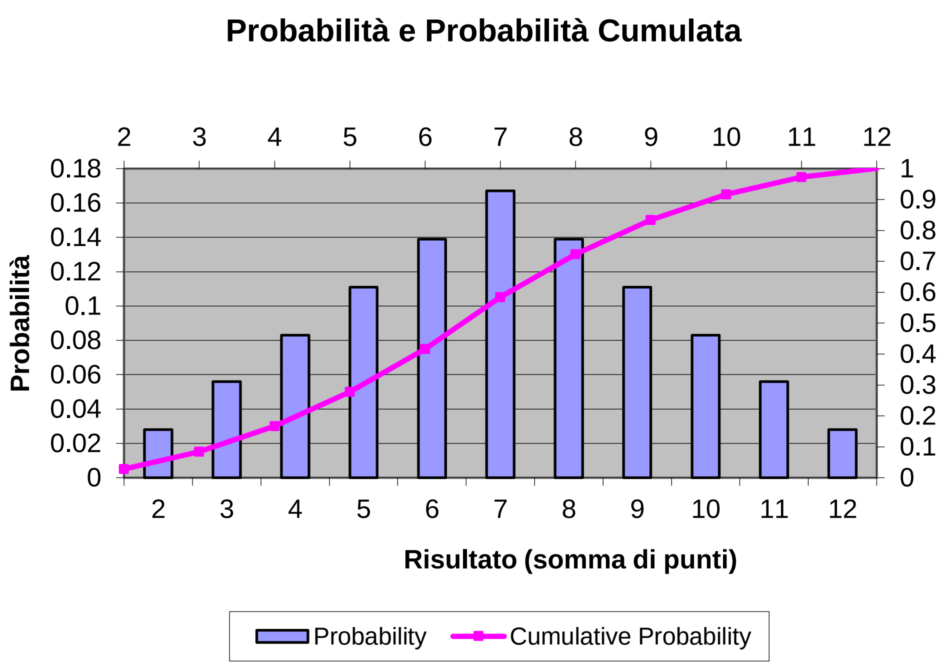
Esercizi I.(a): Revisione di statistica descrittiva e elementi di calcolo delle probabilità

1) Supponete si traggano due dadi e lasciate che M indichi la somma del numero di punti sui due dadi (quindi M è un numero compreso tra 2 e 12) non truccati. La tabella seguente elenca tutti i possibili risultati per la variabile casuale M insieme ai relativi valori della distribuzione delle probabilità e della distribuzione delle probabilità cumulative.

Risultato (somma di punti)	2 = 1 + 1	3 = 1 + 2 = 2 + 1	4 = 1 + 3 = 3 + 1 = 2 + 2	5 = 1 + 4 = 4 + 1 = 2 + 3 = 3 + 2	6 = 1 + 5 = 5 + 1 = 2 + 4 = 4 + 2 = 3 + 3	7 = 1 + 6 = 6 + 1 = 2 + 5 = 5 + 2 = 3 + 4 = 4 + 3	8 = 2 + 6 = 6 + 2 = 3 + 5 = 5 + 3 = 4 + 4	9 = 3 + 6 = 6 + 3 = 4 + 5 = 5 + 4	10 = 4 + 6 = 6 + 4 = 5 + 5	11 = 5 + 6 = 6 + 5	12 = 6 + 6
Distribuzione di probabilità	0,028 $=\frac{1}{6}\times\frac{1}{6}$	0,056 $=\frac{1}{6}\times\frac{1}{6}\times 2$	0,083 $=\frac{1}{6}\times\frac{1}{6}\times 3$	0,111 $=\frac{1}{6}\times\frac{1}{6}\times 4$	0,139 $=\frac{1}{6}\times\frac{1}{6}\times 5$	0,167 $=\frac{1}{6}\times\frac{1}{6}\times 6$	0,139 $=\frac{1}{6}\times\frac{1}{6}\times 5$	0,111 $=\frac{1}{6}\times\frac{1}{6}\times 4$	0,083 $=\frac{1}{6}\times\frac{1}{6}\times 3$	0,056 $=\frac{1}{6}\times\frac{1}{6}\times 2$	0,028 $=\frac{1}{6}\times\frac{1}{6}$
Distribuzione di probabilità cumulativa	0,028	0,083	0,167	0,278	0,417	0,583	0,722	0,833	0,912	0,972	1,000

A) Disegnare entrambe le distribuzioni mostrate nella tabella di cui sopra.

Risposta:



B) Calcolare il valore atteso e la deviazione standard per M.

Risposta: 7,0; 2,42.

	2	0.028	0.056	-5.01	25.07
	3	0.056	0.168	-4.01	16.06
	4	0.083	0.332	-3.01	9.04
	5	0.111	0.555	-2.01	4.03
	6	0.139	0.834	-1.01	1.01
	7	0.167	1.169	-0.01	0.00
	8	0.139	1.112	0.99	0.99
	9	0.111	0.999	1.99	3.97
	10	0.083	0.830	2.99	8.96
	11	0.056	0.616	3.99	15.94
	12	0.028	0.336	4.99	24.93
	Somma	1.00	7.01		
	Media	7.01			
	Varianza	5.85			
	deviazione standard	2.42			

2) Qual è la probabilità dei risultati seguenti?

a. $\Pr(M = 7)$
b. $\Pr(M = 2 \text{ o } M = 10)$
c. $\Pr(M = 4 \text{ o } M \neq 4)$
d. $\Pr(M = 6 \text{ e } M = 9)$
e. $\Pr(M < 8)$
f. $\Pr(M = 6 \text{ o } M > 10)$

Risposta:

(a) 0,167 ovvero $\frac{6}{36} = \frac{1}{6}$;

(b) 0,111 ovvero $\frac{4}{36} = \frac{1}{9}$;

(c) 1;

(d) 0;

(e) 0,583;

(f) 0,222 ovvero $\frac{8}{36} = \frac{2}{9}$.

