VARIABILI ALEATORIE

Psicometria 1 - Lezione 8
Lucidi presentati a lezione

AA 2000/2001 dott. Corrado Caudek

Una ***variabile aleatoria*** è una funzione avente come dominio lo spazio campione associato ad un esperimento e come codominio l'insieme dei numeri reali.

Una variabile aleatoria si dice ***discreta*** se può assumere solo un numero discreto di valori;

si dice ***continua*** se può assumere tutti gli infiniti valori di $\Re$, o di un suo intervallo $[a,b]$.

Esempio Un esperimento consiste nel lancio di due monete. Sia Y il numero di volte in cui l'esito “testa” viene osservato in ciascuna prova dell'esperimento.

Si identifichino i punti campione di S, si assegni un valore y a ciascun punto campione e si identifichino i punti campione associati a ciascuno dei valori che Y può assumere.

Ciascuna prova dell’esperimento può essere denotata da una coppia ordinata di simboli che identificano l’esito del lancio della prima e della seconda moneta.

Ad esempio, *TC* indicherà l’esito “testa” per il lancio della prima moneta e l’esito “croce” per il lancio della seconda moneta.

I quattro punti campione in S sono:

$$E_1 : \{TT\}, \; E_2 : \{TC\}, \; E_3 : \{CT\}, \; E_4 : \{CC\}.$$

Il valore da assegnare a Y in corrispondenza di ciascun punto campione dipende dal numero di volte in cui viene osservato l'esito "testa".

Nel caso dell'evento $E_1 : \{TT\}$, l'esito "testa" viene osservato due volte e quindi a Y viene assegnato il valore 2.

Y sarà uguale a 1 in corrispondenza degli eventi E_2 e E_3 dato che, in questi casi, l'esito “testa” viene osservato una sola volta.

Quando l'esperimento dà luogo all'evento E_4, infine, Y sarà uguale a 0.

In conclusione, la variabile Y può assumere tre valori (Y = 0, 1, 2) e ciascuno di essi può essere considerato un evento composto definito dal seguente insieme di punti campione:

$$\{Y=0\}=\{E_4\} \qquad \{Y=1\}=\{E_2, E_3\} \qquad \{Y=2\}=\{E_1\}.$$

Notazione

Solitamente le variabili aleatorie sono indicate con le lettere maiuscole, mentre gli specifici valori che assumono vengono indicati dalle lettere minuscole.

Nel caso dell'esperimento costituito dal lancio di un dado, ad esempio, Y indica uno qualunque dei 6 valori numerici che possono essere prodotti, mentre y denota lo specifico valore che viene osservato quando il dado è stato tratto.

Distribuzione di probabilità per una variabile aleatoria discreta

COME SI ASSOCIANO LE PROBABILITA' ALLE VARIABILI ALEATORIE?

La probabilità che la variabile Y assuma il valore y, $P(Y = y)$, è definita come la somma delle probabilità di tutti i punti campione in S a cui viene assegnato il valore y.

Generalmente, si denota $P(Y = y)$ con $p(y)$.

Si noti che *p(y)* non è altro che una funzione che assegna una probabilità a ciascun valore *y* ed è chiamata *funzione* (o distribuzione) *di probabilità* di *Y*.

La funzione di probabilità di una variabile aleatoria discreta *Y* può essere rappresentata con una formula, una tabella o un diagramma che associa la probabilità *p(y)* a ciascun valore *y*.

Una funzione di probabilità $p(y)$ deve rispettare le seguenti condizioni:

$$0 \leq p(y) \leq 1$$

$$\sum_{y} p(y) = 1.0$$

La prima condizione significa che a ciascun valore della variabile aleatoria discreta Y viene assegnata una probabilità non negativa;

la seconda condizione significa che la somma delle probabilità assegnate a tutti i valori y deve essere uguale a 1.

Esempio Un esperimento consiste nel lancio di due dadi. *Sia* Y *la somma dei punti osservati quando i due dadi sono stati tratti.*

Si trovi la distribuzione di probabilità di Y.

