PROBABILITÀ – SCHEDA N. 5
SOMMA E DIFFERENZA DI DUE VARIABILI ALEATORIE DISCRETE

1. Distribuzione congiunta

Ci sono situazioni in cui un esperimento casuale non si può modellare con una sola variabile casuale, perché quello che interessa sono proprio le relazioni presenti tra due o più grandezze. Ad esempio nello studio di possibili cause di tumore potremmo voler indagare il rapporto tra il numero medio di sigarette fumate in un giorno e l'età in cui viene riscontrata questa patologia.
Per specificare la relazione tra due variabili aleatorie discrete il punto di partenza è quello di estendere il concetto di densità di probabilità.

La **densità congiunta** di X e Y, che indichiamo con $f_{XY}(x,y)$, è la probabilità che X sia uguale ad x e Y sia uguale a y:

$$f_{XY}(x,y) = P(X = x, Y = y)$$

che può essere schematizzata con una tabella di contingenza "a due entrate".

Come abbiamo visto nel caso della probabilità di due eventi, alla tabella della densità congiunta (Tabella 1 della scheda 1), possono essere aggiunte una riga e una colonna che rappresentano le **densità di probabilità** (**marginali**) di X e Y.

Due variabili aleatorie X e Y sono **indipendenti** se

$$P(X=x \cap Y=y)=P(X=x)P(Y=y),$$

per ogni coppia di valori x e y assunti da X e Y.

2. Densità di probabilità della somma di due variabili aleatorie discrete

ESEMPIO 1. Uno studente ha registrato i tempi che impiega per andare a scuola facendo prima un tragitto a piedi e poi uno in autobus. Indichiamo con X il tempo (in minuti) del percorso a piedi e di attesa dell'autobus e con Y il tempo (in minuti) del percorso in autobus. Utilizziamo le frequenze registrate come valutazione delle probabilità dei tempi di percorrenza. Per semplicità suddividiamo i tempi in classi e consideriamo il valore centrale della classe.

X	5	7.5	10
$f_X(x)$	0.20	0.50	0.30

Y	5	10	15	20
$f_X(x)$	0.10	0.30	0.40	0.20

Tabella 1. Densità di probabilità (marginali) di X e Y, tempi di percorrenza

Siamo interessati a studiare il tempo totale per andare a scuola, ovvero la densità di probabilità di $X+Y$.

Se consideriamo X e Y indipendenti, possiamo calcolare la tabella della distribuzione congiunta.

X\Y	5	10	15	20	totale X
5	0.02	0.06	0.08	0.04	0.20
7.5	0.05	0.15	0.20	0.10	0.50
10	0.03	0.09	0.12	0.06	0.30
totale Y	0.10	0.30	0.40	0.20	

Tabella 2. Densità congiunta di X e Y, tempi di percorrenza

Quali valori può assumere la variabile aleatoria $X+Y$? I valori di tutte le somme possibili, cioè

$X+Y$	5	10	15	20
5	10	15	20	25
7.5	12.5	17.5	22.5	27.5
10	15	20	25	30

Tabella 3. Valori assunti da $X+Y$. Sono evidenziati i valori uguali.

Utilizzando le tabelle 2 e 3 possiamo calcolare la densità di $X+Y$.

$X+Y$	10	12.5	15	17.5	20	22.5	25	27.5	30
$f_{X+Y}(k)$	0.02	0.05	0.06 + 0.03	0.15	0.08 + 0.09	0.20	0.04 + 0.12	0.10	0.06

Tabella 4. Densità di probabilità di $X+Y$.

ESEMPIO 2. In una popolazione il 15% delle coppie non ha figli, il 20% ne ha uno, il 35% ne ha due e il 30% ne ha tre. Quindi ipotizziamo che nessuna coppia abbia più di tre figli (infatti 15%+20%+35%+30%=100%). Assumiamo che in ogni famiglia il genere dei figli (maschio, femmina) sia equiprobabile e indipendente da quello dei fratelli. Si seleziona una famiglia a caso e si denotano con F e M il numero di femmine e di maschi presenti tra i figli in tale famiglia.
Calcoliamo la densità congiunta.

- probabilità di nessun figlio: $P(F=0, M=0)=0.15$
- probabilità di un solo figlio femmina:

$$P(F=1, M=0)=P(F+M=1, F=1)=P(F=1 \mid F+M=1)P(F+M=1)=0.50\times 0.20=0.10$$

- probabilità di un due figli femmina:

$$P(F=2, M=0)=P(F+M=2, F=2)=P(F=2 \mid F+M=2)P(F+M=2)=0.50^2\times 0.35=0.0875$$

e così via. Verificate che anche gli altri valori della tabella siano corretti.