3) Si acceda al sito del libro di testo di Stock & Watson al link seguente: **http://www.pearsonglobaleditions.com/Stock**
Si ottenga il dataset CPS (ch8_cps. xls) relativo ad un esempio trattato nel capitolo 8, aprendolo in un programma di foglio di calcolo come Excel. Per l'esercizio, utilizzare le prime 500 osservazioni relative ai guadagni orari medi (*Ahe*). Si utilizzino statistiche riepilogative, quali media, meadiana, varianza e asimmetria, per descrivere la distribuzione dei guadagni. Si produce inoltre una distribuzione di frequenza ("istogramma") utilizzando classi di misura dei guadagni di ragionevoli dimensioni.

Risposta:

Ahe	
Media	19,79
Errore standard	0,51
Mediana	16,83
Modalità	19,23
Deviazione standard	11,49
Varianza del campione	131,98
Curtosi	0,23
Asimmetria	0,96
Gamma	58,44
Minimo	2,14
Massimo	60,58
Somma	9897,45
Conteggio	500,00

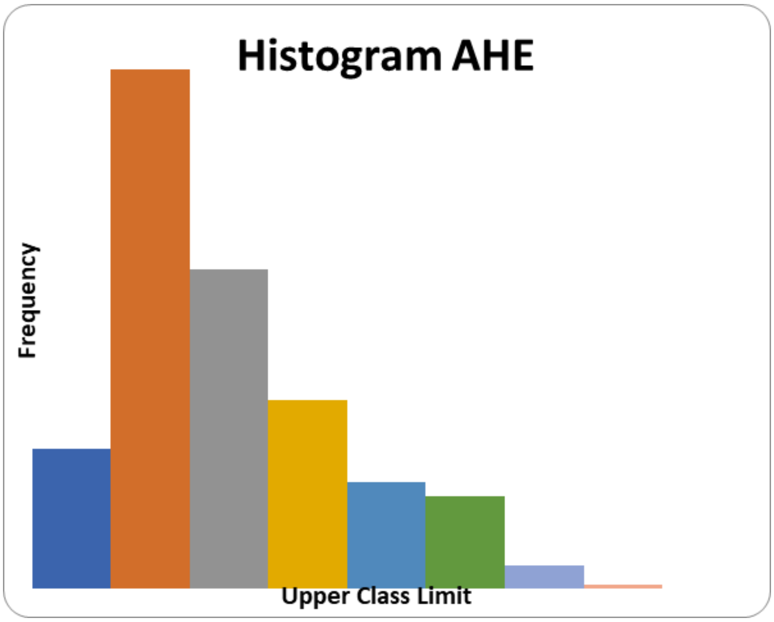
La media è \$19,79. La mediana (\$16,83) è inferiore alla media, suggerendo che la media è stata tirata su da individui con guadagni orari media relativamente elevati. Ciò è confermato dalla misura di asimmetria, che è positiva e quindi suggerisce una distribuzione con una coda lunga alla destra. La varianza è $\$^2$131,96, mentre la deviazione standard è \$11,49.

Per generare la distribuzione di frequenza in Excel, è necessario innanzitutto definire il numero di classi di misura e dunque l'ampiezza dei relativi intervalli. Una volta che avete deciso sul numero, la differenza tra il minimo e il massimo nei dati suggerisce la larghezza della classe. Nel seguito, ho optato per 8 intervalli con una ampiezza di ciascuna classe pari a \$8 (poiché i salari minimi in California sono attualmente \$8 e circa lo stesso in altri Stati degli Stati Uniti).

La tabella seguente riporta le frequenze assolute, e le relative frequenze possono essere calcolate in modo semplice e ovvio.

Estremo superiore delle classi di misura ("bins", che non vuol dire "bidoni"!)	*Frequenza*	*Frequenza relativa*
8	50	0,1
16	187	0,374
24	115	0,23
32	68	0,136
40	38	0,076
48	33	0,066
56	8	0,016
66	1	0,002
> 66	0	

La sostituzione delle frequenze relative nella tabella dell'istogramma produce quindi il grafico seguente (dopo aver eliminato gli spazi tra le barre).