Lo spazio campione associato a questo esperimento è $S = \{(i, j) : 1 \le i, j \le 6\}$, ovvero, l'insieme di tutte le coppie ordinate (i, j) di numeri interi $1 \le i \le 6$ e $1 \le j \le 6$ che rappresentano l'esito del lancio di ciascuno dei due dadi.

In S ci sono 36 eventi semplici:

$$E_1 = \{1,1\},\ E_2 = \{1,2\},\ E_3 = \{1,3\},\ E_4 = \{1,4\},\ E_5 = \{1,5\},\ E_6 = \{1,6\},$$

$$E_7 = \{2,1\},\ E_8 = \{2,2\},\ E_9 = \{2,3\},\ E_{10} = \{2,4\},\ E_{11} = \{2,5\},\ E_{12} = \{2,6\},$$

$$E_{13} = \{3,1\},\ E_{14} = \{3,2\},\ E_{15} = \{3,3\},\ E_{16} = \{3,4\},\ E_{17} = \{3,5\},\ E_{18} = \{3,6\},$$

$$E_{19} = \{4,1\},\ E_{20} = \{4,2\},\ E_{21} = \{4,3\},\ E_{22} = \{4,4\},\ E_{23} = \{4,5\},\ E_{24} = \{4,6\},$$

$$E_{25} = \{5,1\},\ E_{26} = \{5,2\},\ E_{27} = \{5,3\},\ E_{28} = \{5,4\},\ E_{29} = \{5,5\},\ E_{30} = \{5,6\},$$

$$E_{31} = \{6,1\},\ E_{32} = \{6,2\},\ E_{33} = \{6,3\},\ E_{34} = \{6,4\},\ E_{35} = \{6,5\},\ E_{36} = \{6,6\}.$$

In corrispondenza di ciascun evento semplice dello spazio campione S la variabile Y assume i seguenti valori:

$$\{Y=2\}=\{E_1\},$$

$$\{Y=3\}=\{E_2, E_7\},$$

$$\{Y=4\}=\{E_3, E_8, E_{13}\},$$

$$\{Y=5\}=\{E_4, E_9, E_{14}, E_{19}\},$$

$$\{Y=6\}=\{E_5, E_{10}, E_{15}, E_{20}, E_{25}\},$$

$$\{Y=7\}=\{E_6, E_{11}, E_{16}, E_{21}, E_{26}, E_{31}\},$$

$$\{Y=8\}=\{E_{12}, E_{17}, E_{22}, E_{27}, E_{32}\},$$

$$\{Y=9\}=\{E_{18}, E_{13}, E_{28}, E_{33}\},$$

$$\{Y=10\}=\{E_{24}, E_{29}, E_{34}\},$$

$$\{Y=11\}=\{E_{30}, E_{35}\},$$

$$\{Y=12\}=\{E_{36}\}.$$

Se i dadi non sono truccati, tutti gli eventi semplici dello spazio campione S sono equiprobabili.

A ciascun evento semplice $E_i \in \mathrm{S}$ può quindi essere assegnata la stessa probabilità, ovvero 1/ 36.

Una volta trovate le probabilità degli eventi semplici in S, la probabilità da assegnare a ciascun valore y si calcola facendo la somma delle probabilità di tutti i punti campione in S a cui è associato il valore y.

$$P(Y=2)=P(E_1)=1/36$$

$$P(Y=3)=P(E_1)+P(E_7)=2/36$$

$$P(Y=4)=P(E_1)+P(E_7)+P(E_{13})=3/36$$

In questo modo la distribuzione di probabilità di *Y* diventa:

Y	2	3	4	5	6	7	8	9	10	11	12
$p(y)$	1/ 36	2/ 36	3/ 36	4/ 36	5/ 36	6/ 36	5/ 36	4/ 36	3/ 36	2/ 36	1/ 36

Rappresentazione grafica di una distribuzione di probabilità

Supponiamo di ripetere molte volte un esperimento i cui risultati possibili sono dei numeri y compresi nell'intervallo $a \le y \le b$.

