 e L.2.500.000.
   b) Determinate la probabilità che il reddito medio di un campione di 60 famiglie prese a caso con estrazione bernoulliana non superi L.1.900.000.
   c) Determinate la probabilità che il reddito medio di un campione di 100 famiglie prese a caso con estrazione bernoulliana sia inferiore di L. 1.500.000 oppure superiore di L.3.000.000.

Gruppo C

1) Considerate un'urna contenente 5 palline recanti i numeri 0, 3, 10, 30 e 32. Supponete di estrarre tre palline in blocco.
   a) Determinate lo spazio campionario S.
   b) Individuate l'evento E corrispondente alla proposizione *e*: "la somma *s* dei numeri scritti sulle palline divisibile per 5.
   c) Determinate P(E).
   d) Considerate la variabile casuale X che associa ad ogni elemento di S:
      1 se $s \leq 34$; 2 se $35 \leq s \leq 49$, 3 se $s \geq 50$.
      Costruite la tabella di distribuzione ed il diagramma della funzione di distribuzione.
   e) Completate la precedente tabella inserendo la funzione di ripartizione, rappresentandola graficamente.
   f) Calcolate M(X), V(X) e s.

2) Il peso medio di un gruppo di adulti è di 78*kg* con scarto quadratico medio di 21*kg*.
   a) Determinate la probabilità che il peso medio di un campione di 41 persone prese a caso con estrazione bernoulliana sia compreso tra 70*kg* e 75*kg*.
   b) Determinate la probabilità che il peso medio di un campione di 60 persone prese a caso con estrazione bernoulliana non superi 79*kg*.
   c) Determinate la probabilità che il peso medio di un campione di 100 persone prese a caso con estrazione bernoulliana sia inferiore di 69*kg* oppure superiore di 80*kg*.

Gruppo D

1) Considerate un'urna contenente 5 palline recanti i numeri 0, 20, 32, 63 e 100. Supponete di estrarre tre palline in blocco.
   a) Determinate lo spazio campionario S.
   b) Individuate l'evento E corrispondente alla proposizione *e*: "la somma *s* dei numeri scritti sulle palline divisibile per 10.
   c) Determinate P(E).
   d) Considerate la variabile casuale X che associa ad ogni elemento di S:

1 se $s \leq 120$; 2 se $121 \leq s \leq 170$, 3 se $s \geq 171$.
Costruite la tabella di distribuzione ed il diagramma della funzione di distribuzione.

e) Completate la precedente tabella inserendo la funzione di ripartizione, rappresentandola graficamente.
f) Calcolate M(X), V(X) e s.

2) L'altezza media di un gruppo di bambini è di 93*cm* con scarto quadratico medio di 12*cm*.
   a) Determinate la probabilità che l'altezza media di un campione di 33 bambini presi a caso con estrazione bernoulliana sia compreso tra 70*cm* e 90*cm*.
   b) Determinate la probabilità che l'altezza media di un campione di 65 bambini presi a caso con estrazione bernoulliana non superi 85*cm*.
   c) Determinate la probabilità che l'altezza media di un campione di 112 bambini presi a caso con estrazione bernoulliana sia inferiore di 60*cm* oppure superiore di 100*cm*.

VERIFICA 2

Gruppo A

1) Risolvete il seguente esercizio utilizzando sia la distribuzione bernoulliana che approssimandola mediante la distribuzione normale.
   Sapendo che l'indice di occupazione di un paese è del 60%,.
   ◊ Determinate la probabilità che in un campione di 80 persone, prese con estrazione bernoulliana, 7 siano disoccupate.
   ◊ Determinate la probabilità che in un campione di 40 persone, prese con estrazione bernoulliana, i disoccupati siano in numero minore di 5.
   ◊ Determinate la probabilità che su 300 persone, prese con estrazione bernoulliana, gli occupati siano più di 200 ma meno di 280.

2) In un collegio elettorale si presentano alle elezioni due candidati. Da un sondaggio effettuato su un campione di 1500 persone è risultato che il primo candidato ha ottenuto una preferenza di 790 persone. Stimate per intervallo la percentuale dei voti che avrà il primo candidato con un grado di fiducia uguale a 0.95.

3) Una ditta costruttrice di cavi garantisce una possibilità di carico di 2500*kg*. Proviamo 110 cavi ottenendo un carico di rottura medio di 2460*kg* e uno scarto quadratico medio di 100*kg*. Verificare l'ipotesi, al livello di significatività uguale a 0.01, che la differenza riscontrata non sia significativa.

4) Due campioni di 200 pezzi costruiti da una ditta e di 150 pezzi costruiti da una seconda ditta hanno mostrato rispettivamente 10 e 8 pezzi difettosi.
   Esiste una differenza significativa fra l'affidabilità delle due ditte, con un livello di significatività uguale a 0.01?

5) È stato condotto un sondaggio d'opinione su un campione di 100 bambini per stabilire se preferissero o meno una data marca di cioccolata. Si sono osservati i dati di seguito riportati

	preferenza	non preferenza
bambini	33	21
bambine	36	10

Verificate se c'è differenza significativa tra le preferenze dei bambini e quella delle bambine con livello di significatività dello 0.01.