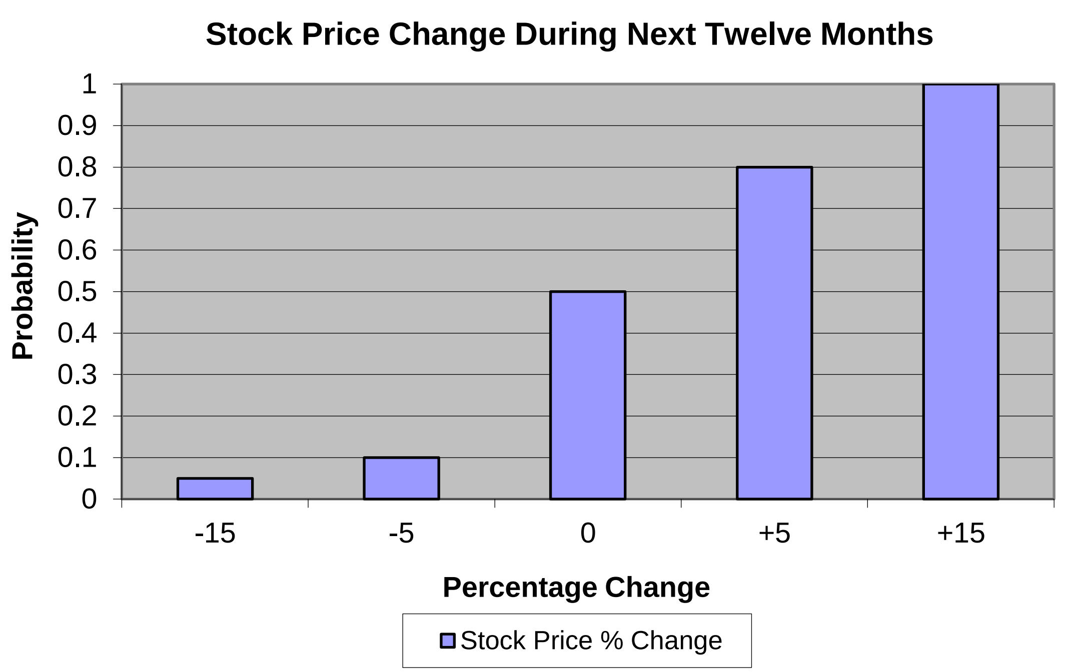
4) Nel considerare l'acquisto di un certo titolo azionario, si attribuiscono le seguenti probabilità di possibili variazioni del prezzo delle azioni nel corso dell'anno successivo.

Cambio di prezzo del titolo azioanrio durante i prossimi dodici mesi (%)	**Probabilità**
+ 15	0,2
+ 5	0,3
0	0,4
-5	0,05
-15	0,05

Qual è il valore atteso, la varianza e la deviazione standard? Qual è il risultato più probabile? Si produca uno schizzo della funzione di distribuzione cumulativa.

.

Risposta: E(Y) = 3,5; $\sigma_Y^2 = 8{,}49$; σ_Y = 2,91; molto probabilmente: 0.



5) Le probabilità e le frequenze relative sono concetti legati l'uno all'altro in quanto la probabilità di un risultato è la frequenza con cui il risultato si verifica nel lungo periodo (ovvero in un numero di campionamenti infiniti dalla stessa popolazione). Quindi i concetti delle distribuzioni congiunte, marginali e di probabilità condizionale derivano dai concetti relativi delle distribuzioni di frequenza.

Supponiamo dunque che siate interessati a indagare il rapporto tra l'età dei capo-famiglia e i guadagni settimanali delle famiglie. Si considerino a tale scopo dei dati sul numero di occorrenze di diversi valori di queste variabili raggruppate per età e reddito. I dati sono relativi a 1.744 individui. Si pensi a questi individui come a una popolazione che si desidera descrivere, piuttosto che un campione tramite il quale dedurre i tratti di una popolazione più grande. Dopo aver ordinato i dati in classi rispetto ai valori assunti dalle due variabili, si genera la tabella seguente:

Tabella 1
Frequenze assolute congiunte di età e reddito per una popolazione di 1.744 famiglie

		Età del capofamiglia				
		X_1	X_2	X_3	X_4	X_5
	Reddito familiare	16-Under 20	20-Under 25	25-Under 45	45-Under 65	65 e >
Y_1	\$0-sotto \$200	80	76	130	86	24
Y_2	\$200-sotto \$400	13	90	346	140	8
Y_3	\$400-sotto \$600	0	19	251	101	6
Y_4	\$600-sotto \$800	1	11	110	55	1
Y_5	\$800 e >	1	1	108	84	2

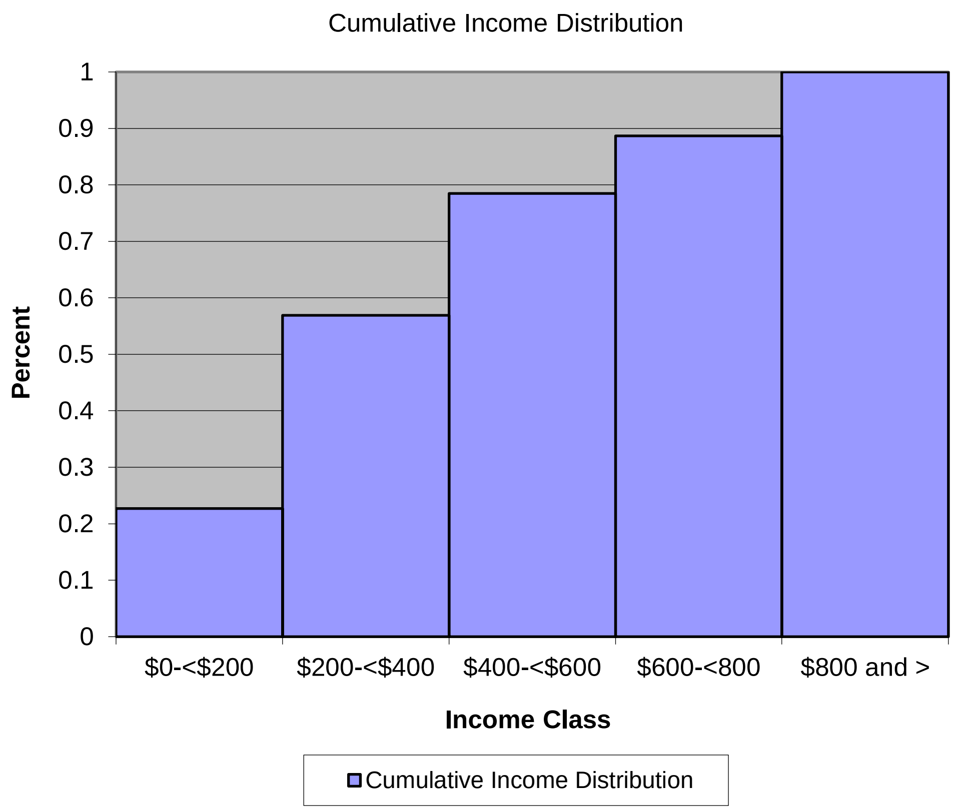
La media del gruppo di reddito che guadagna non meno di \$800 è \$1.050.