Dividiamo l'intervallo $[a, b]$ in n sottointervalli di lunghezza uguale, diciamo $a = y_1 < y_2 < y_3 < \ldots < y_n = b$, e contiamo il numero di volte m_k in cui il risultato dell'esperimento è un numero compreso tra y_{k-1} e y_k.

Se questo modo di suddividere i dati è rappresentato con una serie di rettangoli contigui aventi altezza m_k sopra l'intervallo k-esimo, allora il diagramma risultante è chiamato *istogramma*.

La distribuzione di probabilità di una variabile aleatoria discreta Y può essere rappresentata in maniera grafica costruendo un istogramma che associa a ciascun valore $Y = y$ un rettangolo avente ampiezza $1/n$ (dove n è il numero di valori che la variabile aleatoria può assumere) e altezza $p(y)$.

In questo modo, l'area di ciascun rettangolo sarà uguale alla probabilità $P(Y = y)$ e l'area totale di tutti i rettangoli che formano l'istogramma sarà uguale a 1.

Ne segue che la probabilità $P(c \le Y \le d)$, con $a \le c < d \le b$, sarà uguale alla somma delle aree dei rettangoli dell'istogramma nell'intervallo $[c, d]$.

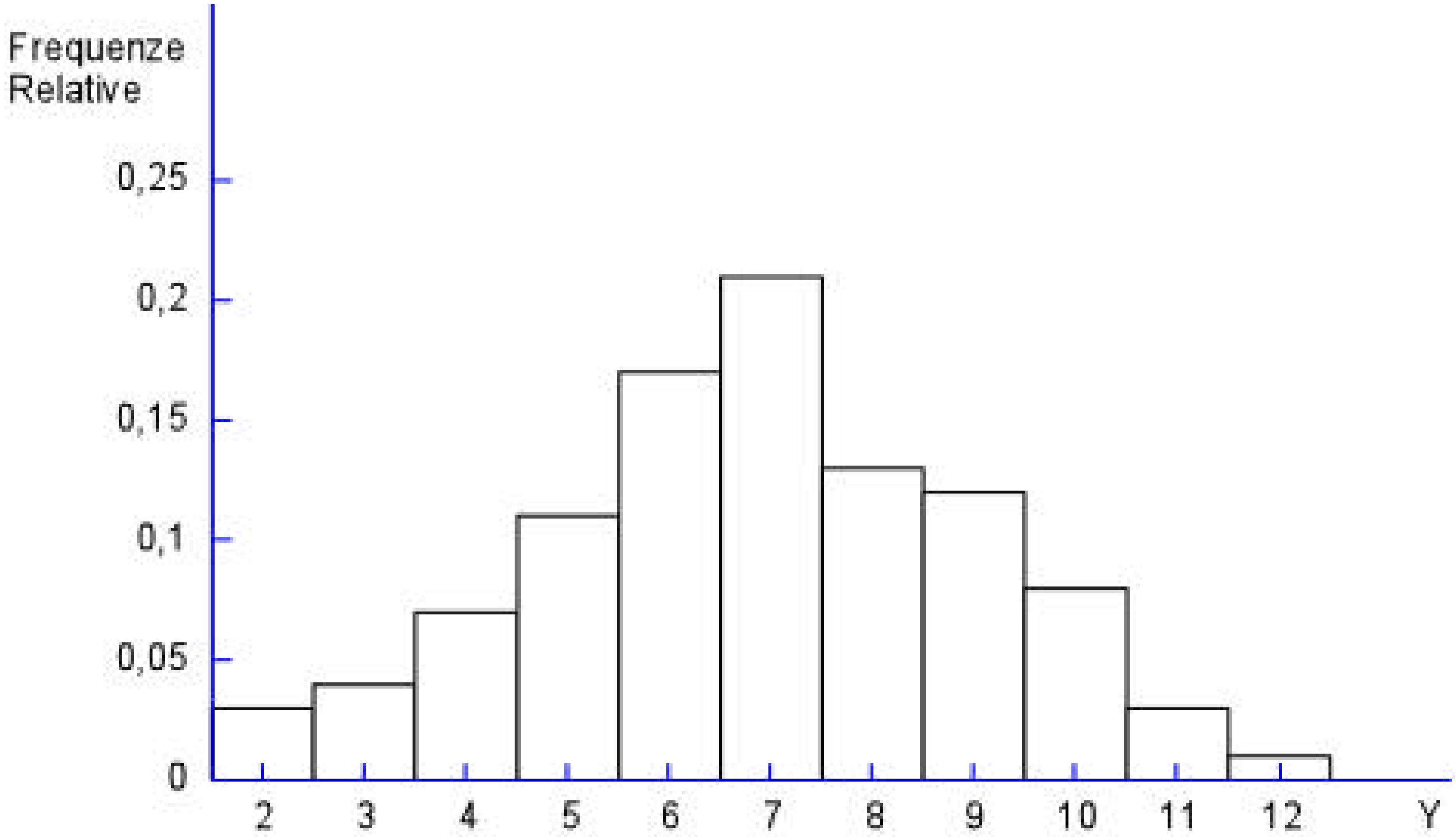
Esempio Se l'esperimento descritto nell'esempio precedente fosse effettivamente eseguito, un singolo valore Y verrebbe osservato.