F\M	0	1	2	3	totale F
0	0.1500	0.1000	0.0875	0.0375	0.3750
1	0.1000	0.1750	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2000
3	0.0375	0	0	0	0.0375
totale M	0.375	0.3875	0.2000	0.0375	

Tabella 5. Densità congiunta del numero di figli maschi e femmine per famiglia

Osserviamo che le due variabili F e M non sono indipendenti. Infatti ad esempio $P(F=3,\ M=0)\neq P(F=3)\ P(M=0)$ ossia $0{,}0375\neq(0{,}5)^3$

ESEMPIO 3. Abbiamo già visto, in schede precedenti, il problema del lancio di due dadi. Indichiamo con X e Y le variabili aleatorie che rappresentano il numero di pallini sulla faccia in alto rispettivamente del primo e del secondo dado. Consideriamo la variabile aleatoria $X+Y$ che rappresenta la somma dei pallini sulle facce di ciascuno dei due dadi.
Quali valori può assumere e qual è la sua densità di probabilità?
$X+Y$ assume tutti i valori interi da 2 (caso 1+1) fino a 12 (caso 6+6); con quali probabilità?
Consideriamo, per esempio, il caso di $X+Y$ = 4. La somma dei pallini sulle due facce è 4 nei casi 1+3, 2+2 e 3+1. Attenzione: bisogna considerare sia 3+1 (3 sul primo dado e 1 sul secondo) che 1+3 (1 sul primo dado e 3 sul secondo).
Allora

$$P(X+Y=4) = P(X=1,Y=3) + P(X=2,Y=2) + P(X=3,Y=1)$$

I lanci e i risultati sono indipendenti; quindi:

$$P(X+Y=4) = P(X=1)P(Y=3) + P(X=2)P(Y=2) + P(X=3)P(Y=1) = 3/36 = 1/12.$$

X\Y	1	2	3	4	5	6	
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

Tabella 6. Distribuzione congiunta del risultato del lancio di due dadi con evidenziate le probabilità di ottenere come somma 2, 5 e 8.

Calcolate la densità della variabile aleatoria $X+Y$.

$X+Y$	2	3	4	5	6	7	8	9	10	11	12
$f_{X+Y}(k)$											

Quale punteggio ha probabilità massima?

ESEMPIO 4. Consideriamo l'acquisto di pane di una famiglia di una certa zona della città in due giorni consecutivi dal lunedì al venerdì. Indichiamo con X_1 l'acquisto di pane in chilogrammi nel primo giorno e con X_2 l'acquisto, sempre in chilogrammi, nel giorno successivo. Si sono rilevati i dati su tutte le famiglie per un mese e si utilizzano questi dati come probabilità che una famiglia, scelta a caso nella zona, compri determinate quantità di pane.

Si è trovato che le distribuzioni marginali delle due variabili sono uguali (cioè le famiglie comprano il pane in media con la stessa distribuzione)

X	0	$\leq$ 0. 5	(0.5, 1.0]	(1.0, 1.5]	> 1.5
f	0.20	0.40	0.30	0.05	0.05

Per semplicità attribuiamo a X_1 e a X_2 i valori medi delle classi di chilogrammi di pane comprati:

X	0	0.250	0.750	1.250	2.000
f	0.20	0.40	0.30	0.05	0.05

È stata rilevata anche la seguente distribuzione condizionata di X_2 rispetto a X_1 (profili riga):

	0	0.250	0.750	1.250	2.000
0	0.050	0.475	0.400	0.050	0.025
0.250	0.250	0.450	0.250	0.025	0.025
0.750	0.200	0.350	0.367	0.050	0.033
1.250	0.200	0.300	0.100	0.200	0.200
2.000	0.400	0.100	0.100	0.100	0.300
	0.20	0.40	0.30	0.05	0.05

Tabella 7. Distribuzione condizionata dell'acquisto di pane rispetto a quanto acquistato il giorno precedente.

Calcoliamo la distribuzione congiunta di X_1 (in riga) e X_2 (in colonna).