a. Calcolare le frequenze relative congiunte e marginali. Tracciare la distribuzione del reddito cumulativa.

Risposta: Le frequenze relative congiunte e marginali sono indicate nella tabella in basso. Quelle congiunte sono nelle celle all'interno del bordo. Quelle marginali del reddito familiare sono nella colonna esterna a destra che riporta i totali per riga. Quelle marginali dell'età sono nella riga esterna in basso che riporta i totali per colonna. Ad esampio, il 5,2 per cento (= 90/1.744) degli individui sono di età compresa tra 20 e 24 anni e guadagnano tra \$200 e \$400, mentre il 21,6 per cento (= (0+19+251+101+6)/ 1.744 = 377/1.744) degli individui guadagna fra \$400 e \$600.

Tabella 2
Frequenze relative congiunte e marginali di età e reddito, 1.744 famiglie

		Età del capofamiglia					
		X_1	X_2	X_3	X_4	X_5	
Reddito familiare		16-Under 20	20-Under 25	25-sotto 45	45-sotto 65	65 e >	**Totale**
Y_1	\$0-sotto \$200	0,046	0,044	0,075	0,049	0,014	0,227
Y_2	\$200-sotto \$400	0,007	0,052	0,198	0,080	0,005	0,342
Y_3	\$400-sotto \$600	0,000	0,011	0,144	0,058	0,003	0,216
Y_4	\$600-sotto \$800	0,001	0,006	0,063	0,032	0,001	0,102
Y_5	\$800 e >	0,001	0,001	0,062	0,048	0,001	0,112
	Totale	0,055	0,114	0,542	0,267	0,024	



Qui di seguito è disponibile un link a un foglio elettronico di MS Excel con i calcoli:

		X_1	X_2	X_3	X_4	X_5	
		16-Under 20	20-Under 25	25-sotto 45	45-sotto 65	65 e >	
Y_1	$0-sotto $200	4.6%	4.4%	7.5%	4.9%	1.4%	22.7%
Y_2	$200-sotto $40	0.7%	5.2%	19.8%	8.0%	0.5%	34.2%
Y_3	$400-sotto $60	0.0%	1.1%	14.4%	5.8%	0.3%	21.6%
Y_4	$600-sotto $80	0.1%	0.6%	6.3%	3.2%	0.1%	10.2%
Y_5	$800 e >	0.1%	0.1%	6.2%	4.8%	0.1%	11.2%
		5.4%	11.3%	54.2%	26.7%	2.4%	100.0%

b. Calcolare le frequenze relative di reddito condizionali per le due categorie di età 16-Under 20, e 45-under 65. Calcolare il reddito medio delle famiglie per entrambe le categorie di età.

Risposta: Le frequenze calcolate sono riportate nella tabella qui sotto.
Sulla base di tali frequenze, il reddito medio delle famiglie per la categoria di età 16-Under 20 è di circa $144, mentre è di circa $489 per la categoria di età 45-under 65.

Tabella 3
Frequenze relative di reddito per le categorie di età "16-Under 20" e "45-sotto 65"

		Età del capofamiglia	
		X_1	X_4
Reddito familiare		16-Under 20	45-sotto 65
Y_1	$0-sotto $200	84.21%	18.45%
Y_2	$200-sotto $400	13.68%	30.04%
Y_3	$400-sotto $600	0.00%	21.67%
Y_4	$600-sotto $800	1.05%	11.80%
Y_5	$800 e >	1.05%	18.03%

Qui di seguito è disponibile un link a un foglio elettronico di MS Excel con i calcoli:

		X_1	X_4
		16-Under 20	45-sotto 65
Y_1	$0-sotto $200	83.64%	18.35%
Y_2	$200-sotto $400	12.73%	29.96%
Y_3	$400-sotto $600	0.00%	21.72%
Y_4	$600-sotto $800	1.82%	11.99%
Y_5	$800 e >	1.82%	17.98%
		1	1

c. Se il reddito delle famiglie e l'età del capo della famiglia fossero distribuiti in modo indipendente, che cosa vi aspettereste riguardo al rapporto tra le due distribuzioni condizionali di cui sopra e la corrispondente distribuzione marginale? È questo il caso?

Risposta: Dovrebbero essere identiche. Chiaramente non è questo il caso. Si confrontino ad esempio la distribuzione condizionale del reddito per le due classi di età di cui sopra con la distribuzione nella ultima colonna di Tabella 2. Quindi le due variabili (reddito delle famiglie e l'età del capo della famiglia non sono indipendentemente distribuiti).

d. Il tuo libro di testo ti ha dato una definizione alternativa di indipendenza che non comporta distribuzioni di frequenza relativa condizionale. Qual è la definizione? Pensi che l'età e il reddito sono indipendenti qui, utilizzando questa definizione?