Il valore di Y potrebbe essere 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 o 12.

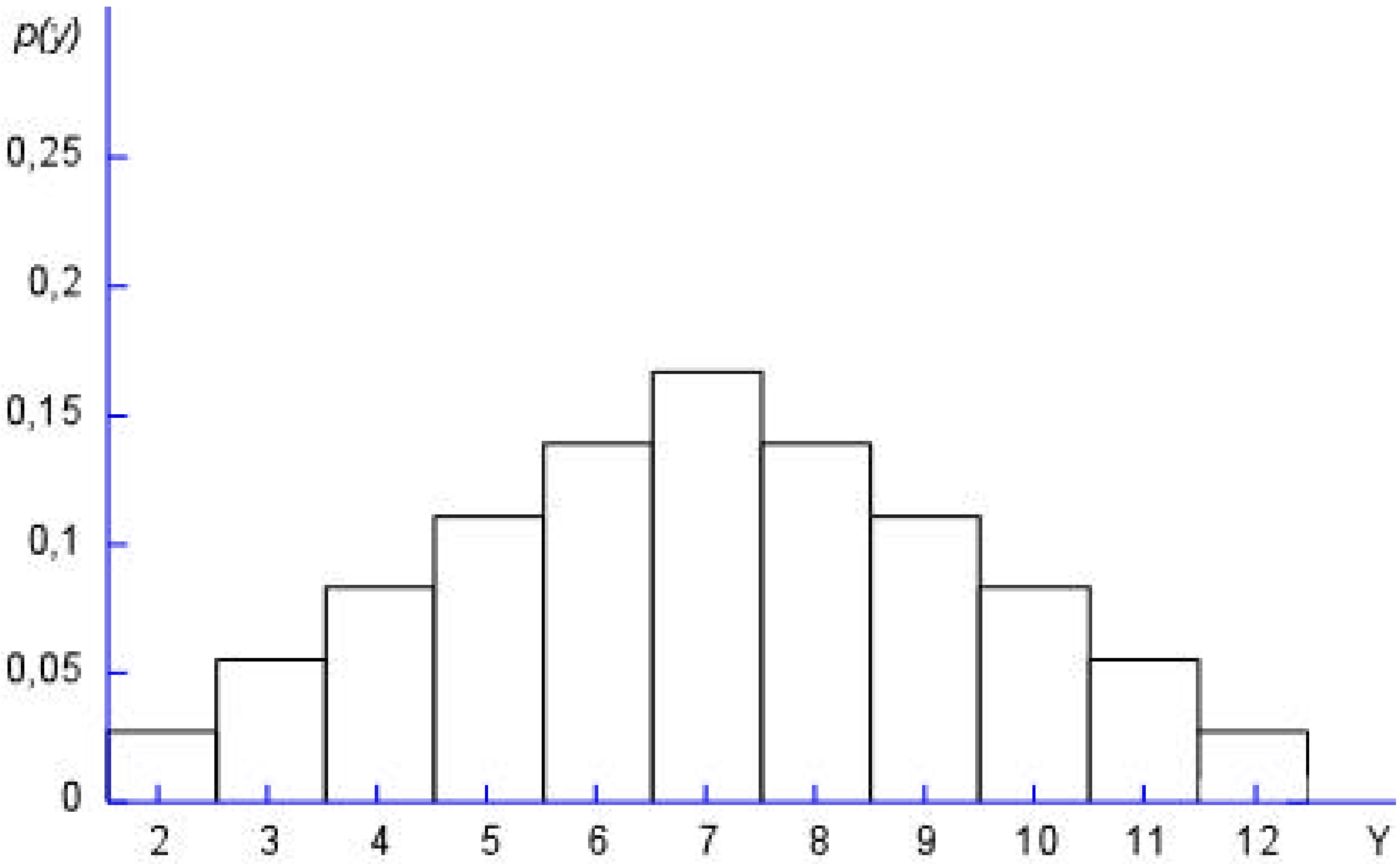
Se l'esperimento venisse ripetuto n volte, n valori y verrebbero osservati.