Come sappiamo la probabilità di comprare la coppia di quantità di pane nei due giorni consecutivi (x_1, x_2) è:

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)\,P(X_2 = x_2 \mid X_1 = x_1)$$

quindi, ad esempio, la probabilità che il primo giorno non sia stato comprato pane **e** il secondo giorno si compri:

0 kg di pane è: 0.20 x 0.050 = 0.010
0.250 kg di pane è: 0.20 x 0.475 = 0.095
0.750 kg di pane è: 0.20 x 0.400 = 0.080
1.250 kg di pane è: 0.20 x 0.050 = 0.010
2.000 kg di pane è: 0.20 x 0.025 = 0.005

Controlla la tabella della distribuzione congiunta riportata qui sotto:

	0	0.250	0.750	1.250	2.000
0	0.010	0.095	0.080	0.010	0.005
0.250	0.100	0.180	0.100	0.010	0.010
0.750	0.060	0.105	0.110	0.015	0.010
1.250	0.010	0.015	0.005	0.010	0.010
2.000	0.020	0.005	0.005	0.005	0.015

Tabella 8. Distribuzione congiunta dell'acquisto di pane in due giorni consecutivi.

Consideriamo ora la variabile aleatoria che indica la **somma** del pane comprato nei due giorni:

$$T = X_1 + X_2.$$

La variabile aleatoria T può assumere valori da 0 a 4 kg così come riportato nella seguente tabella della somma dei valori di T

	0	0.250	0.750	1.250	2.000
0	0.000	0.250	0.750	1.250	2.000
0.250	0.250	0.500	1.000	1.500	2.250
0.750	0.750	1.000	1.500	2.000	2.750
1.250	1.250	1.500	2.000	2.500	3.250
2.000	2.000	2.250	2.750	3.250	4.000

Tabella 9. Valori assunti dalla somma del pane comprato in due giorni consecutivi.

quindi riassumendo (completa):

	casi
0	$X_1=0$ $X_2=0$
0.250	($X_1=0$ $X_2=0.250$) ($X_1=0.250$ $X_2=0$)
0.500	($X_1=0.250$ $X_2=0.250$)
0.750	($X_1=0.250$ $X_2=0.500$) ($X_1=0.500$ $X_2=0.250$)
1.000	
0.250	
0.750	

	casi
1.250	
1.500	
2.000	
2.250	
2.750	
3.250	
4.000	

Possiamo ora calcolare la funzione di densità di T (completa):

T	f_T
0	0.010
0.250	0.095 + 0.10 = 0.195
0.500	0.180
0.750	
1.000	
0.250	
0.750	

T	f_T
1.250	
1.500	
2.000	
2.250	
2.750	
3.250	
4.000	

Tabella 10. Densità di probabilità di T*.*

Controlla se hai fatto i calcoli esatti calcolando la somma dei valori della densità di T, che deve risultare uguale a 1.

Dagli esempi precedenti deduciamo che NON possiamo ricavare la densità di probabilità di $X+Y$, note le densità di X e Y. È necessario conoscere la densità congiunta. Infatti la probabilità che $X+Y$ sia uguale a k coincide con la somma di tutte le probabilità che X sia uguale ad a e Y sia uguale a b, con la condizione che $a+b$ sia uguale a k.

3. Densità di probabilità della differenza di due variabili aleatorie discrete

Quanto abbiamo visto per la somma di due variabili aleatorie, vale in modo analogo anche per la differenza o il prodotto.

ESEMPIO 5. Consideriamo nuovamente il lancio di due dadi e la variabile aleatoria T che indica il valore assoluto della differenza dei pallini sulle due facce:

$$T = |X - Y|$$

La variabile aleatoria T assume valori tra 0 e 5.
Quale è la densità di T calcolata in 0? Ovvero, qual è la probabilità che T assuma il valore 0? È la probabilità di ottenere le coppie con valori uguali e quindi 6 x 1/36, cioè 1/6.

Quale è la densità di T calcolata in 2?

X\Y	1	2	3	4	5	6	
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

Tabella 11. Distribuzione congiunta del risultato del lancio di due dadi|con evidenziate le probabilità di ottenere come differenza 0.

4. Valore atteso e varianza della somma e del prodotto di due variabili aleatorie. Covarianza fra due variabili aleatorie.

Quando X e Y sono due variabili aleatorie definite sullo stesso spazio campionario si ha:

a. E(X+Y)=E(X)+E(Y) Non è quindi necessario conoscere la densità congiunta di X e Y.
b. Se X e Y sono **indipendenti** allora E(XY)=E(X)E(Y)

Per la varianza della somma non vale l'analogo dell'identità a. In questo caso è necessario conoscere anche la covarianza fra X e Y. Infatti, svolgendo i calcoli, si ottiene:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2\mathrm{Cov}(X,Y) + \mathrm{Var}(Y),$$

dove

$$\mathrm{Cov}(X,Y)=E((X-E(X))(Y-E(Y)))$$

è la **covarianza** fra le due variabili X e Y. Utilizzando la densità congiunta si può vedere che

$$\mathrm{Cov}(X,Y)=\sum_{i,j}(x_i - E(X))(y_j - E(Y))f_{XY}(x_i, y_j)$$

E ancora Cov(X,Y)=E(XY)-E(X)E(Y).