Risposta: La definizione alternativa è basata sulla condizione seguente: $\Pr(Y = y, X = x) = \Pr(Y = y)\Pr(X = x)$ per ciascun valore di x e y. Possiamo verificare se vale questa condizione moltiplicando le due probabilità marginali per ciascuna possibile coppia di valori di X e Y e vedere se il risultato è uguale alla loro probabilità congiunta. Per esempio, poiché $\Pr(Y = Y_3) = 0.216$ e $\Pr(X = X_3) = 0.542$, il loro prodotto è 0,117, che non è uguale alla loro probabilità congiunta (ovvero 0,144). Per chi abbia già studiato statistica inferenziale, si ricorda che, poiché stiamo trattando i dati come riferiti a una popolazione piuttosto che a un campione, non dobbiamo preoccuparci di verificare se la differenza tra 0,117 e 0,144 sia dovuta a errore di campionamento (ovevro se siano differenti in maniera statisticamente significativa).

e. Si calcolino i valori attesi del reddito condizionando all'appartenenza della famiglia alla prima ed alla quarta classe di misura dell'età del capofamiglia. Ai fini del calcolo dei valori attesi richiesti, si attribuisca a ciascuna classe di misura del reddito un valore del reddito stesso pari al valore centrale della classe di misura stessa con l'eccezione della ultima classe di mura, a cui si attribuisca un valore di $900.

Alla pagina che segue è disponibile un link a un foglio elettronico di MS Excel con i calcoli relativi al valore atteso, oltra che alle sotto-domande precedenti:

			X_1	X_2	X_3	X_4	X_5	
			16-Under 20	20-Under 25	25-sotto 45	45-sotto 65	65 e >	
	Y_1	$0-sotto $200	80	76	130	86	24	
	Y_2	$200-sotto $400	13	90	346	140	8	
	Y_3	$400-sotto $600	0	19	251	101	6	
	Y_4	$600-sotto $800	1	11	110	55	1	
	Y_5	$800 e >	1	1	108	84	2	
		Somma	1,744					

			X_1	X_2	X_3	X_4	X_5	
			16-Under 20	20-Under 25	25-sotto 45	45-sotto 65	65 e >	
	Y_1	$0-sotto $200	4.6%	4.4%	7.5%	4.9%	1.4%	22.7%
	Y_2	$200-sotto $400	0.7%	5.2%	19.8%	8.0%	0.5%	34.2%
	Y_3	$400-sotto $600	0.0%	1.1%	14.4%	5.8%	0.3%	21.6%
	Y_4	$600-sotto $800	0.1%	0.6%	6.3%	3.2%	0.1%	10.2%
	Y_5	$800 e >	0.1%	0.1%	6.2%	4.8%	0.1%	11.2%
			5.4%	11.3%	54.2%	26.7%	2.4%	100.0%

			X_1	X_4
			16-Under 20	45-sotto 65
100	Y_1	$0-sotto $200	84.2%	18.5%
300	Y_2	$200-sotto $400	13.7%	30.0%
500	Y_3	$400-sotto $600	0.0%	21.7%
700	Y_4	$600-sotto $800	1.1%	11.8%
900	Y_5	$800 e >	1.1%	18.0%
			100.0%	100.0%
			E(Y|X=X_1)	E(Y|X=X_4)
			142.11	461.80

6) Avete probabilmente sentito parlare della cosiddetta teoria "catch-up" sostenuta da storici dell'economia, per cui le nazioni che sono più indietro nel reddito pro-capite tendono a crescere più velocemente in seguito. Per mettere la teoria alla prova, si raccolgono i dati sul reddito pro capite per due anni, 1960 e 1990, per 24 paesi OCSE, espresso come frazione del reddito pro-capite degli USA. Si pensi a questi paesi come una popolazione che si desidera descrivere, piuttosto che un campione da cui si desidera dedurre il comportamento di una popolazione più grande. I dati rilevanti per questa domanda sono i seguenti:

Y	X_1	X_2	$Y \times X_1$	Y^2	X_1^2	X_2^2
0,023	0,770	1,030	0,018	0,00053	0,593	1,0609
0,014	1,000	1,000	0,014	0,00020	1,000	1,0000
....	....	....	....	....	....	....
0,041	0,200	0,450	0,008	0,00168	0,040	0,2025
0,033	0,130	0,230	0,004	0,00109	0,017	0,0529
0,625	13,220	17,800	0,294	0,01877	8,529	13,9164

Nella tabella qui sopra, X_1 e X_2 sono il reddito pro-capite espresso come frazione di quello degli Stati Uniti negli anni 1960 e 1990, rispettivamente, e Y ne è il tasso di crescita annuo medio nel periodo 1960-1990. I numeri nell'ultima riga rappresentano le somme delle colonne di cui sopra.

A. Calcolare la media, varianza e la deviazione standard di Y, X_1 e X_2. Perchè si possa concludere che un effetto catch-up sia presente, quale rapporto ci deve essere tra le due deviazioni standard? È questo il caso?

<u>Risposta</u>: Le medie di Y, X_1 e X_2 sono, rispettivamente, 0,625/24 = 0,026, 13,220/24 = 0,55 e 17,800/24 = 0,74.

Le varianze di Y, X_1, X_2 si possono calcolare utilizzando la formula
$Var(X) = E(X^2) - E(X)^2$,

dove X è una qualunque variabile aleatoria con momenti primi e secondi definiti.