Supponiamo che, dopo avere eseguito l'esperimento 100 volte, i seguenti risultati vengano osservati.

Y	Frequenze osservate in 100 lanci	Frequenze relative
$y_1 = 2$	3	.03
$y_2 = 3$	4	.04
$y_3 = 4$	7	.07
$y_4 = 5$	11	.11
$y_5 = 6$	17	.17
$y_6 = 7$	21	.21
$y_7 = 8$	13	.13
$y_8 = 9$	12	.12
$y_9 = 10$	8	.08
$y_{10} = 11$	3	.03
$y_{11} = 12$	1	.01


Frequenze
Relative
0,25
0,2
0,15
0,1
0,05
0
2
3
4
5
6
7
8
9
10
11
12
Y

La distribuzione di probabilità che abbiamo definito nell’esempio precedente è:


p(y)
0,25
0,2
0,15
0,1
0,05
0
2
3
4
5
6
7
8
9
10
11
12
Y

Si noti la somiglianza tra i due istogrammi.

In seguito verrà presentato un teorema in cui si dimostra che, all'aumentare del numero delle ripetizioni di un esperimento, la distribuzione delle frequenze empiriche si approssima sempre più alla distribuzione teorica di probabilità.

… a che cosa serve una distribuzione di probabilità?

La distribuzione di probabilità di una variabile aleatoria fornisce *un modello teorico della distribuzione di frequenza di una popolazione di dati empirici*.

Se il modello fornisce una rappresentazione adeguata dei fenomeni considerati, la distribuzione teorica e quella empirica saranno equivalenti.

Poniamoci ora il problema di determinare la media e la varianza di una variabile aleatoria discreta.

VALORE ATTESO DI UNA VARIABILE ALEATORIA DISCRETA

Sia Y una variabile aleatoria discreta con una funzione di probabilità $p(y)$.

Il *valore atteso* (o *speranza matematica*) di Y è definito come:

$$E(Y) = \sum_{y} y p(y)$$

Se $p(y)$ fornisce un'accurata rappresentazione della distribuzione di frequenza di una popolazione di dati empirici, allora il valore atteso diventa uguale alla media m della popolazione, $E(Y)=m$.

Il valore atteso ha quindi lo stesso significato della media aritmetica, l’unica differenza essendo il fatto che il valore atteso si riferisce ad una distribuzione teorica di probabilità mentre la media aritmetica si riferisce a una distribuzione empirica di dati.

Esempio Un esperimento consiste nel lancio di un dado non truccato. Sia X il numero di punti osservati sulla faccia superiore dopo che il dado è stato tratto.

Si calcoli il valore atteso di X.