La covarianza può assumere valori reali positivi (in questo caso X e Y sono **positivamente correlate**), reali negativi (X e Y sono **negativamente correlate**) e può risultare nulla (X e Y sono **non correlate**). L'indice

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

detto **correlazione di X e Y**, è un numero compreso fra –1 e 1.
Tale valore *misura* il legame lineare fra X e Y. In particolare se $\rho(X,Y)=\pm 1$ allora X e Y sono ottenute una dall'altra per trasformazione lineare.

Nel caso in cui X e Y sono indipendenti la covarianza è 0.
Quando X e Y NON sono indipendenti, bisogna calcolare esplicitamente E(XY) (con la densità congiunta), E(X) e E(Y).

ESEMPIO 6. Riprendiamo l'Esempio 2 del numero di figli per famiglia. Quanto vale la covarianza fra F e M?
Riportiamo la densità congiunta e le marginali di F e M, cioè la Tabella 5

F\M	0	1	2	3	totale F
0	0.1500	0.1000	0.0875	0.0375	0.3750
1	0.1000	0.1750	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2000
3	0.0375	0	0	0	0.0375
totale M	0.375	0.3875	0.2000	0.0375	

Il valore atteso della variabile aleatoria FM è:

E(FM) = 0 x 0 x 0.1500 + 0 x 1 x 0.1000 + 0 x 2 x 0.0875 + 0 x 3 x 0.0375 +
1 x 0 x 0.1000 + 1 x 1 x 0.1750 + 1 x 2 x 0.1125 + 1 x 3 x 0 +
2 x 0 x 0.0875 + 2 x 1 x 0.1125 + 2 x 2 x 0 + 2 x 3 x 0 +
3 x 0 x 0.0375 + 3 x 1 x 0 + 3 x 2 x 0 + 3 x 3 x 0
= 1 x 0.1750 + 2 x 0.1125 + 2 x 0.1125 = 0.625

I valori attesi di F e M sono:
E(F) = E(M) = 0 x 0.375 + 1 x 0.3875 + 2 x 0.2000 + 3 x 0.0375 = 0.9
Quindi la covarianza fra F e M è: Cov(F, M) = E(FM) - E(F) E(M) = 0.625 $-(0.9)^2$ = - 0.185

ESEMPIO 7. Osservate la tabella della distribuzione congiunta delle variabili aleatorie dell'Esempio 4 sull'acquisto del pane. Che cosa potete dire relativamente al segno della covarianza? Giustificate la risposta e poi verificate la sua correttezza facendo i conti.

E1. Esperienza. Simulazione della somma di due variabili aleatorie indipendenti: il risultato del lancio di due dadi

Simulate 1000 lanci di due dadi (in due colonne).
Considerando che sono indipendenti, calcolate la somma dei due risultati.
Calcolate la distribuzione della somma e confrontatela con i risultati teorici ottenuti.

E2. Esperienza. Simulazione della somma di due variabili aleatorie non indipendenti

Simulate 1000 realizzazioni del della somma di due variabili aleatorie X e Y che hanno la seguente densità congiunta.

X\Y	0	1	2
-1	0.15	0.10	0.05
0	0.10	0.20	0.05
1	0.10	0.15	0.10

Per simulare la distribuzione congiunta ci sono due strategie:

1. Prima si simula la variabile X e, a seconda del risultato, si simula la variabile Y con le probabilità condizionata riga (da costruire). [Questa strategia è piuttosto lunga]
2. Si considera una variabile aleatoria con 9 valori possibili (le coppie di X e Y) e con la densità della congiunta. Si calcola quindi la funzione di distribuzione cumulata.

$t=(x,y)$	$f_T(t)$	$F_T(t)$	$z=x+y$
(-1,0)	0.15	0.15	-1
(-1,1)	0.10	0.25	0
(-1,2)	0.05	0.30	1
(0,0)	0.10	0.40	0
(0,1)	0.20	0.60	1
(0,2)	0.05	0.65	2
(1,0)	0.10	0.75	1
(1,1)	0.15	0.90	2
(1,2)	0.10	1	3

Generate quindi 1000 realizzazioni di T. Nella scheda 3 abbiamo visto come simulare le realizzazioni di una qualunque variabile aleatoria discreta bidimensionale (X,Y), funzione di X e Y. In particolare nel caso della somma, si associa a ogni realizzazione di T la somma degli x e y corrispondenti. Si calcolano quindi le frequenze dei valori ottenuti.

Confrontate quanto ottenuto con i valori (teorici) della densità di probabilità di X+Y.