Esse sono dunque, rispettivamente,

$Var(Y) = E(Y^2) - E(Y)^2 = \frac{0,01877}{24} - 0,026^2 = 0,0001$

$Var(X_1) = 0,0520$

$Var(X_2) = 0,0298$

Le deviazioni standard richieste sono dunque $\sigma(Y) = \sqrt{0,0001} = 1,0299$, con deviazioni standard di 0,2279 e 0,1726. Perchè si possa concludere che un effetto catch-up sia presente, la deviazione standard dovrebbe ridursi nel tempo. Questo è effettivamante il caso.

B. Calcolare la covarianza e correlazione tra Y e X_1. Quale segno devono avere perché siano prova di un effetto catch-up?

<u>Risposta</u>: La covarianza si può calcolate utilizzando la formula

$$Cov(X,Y) = E(XY) - E(X)E(Y),$$

dove X e Y sone due qualunque variabili aleatorie con momenti primi e secondi definiti.

La covarianza tra Y e X_1 è dunque

$$\begin{aligned} Cov(X_1, Y) &= E(X_1 Y) - E(X_1)E(Y) \\ &\approx 0{,}294/24 - 0{,}55 \times 0{,}026 \\ &= -0.002095 \end{aligned}$$

Il coefficiente di correlazione tra X_1 e Y è

$$\rho(X_1, Y) := \frac{Cov(X_1, Y)}{\sigma(X_1)\sigma(Y)} \approx \frac{-0.002095}{0.01 \times 0.228} = -0.90$$

Il coefficiente di correlazione è dunque -0.90 (-0,88 se avete lavorato solo con tre decimali). Deve essere negativo perché si possa parlare di un effetto catch-up. Se i paesi che erano relativamente avanti nel periodo iniziale in termini di reddito pro-capite crescono di relativamente meno nel tempo, poi alla fine i ritardatari recuperano (catch-up).

Il solito foglio Excel con le soluzioni è disponibile qui di seguito:

	Y	X_1	X_2	$Y \times X_1$	Y^2	X_1^2	X_2^2
	0.023	0.770	1030.000	0.018	0.001	0.593	10609.000
	0.014	1.000	1.000	0.014	0.000	1.000	1.000
	....	....	....	....	....	....	....
	0.041	0.200	0.450	0.008	0.002	0.040	0.203
	0.033	0.130	0.230	0.004	0.001	0.017	0.053
somme per colonna	0.625	13.220	17.800	0.294	0.019	8.529	13.916
medie	0.026	0.551	0.742	0.012	0.001	0.355	0.580
varianze	0.010%	0.052	0.030				
deviazioni standard	0.010	0.228	0.173				
E(X_1,Y)	1.225%						
E(X_1)E(Y)	1.434%						
covarianza(X_1,Y)	-0.002095	-0.209%					
correlazione(X_1,Y)	-0.901449						

7) Per volontà di Alfred Nobel, ci sono cinque premi Nobel assegnati ogni anno per meriti eccezionali nella chimica, nella fisica, nella fisiologia o nella medicina, nella letteratura e nella pace. Nel 1968, la banca di Svezia aggiunse un premio in scienze economiche in memoria di Alfred Nobel. La tabella di seguito descrive la distribuzione di probabilità congiunta relativa ai caratteri "essere destinatari di un premio Nobel in economia o in una delle altre cinque discipline" e "cittadinanza" dei destinatari di premi Nobel nel periodo 1969-2001. Si pensi ai dati come volti a descrivere una popolazione (quella dei destinatari), piuttosto che un campione da cui si desidera dedurre il comportamento di una popolazione più grande.

Distribuzione congiunta di premi Nobel a seconda della disciplina (discipline economiche e non economiche) e della cittadinanza, 1969-2001

	Cittadino degli Stati Uniti ($Y=0$)	Cittadino non statunitense ($Y=1$)	Totale
Premio Nobel di economia ($X=0$)	0,118	0,049	0,167
Fisica, chimica, medicina, letteratura e premio Nobel di pace ($X=1$)	0,345	0,488	0,833
Totale	0,463	0,537	1,00

A. Calcolare $E(Y)$ e interpretare il numero risultante.

Risposta: $E(Y)=0.537$, ovvero il 53,7 per cento dei vincitori del premio Nobel erano cittadini non statunitensi.

B. Calcola e interpreta $E(Y \mid X=1)$ E $E(Y \mid X=0)$.

Risposta: $E(Y \mid X=1)=0.586$ vuol dire che il 58,6 per cento dei vincitori del premio Nobel nelle discipline non di economia erano cittadini non-USA. $E(Y \mid X=0)=0.293$ vuol dire che il 29,3 per cento dei vincitori del premio Nobel di economia erano cittadini non-USA.

C. Un vincitore del premio Nobel casualmente selezionato riporta che è un cittadino non statunitense. Qual è la probabilità che questa persona abbia vinto il premio Nobel per l'economia? Qual è la probabilità che abbia vinto un premio Nobel nelle altre cinque discipline?

Risposta: C' è una probabilità di 9,1 per cento che abbia vinto il premio Nobel per l'economia e una probabilità di 90,9 per cento che abbia vinto un premio Nobel in una delle altre cinque discipline.