Dato che il dado non è truccato, la distribuzione di probabilità di X è uniforme: $P(X = x_i) = \frac{1}{6}$, con i = 1, …, 6.

Il valore atteso di X è:

$$E(X) = \sum xp(x) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

Supponiamo di ripetere l'esperimento 6 milioni di volte e di registrare i valori che X assume in ciascuno di questi lanci. Quale sarà la media dei valori X così osservati?

Dato che la distribuzione di probabilità di X è uniforme, possiamo aspettarci di osservare approssimativamente 1 milione di volte ciascuno dei valori che X può assumere. Facendo la media, otteniamo

$$m \approx \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(1.000.000)(1) + (1.000.000)(2) + \ldots + (1.000.000)(6)}{6.000.000} \approx 3{,}5$$

… il che conferma che il valore atteso è equivalente alla media aritmetica.

PROPRIETA’ DEL VALORE ATTESO

1 Il valore atteso di una costante c è uguale a c:

$$E(c) = c$$

2 Il valore atteso di una variabile aleatoria discreta Y moltiplicata per una costante c è uguale al valore della costante moltiplicato per il valore atteso della variabile aleatoria:

$$E(cY) = cE(Y)$$

3 Il valore atteso della somma di due variabile aleatorie discrete X e Y è uguale alla somma dei rispettivi valori attesi:

$$E(X+Y)=E(X)+E(Y)$$

In maniera equivalente, il valore atteso della differenza di due variabile aleatorie discrete è uguale alla differenza dei rispettivi valori attesi:

$$E(X - Y) = E(X) - E(Y)$$

Esempio. Consideriamo nuovamente l'esperimento precedente.

Sia X l'esito prodotto dal lancio del primo dado e W l'esito prodotto dal lancio del secondo dado.

Sia $Y = X + W$.

Si trovi il valore atteso di Y.

$$E(Y) = \sum yp(y) = 2\left(\frac{1}{36}\right) + 3\left(\frac{2}{36}\right) + 4\left(\frac{3}{36}\right) + \ldots + 10\left(\frac{3}{36}\right) + 11\left(\frac{2}{36}\right) + 12\left(\frac{1}{36}\right) = 7$$

Notate che lo stesso risultato si poteva trovare usando una delle proprietà del valore atteso descritte in precedenza.

$$E(X + W) = E(X) + E(W) = 3.5 + 3.5 = 7$$

4 Se due variabili aleatorie discrete X e Y sono *indipendenti*, allora il valore atteso del loro prodotto è uguale al prodotto dei rispettivi valori attesi:

$$E(X \times Y) = E(X)E(Y)$$

Esempio Una moneta è lanciata due volte. La variabile aleatoria W_1 assume valore 1 se l'esito del primo lancio è "testa" e zero altrimenti. La variabile W_2 assume valore 1 se l'esito del secondo lancio è "testa" e zero altrimenti. Si trovi il valore atteso del prodotto $W_1 \cdot W_2$.

Le variabili W_1 e W_2 sono indipendenti e hanno entrambe valore atteso uguale a ½.

Di conseguenza,

$$E(W_1 \cdot W_2) = E(W_1) \cdot E(W_2) = 1/2 \cdot 1/2 = 1/4 \; .$$

Esempio Si consideri l'esperimento consistente nel singolo lancio di una moneta.

Sia $X = 1$ se viene osservato l'esito "testa" e $X = 0$ se viene osservato l'esito "croce".

Sia $Y = 1 - X$.

Per entrambe queste variabili, il valore atteso sarà $E(X) = E(Y) = ½$. Il prodotto $X \cdot Y$, però, è uguale a zero per entrambi gli esiti che il lancio della moneta può produrre.

Quindi, $E(X \cdot Y) \neq E(X)E(Y)$ se le variabili aleatorie non sono indipendenti.

VARIANZA DI UNA VARIABILE ALEATORIA DISCRETA

La varianza di una distribuzione di misure è stata definita come la media degli scostamenti al quadrato di ciascuna misura dalla media.