Gruppo B

1) Risolvete il seguente esercizio utilizzando sia la distribuzione bernoulliana che approssimandola mediante la distribuzione normale.
   Sapendo che l'indice di disoccupazione di un paese è del 12%,.
   - ◊ Determinate la probabilità che in un campione di 80 persone, prese con estrazione bernoulliana, 7 siano disoccupate.
   - ◊ Determinate la probabilità che in un campione di 40 persone, prese con estrazione bernoulliana, i disoccupati siano in numero minore di 5.
   - ◊ Determinate la probabilità che su 400 persone, prese con estrazione bernoulliana, gli occupati siano più di 300 ma meno di 350.

2) In un collegio elettorale si presentano alle elezioni due candidati. Da un sondaggio effettuato su un campione di 1100 persone è risultato che il primo candidato ha ottenuto una preferenza di 800 persone. Stimate per intervallo il numero di voti che avrà il primo candidato con un grado di fiducia uguale a 0.99.

3) Una ditta costruttrice di cavi garantisce una possibilità di carico di 2200*kg*. Proviamo 90 cavi ottenendo un carico di rottura medio di 2150*kg* e uno scarto quadratico medio di 69*kg*. Verificare l'ipotesi, al livello di significatività uguale a 0.05, che la differenza riscontrata non sia significativa.

4) Due campioni di 150 pezzi costruiti da una ditta e di 220 pezzi costruiti da una seconda ditta hanno mostrato rispettivamente 10 e 14 pezzi difettosi.
   Esiste una differenza significativa fra l'affidabilità delle due ditte, con un livello di significatività uguale a 0.05?

5) È stato condotto un sondaggio d'opinione su un campione di 120 studenti per stabilire se preferissero o meno un dato cantante. Si sono osservati i dati di seguito riportati

età	preferenza	non preferenza
15 - 17	43	21
18 - 20	46	10

   Verificate se c'è differenza significativa tra le preferenze dei ragazzi di età compresa tra i 15 e i 17 anni e quelli di età compresa tra i 18 ed i 20 anni con livello di significatività dello 0.01.

Gruppo C

1) Risolvete il seguente esercizio utilizzando sia la distribuzione bernoulliana che approssimandola mediante la distribuzione normale. Per 750 volte lanciamo due monete.
   - ◊ Determinate la probabilità che per 320 volte escano una Croce e una Testa.
   - ◊ Determinate la probabilità che il numero di volte che escano due Croci sia maggiore o uguale di 500.
   - ◊ Determinate la probabilità che il numero di volte che escano due Croci sia minore di 250.

2) Da una scatola si estraggono in modo bernoulliano 100 pezzi rilevandone 7 difettosi. Stimate per intervallo la percentuale di pezzi difettosi presenti nella scatola con un grado di fiducia uguale a 0.99.

3) Una casa farmaceutica asserisce che un suo prodotto è in grado di guarire una data malattia nel 90% dei casi. In un campione di 120 pazienti che soffrono di questa malattia la medicina ne ha guariti 108. Determinate, con un livello di significatività uguale a 0.01, se la casa farmaceutica può asserire la precedente affermazione legittimamente.

4) In due campioni di 110 lampadine costruiti da una ditta e di 105 lampadine costruiti da una seconda ditta sono state rilevate 6 lampadine non funzionanti.
Esiste una differenza significativa fra l'affidabilità delle due ditte di lampadine, con un livello di significatività uguale a 0.01?

5) È stato condotto un sondaggio d'opinione su un campione di 110 persone per stabilire quale quotidiano viene acquistato. Si sono osservati i dati di seguito riportati

	quotidiano A	quotidiano B
lettori	15	29
lettrici	23	33

Verificate se c'è differenza significativa tra le preferenze dei maschi e quelle delle femmine con livello di significatività dello 0.01.

Gruppo D

1) Risolvete il seguente esercizio utilizzando sia la distribuzione bernoulliana che approssimandola mediante la distribuzione normale. Per 700 volte lanciamo un dado.
   - ◊ Determinate la probabilità che per 200 volte esca un numero pari.
   - ◊ Determinate la probabilità che il numero di volte che esca un numero dispari sia maggiore o uguale di 400.
   - ◊ Determinate la probabilità che il numero di volte che esca un numero multiplo di tre sia minore di 200.

2) Da una scatola si estraggono in modo bernoulliano 45 pezzi rilevandone 2 difettosi. Stimate per intervallo la percentuale di pezzi difettosi presenti nella scatola con un grado di fiducia uguale a 0.95.

3) Una casa farmaceutica asserisce che un suo prodotto è in grado di guarire una data malattia nel 95% dei casi. In un campione di 97 pazienti che soffrono di questa malattia la medicina ne ha guariti 88. Determinate, con un livello di significatività uguale a 0.05, se la casa farmaceutica può asserire la precedente affermazione legittimamente.

4) In due campioni di 100 pile costruiti da una ditta e di 87 pile costruiti da una seconda ditta sono state rilevate rispettivamente 6 e 3 pile non funzionanti.
Esiste una differenza significativa fra l'affidabilità delle due ditte di lampadine, con un livello di significatività uguale a 0.05?

6) È stato condotto un sondaggio d'opinione su un campione di 130 famiglie per stabilire se preferissero o meno un dato programma televisivo. Si sono osservati i dati di seguito riportati

	preferenza	non preferenza
famiglie del nord	43	21
famiglie del sud	46	20

Verificate se c'è differenza significativa tra le preferenze delle famiglie del nord e quelle del sud con livello di significatività dello 0.01.