D. Mostrare che la distribuzione congiunta sarebbe simile se le due categorie fossero indipendenti.

Risposta:
Distribuzione congiunta di premi Nobel in discipline economiche e non-economia, e cittadinanza, 1969-2001, sotto assunzione di indipendenza

	Cittadino degli Stati Uniti ($Y = 0$)	Non= Cittadino degli Stati Uniti ($Y = 1$)	Totale
Premio Nobel di economia ($X = 0$)	0,077	0,090	0,167
Fisica, chimica, medicina, letteratura e premio Nobel di pace ($X = 1$)	0,386	0,447	0,833
Totale	0,463	0,537	1,00

Per i calcoli dettagliati, si veda il foglio di calcolo al link di seguito:

		Cittadino USA	Cittadino non-USA	Totale
	Premio di economia	0,118	0,049	0,167
	Premio non di economi	0,345	0,488	0,833
	Totale	0,463	0,537	1,00
	Y	0	1	
X		Cittadino USA (Y = 0)	Cittadino non-USA (Y = 1)	Totale
0	Premio di economia (X = 0)	0.118	0.049	0.167
1	Premio non di economia (X = 1)	0.345	0.488	0.833
	Totale	0.463	0.537	1
A.	E(Y)	0.537		
B.	E(Y|X=0)	0.293		
	E(Y|X=1)	0.586		
C.	Pr(X=0|Y=1)	0.091		
	Pr(X=1|Y=1)	0.909		
D.				
	Y	0	1	
X		Cittadino USA (Y = 0)	Cittadino non-USA (Y = 1)	Totale
0	Premio di economia (X = 0)	0.077	0.090	0.167
1	Premio non di economia (X = 1)	0.386	0.447	0.833
	Totale	0.463	0.537	1

Appendice: Esercizi addizionali

2.7 In una data popolazione di coppie maschio/femmina, dove entrambi sono percettori di reddito, le retribuzioni maschili hanno una media di 50.000$ all'anno e una deviazione standard di 15.000$. Le retribuzioni femminili hanno una media di 48.000$ all'anno e una deviazione standard di 13.000$. La correlazione tra le retribuzioni maschili e femminili all'interno di una coppia è di 0,90. Siano C le retribuzioni complessive di una coppia selezionata a caso.

a. Qual è la media di C?

b. Qual è la covarianza tra la retribuzione maschile e quella femminile?

c. Qual è la deviazione standard di C?

d. Si convertano le risposte ai punti (a)-(c) da $ (dollari) a € (euro).

Risposta:

2.7. Utilizzando una notazione ovvia, $C = M + F$; quindi $\mu_C = \mu_M + \mu_F$ e $\sigma_C^2 = \sigma_M^2 + \sigma_F^2 + 2\,\mathrm{cov}(M, F)$. Ciò implica

(a) $\mu_C = 50 + 48 = \$98.000$ per anno.

(b) $\mathrm{corr}(M, F) = \dfrac{\mathrm{cov}(M, F)}{\sigma_M \sigma_F}$, quindi $\mathrm{cov}(M, F) = \sigma_M \sigma_F \mathrm{corr}(M, F)$. Quindi $\mathrm{cov}(M, F) = 15 \times 13 \times 0{,}9 = 175{,}50$, dove le unità sono migliaia di dollari l'anno al quadrato.

(c) $\sigma_C^2 = \sigma_M^2 + \sigma_F^2 + 2\,\mathrm{cov}(M, F)$, perciò $\sigma_C^2 = 15^2 + 13^2 + 2(175{,}5) = 745$, e $\sigma_C = \sqrt{745} = 27{,}295$ migliaia di dollari l'anno.

(d) Per prima cosa occorre cercare il tasso di cambio euro/dollaro attuale sul Wall Street Journal, sul sito web della Federal Reserve o consultando altre fonti di informazioni finanziarie. Supponiamo che il tasso di cambio sia e (per esempio $e = 0{,}75$ euro per dollaro); quindi 1 dollaro vale e euro. La media è quindi $e \times \sigma_C$ (in migliaia di euro l'anno) e la deviazione standard è $e \times \sigma_C$ (in migliaia di euro l'anno). La correlazione è indipendente dall'unità di misura ed è invariata.

Nota: Lo scopo principale di questo esercizio è quello di rammentare che, per due qualunque variabili aleatorie X ed Y,

$$E(X+Y) = E(X) + E(Y) \qquad Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

$$Cov(X,Y) = Corr(X,Y)\sigma(X)\sigma(Y)$$

Qui sopra, si utilizza la notazione consueta per cui (con qualche variazione) di solito, per una data variabile, poniamo X, per il valore atteso si usa indifferentemente $E(X)$ oppure μ_X, per la varianza si usano i simboli σ_X^2 oppure $Var(X)$, per la correlazione con un'atra variabile Y si usa $\rho_{X,Y}$ oppure $Corr(X,Y)$ e per la covarianza si usa $\sigma_{X,Y}$ oppure $Cov(X,Y)$.

2.9 Siano X e Y variabili casuali discrete con la seguente distribuzione congiunta:

		Valore di Y				
		2	**4**	**6**	**8**	**10**
Valore di X	**3**	0,04	0,09	0,03	0,12	0,01
	6	0,10	0,06	0,15	0,03	0,02
	9	0,13	0,11	0,04	0,06	0,01

Ovvero, $\Pr(X = 3, Y = 2) = 0{,}04$ e così via.

a. Si calcolino la distribuzione di probabilità, la media e la varianza di Y.

b. Si calcolino la distribuzione di probabilità, la media e la varianza di Y data $X = 6$.

c. Si calcolino la covarianza e la correlazione tra X e Y.

Risposta:

2.9.