In maniera analoga, la varianza di una variabile aleatoria Y è definita come il valore atteso di $(Y-m)^2$:

$$V(Y) = E\left[(Y - m)^2\right]$$

ovvero

$$V(Y) = \sum_{y} (y - m)^2 \, p(y)$$

La radice quadrata della varianza viene detta *errore standard***.**

Esempio Sia X il numero di punti prodotti dal lancio di un dado. Si trovi la varianza di X.

Il valore atteso di *X*, *E(X)* = 3.5, è stato calcolato in precedenza.

La varianza di *X* diventa quindi

$$V(X) = \sum_x (x-m)^2 \, p(x)$$

$$= (1-3{,}5)^2 \frac{1}{6} + (2-3{,}5)^2 \frac{1}{6} + (3-3{,}5)^2 \frac{1}{6} + (4-3{,}5)^2 \frac{1}{6} + (5-3{,}5)^2 \frac{1}{6} + (6-3{,}5)^2 \frac{1}{6} = \frac{35}{12}$$

La formula $V(Y)=\sum_{y}(y-m)^2 p(y)$ può essere riscritta come:

$$V(Y)=E(Y^2)-m^2$$

$$V(Y) = \sum (y - m)^2 p(y) = \sum \left(y^2 - 2my \ + m^2\right) p(y)$$

$$= \sum y^2 p(y) - 2m \sum yp(y) + m^2 \sum p(y)$$

dato che $\sum yp(y) = E(Y) = m$ e che $\sum p(y) = 1.0$, allora

$$= \sum y^2 p(y) - 2m^2 + m^2 = \sum y^2 p(y) - m^2 = E\left(Y^2\right) - m^2$$

PROPRIETA’ DELLA VARIANZA

1 La varianza di una variabile aleatoria discreta Y moltiplicata per una costante c è uguale alla varianza della variabile aleatoria moltiplicata per la costante innalzata al quadrato:

$$V(cY) = c^2 V(Y)$$

2 La varianza di una variabile aleatoria discreta *Y* non muta se a ciasun valore *y* viene sommata una costante *c*:

$$V(Y+c)=V(Y)$$

3 Se due variabili aleatorie discrete X e Y sono indipendenti, allora la varianza della loro somma è uguale alla somma delle rispettive varianze:

$$V(X+Y)=V(X)+V(Y)$$

In maniera analoga, la varianza della differenza di due variabili aleatorie indipendenti è uguale alla somma delle rispettive varianze:

$$V\left(X-Y\right)=V\left(X\right)+V\left(Y\right)$$

VARIABILI ALEATORIE CONTINUE

In precedenza abbiamo osservato che la distribuzione di probabilità di una variabile aleatoria discreta Y può essere generata assegnando una probabilità $p(y)$ maggiore o uguale a zero a ciascuno dei valori che Y può assumere, in modo tale che $\sum_y p(y) = 1$.

La distribuzione di probabilità di una variabile aleatoria continua, però, non può essere specificata nello stesso modo.

Non è possibile, infatti, assegnare un valore non nullo a ciascuno degli infiniti valori che una variabile aleatoria continua può assumere e, allo stesso tempo, rispettare il vincolo secondo cui la somma di tali probabilità deve essere uguale ad 1.

Per assegnare i valori di probabilità ad una variabile aleatoria continua dobbiamo dunque procedere in un altro modo.

FUNZIONE DI DENSITA'
DI PROBABILITA'

$$f(y) \geq 0$$

$$\int_{-\infty}^{\infty} f(y)\,dy = 1$$

Una volta definita la nozione di funzione di densità, è possibile descrivere una procedura che consente di associare valori di probabilità ad una variabile aleatoria continua.

In precedenza abbiamo osservato che, nel caso di una variabile aleatoria discreta Y, la probabilità $P(c \leq Y \leq d)$ è uguale all'area delle barre dell'istogramma che rappresenta la distribuzione di probabilità di Y nell'intervallo $[c, d]$

Anche nel caso di una variabile aleatoria continua X, dunque, potremmo aspettarci che $P(c \leq X \leq d)$ sia uguale all'area sottesa dalla funzione di densità $f(x)$ nell'intervallo [c, d].