		Valore di Y					**Distribuzione di probabilità di X**
		2	**4**	**6**	**8**	**10**	
Valore di X	3	0,04	0,09	0,03	0,12	0,01	0,29
	6	0,10	0,06	0,15	0,03	0,02	0,36
	9	0,13	0,11	0,04	0,06	0,01	0,35
Distribuzione di probabilità di Y		0,27	0,26	0,22	0,21	0,04	1,00

(a) La distribuzione di probabilità è data nella tabella precedente. Dunque,[1]

$$E(Y) = 2 \times 0{,}27 + 4 \times 0{,}26 + 6 \times 0{,}22 + 8 \times 0{,}21 + 10 \times 0{,}04 = 4{,}98$$

$$E(Y^2) = 2^2 \times 0{,}27 + 4^2 \times 0{,}26 + 6^2 \times 0{,}22 + 8^2 \times 0{,}21 + 10^2 \times 0{,}04 = 30{,}6$$

$Var(Y) = E(Y^2) - E(Y)^2 = 30{,}6 - 4{,}98^2 \cong 30{,}6 - 24{,}8 = 5{,}8$ (dunque $\sigma_Y = \sqrt{5{,}8} = 2{,}41$)

(b) La probabilità di Y condizionata a $X = 6$, ovvero di $Y|X = 6$, è data nella tabella seguente

Valore di Y (y_i)	2	4	6	8	10	Totale (per riga)
Probabilità congiunta = $\Pr(Y = y_i \; and \; X = 6)$ A quanto sommano?	0,10	0,06	0,15	0,03	0,02	0,36

[1] Si ricordi che

$$Var(X) = E(X^2) - [E(X)]^2$$
$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

Probabilità condizionata = $\frac{\Pr(Y=y_i \; and \; X=6)}{\Pr(X=6)}$	0,10/0,36	0,06/0,36	0,15/0,36	0,03/0,36	0,02/0,36	1,00

Ovvero:

Valore di Y (y_i)	2	4	6	8	10	Totale (per riga)
Probabilità condizionata = $\frac{\Pr(Y=y_i \; and \; X=6)}{\Pr(X=6)}$	27.78%	16.67%	41.67%	8.33%	5.56%	100%

E dunque abbiamo:

$$E(Y|X=6) = 2 \times \frac{0,10}{0,36} + 4 \times \frac{0,06}{0,36} + 6 \times \frac{0,15}{0,36} + 8 \times \frac{0,03}{0,36} + 10 \times \frac{0,02}{0,36} \cong 4,94$$

$$E(Y^2|X=6) = 2^2 \times \frac{0,10}{0,36} + 4^2 \times \frac{0,06}{0,36} + 6^2 \times \frac{0,15}{0,36} + 8^2 \times \frac{0,03}{0,36} + 10^2 \times \frac{0,02}{0,36} \cong 29,6$$

$$Var(Y|X=6) = E(Y^2|X=6) - E(Y|X=6)^2 = 29,6 - 4,94^2 \cong 5,19$$

(c) $E(XY) = (3 \times 2 \times 0,04) + (3 \times 4 \times 0,09) + \cdots + (9 \times 10 \times 0,01) = 29,4$

$Cov(X,Y) = E(XY) - E(X)E(Y) = 29,4 - 6,18 \times 4,98 = -1,3764$

$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-1,3764}{\sqrt{5,7276 \times 5,8}}$

Il calcolo di $E(XY)$ di cui sopra è un esempio di quando dobbiamo ricorrere ai computer. Con carta, penna e calamaio ci vorrebbe un bel pò di lavoro ma con il calcolo matriciale e un software capace di implementarlo (praticamente tutti) ci vogliono pochi secondi. Ad esempio, si può cominciare dal creare le seguenti matrici

X = [3 6 9]'
Y = [2 4 6 8 10]'

$$P = \begin{pmatrix} 4\% & 9\% & 3\% & 12\% & 1\% \\ 10\% & 6\% & 15\% & 3\% & 2\% \\ 13\% & 11\% & 4\% & 6\% & 1\% \end{pmatrix}$$

E poi si calcola quanto segue:

$E(XY) = X'PY = 29,4$

Si provi ad implementare questo calcolo in Excel, dopo aver creato le necessarie matrici e vettori (ovvero, P, X e Y).

Nel linguaggio R, per chi cominci ad esserne curioso, ciò può essere svolto con i seguenti comandi:

```
> P = matrix(c(0.04,   0.09,   0.03,   0.12,   0.01,
+           0.1,0.06,   0.15,   0.03,   0.02,
+           0.13,       0.11,   0.04,   0.06,   0.01), nrow = 3,
ncol = 5,byrow=TRUE)
> P
     [,1] [,2] [,3] [,4] [,5]
[1,] 0.04 0.09 0.03 0.12 0.01
[2,] 0.10 0.06 0.15 0.03 0.02
[3,] 0.13 0.11 0.04 0.06 0.01
> Y=c(2,4,6,8,10)
> X=c(3,6,9)
> t(X)%*%P%*%Y
     [,1]
[1,] 29.4
> E_XY = t(X)%*%P%*%Y
```

Per il calcolo della covarianza e della varianza di X, sempre in R, si può procedere come di seguito indicato:

```
> E_X=(rowSums(P))%*%X
> E_X
     [,1]
[1,] 6.18

> E_Y=t(colSums(P))%*%Y
> E_Y
     [,1]
[1,] 4.98

> cov_XY = E_XY - E_X * E_Y
> cov_XY
        [,1]
[1,] -1.3764

> E_XX = t(X^2)%*%rowSums(P)
> E_XX
      [,1]
[1,] 43.92
> V_X = E_XX - E_X^2
> V_X
       [,1]
[1,] 5.7276
```