Infatti, è proprio così.

Probabilità diverse da zero possono quindi essere assegnate ad *intervalli di valori* di una variabile aleatoria continua X.

A ciascuno dei singoli valori che la variabile aleatoria continua può assumere, invece, è sempre associata una probabilità uguale a zero, $P(X = x) = 0$.

Quest’ultima affermazione risulta più facilmente comprensibile se interpretiamo la probabilità in termini geometrici:

l’area sottesa alla funzione di densità *f(x)* nell’intervallo corrispondente ad un punto è necessariamente uguale a zero.

Cosa significa in pratica il fatto che una probabilità uguale a 0 viene associata ad un evento perfettamente legittimo e possibile (ovvero, l'evento corrispondente al fatto che la variabile aleatoria continua assuma un determinato valore)?

Esempio. Sia X la variabile aleatoria continua corrispondente all'altezza in centimetri di uno studente scelto a caso dalla popolazione studentesca universitaria.

Supponiamo di disporre di uno strumento *infinitamente preciso* per misurare l'altezza e poniamo che l'altezza degli studenti universitari vari tra 150 e 210 cm.

La variabile aleatoria X, quindi, potrà assumere qualsiasi valore compreso in questo intervallo.

Consideriamo ora un particolare evento all'interno dello spazio campione, ovvero l'evento $X = 173.128735274$ cm.

Pur essendo questo un valore che X può legittimamente assumere, non sarà facile trovare uno studente che abbia *esattamente* questa altezza.

Potremmo continuare a misurare l'altezza di moltissimi studenti senza mai osservare l'evento in questione. Questo ci porterebbe a concludere tale evento è talmente improbabile che ad esso può essere assegnata una probabilità uguale a 0.

In base al medesimo argomento, di conseguenza, *una probabilità uguale a 0 dovrà essere assegnata a qualunque altro specifico valore che X può assumere.*

Consideriamo ora l'evento: $172 \leq X \leq 174$.

Non sarà difficile, in questo caso, trovare degli studenti universitari la cui altezza è compresa in questo intervallo. Di conseguenza, a questo secondo evento (corrispondente ad un *intervallo di valori*) potrà essere assegnata una probabilità maggiore di 0.

ESERCIZI

E1. Sia *X* una variabile aleatoria che può assumere soltanto i valori 1, 2, 3, 4, 5. La funzione di probabilità *p(x)* è

$$p(x) = \begin{cases} x^2/55 & \text{se } x = 1, 2, 3, 4, 5 \\ 0 & \text{altrimenti} \end{cases}$$

1. Rappresentata in forma grafica la funzione di probabilità di X.

2. Rappresentate graficamente la funzione di probabilità cumulativa di X.

3. Quale è la probabilità che X assuma uno dei valori 2, 3 oppure 4?

4. Quale è la probabilità che X assuma un valore qualunque?

5. Calcolate il valore atteso e la varianza di X.

E2. All’interno di un processo di produzione la temperatura (in Fahereneit) non varia mai di più di 2° dal valore di 62°.

La temperatura, infatti, può essere considerata una variabile aleatoria *F* con la seguente distribuzione

F	**60**	**61**	**61**	**63**	**64**
p(f)	**1/10**	**2/10**	**4/10**	**2/10**	**1/10**

1. Si trovi *E(F)* e *V(F)*.

2. Sia $T = F - 62$. Si trovi *E(T)* e *V(T)*.

3. Si trasformino le temperature in gradi centigradi:
$C = (5/9)(F - 32)$. Si trovi *E(C)* e *V(C)*.

E3. Una moneta onesta viene lanciata 3 volte. Sia Y il numero di teste che viene osservato. Si trovi $V(Y)$.

E4. Un dado viene costruito in modo tale che la probabilità di ottenere il numero x (con *1 £ x £ 6*) è proporzionale a x. Il dado viene lanciato e produce il risultato X. Si trovi la varianza di X.