STATISTICA DESCRITTIVA BIVARIATA

Si parla di Analisi Multivariata quando su ogni unità statistica, appartenente ad una determinata popolazione, si rileva un certo numero ***s*** di caratteri $X_1, X_2, \ldots, X_s$.

Si parla di Analisi Bivariata quando su ogni unità statistica, appartenente ad una determinata popolazione, si rilevano due caratteri X e Y.
Può trattarsi di due caratteri qualitativi (ovvero mutabili), o di due caratteri quantitativi (ovvero variabili), oppure di un carattere qualitativo e di un carattere quantitativo.

Distribuzione bivariata semplice

X	x_1	x_2	…	x_i	…	x_n
Y	y_1	y_2	…	y_i	…	y_n

Distribuzione bivariata doppia o congiunta: TABELLA A DOPPIA ENTRATA (TABELLA DI CONTINGENZA)

	Y						
X	Y_1	Y_2	…	Y_j	…	Y_h	
X_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1h}	$f_{1\bullet}$
X_2	f_{21}	f_{22}	…	f_{2j}	…	f_{2h}	$f_{2\bullet}$
:	:	:		:		:	:
:	:	.		:		:	:
X_i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{ih}	$f_{i\bullet}$
:	:	:		:		:	:
:	:	:		:		:	:
X_k	f_{k1}	f_{k2}	…	f_{kj}	…	f_{kh}	$f_{k\bullet}$
	$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet j}$	…	$f_{\bullet h}$	N

f_{ij} FREQUENZA CONGIUNTA **ASSOLUTA**: è il numero delle volte con cui la coppia di modalità (x_i, y_j) si presenta, ovvero la frequenza con la quale, su di un'unità statistica, il carattere X assume la modalità x_i e contemporaneamente il carattere Y assume la modalità y_j.

$f_{i\bullet} = \sum_{j=1}^{h} f_{ij}$ Frequenza assoluta MARGINALE per Riga (: riferita alla riga i-ma): esprime la frequenza della modalità i-ma del carattere X per riga, senza tener conto delle modalità dell'altro carattere Y.

$f_{\bullet j} = \sum_{i=1}^{k} f_{ij}$ Frequenza assoluta MARGINALE per Colonna (: riferita alla colonna j-ma): esprime la frequenza della modalità j-ma del carattere Y per colonna, senza tener conto delle modalità dell'altro carattere X.

Da cui vale la seguente uguaglianza:

$$N = \sum_{j=1}^{h}\sum_{i=1}^{k} f_{ij} = \sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{h} f_{\bullet j}$$

Distribuzione marginale o ***Distribuzione univariata del carattere X***: data una distribuzione doppia di frequenze relativa alla rilevazione di N unità statistiche, si definisce distribuzione marginale del carattere X la distribuzione statistica semplice delle N unità statistiche secondo il carattere X.

X	f(x)
X_1	$f_{1\bullet}$
…	…
X_i	$f_{i\bullet}$
…	…
X_k	$f_{k\bullet}$
	N

Distribuzione marginale o ***Distribuzione univariata del carattere Y***: data una distribuzione doppia di frequenze relativa alla rilevazione di N unità statistiche, si definisce distribuzione marginale del carattere Y la distribuzione statistica semplice delle N unità statistiche secondo il carattere Y.

Y	f(y)
Y_1	$f_{\bullet 1}$
…	…
Y_j	$f_{\bullet j}$
…	…
Y_h	$f_{\bullet h}$
	N

f_{ij}^R FREQUENZA CONGIUNTA **RELATIVA**: è la proporzione dei casi in cui, su una popolazione di N unità statistiche, la coppia di modalità (x_i, y_j) si presenta:

$$f_{ij}^R = \frac{f_{ij}}{N}$$

$$\sum_{j=1}^{h}\sum_{i=1}^{k} f_{ij}^R = 1$$

NOTAZIONE

Frequenza congiunta assoluta f_{ij} (*notazione Cicchitelli* n_{ij})

Frequenza assoluta MARGINALE per Riga $f_{i\bullet}$ (*notazione Cicchitelli* n_{i0})

Frequenza assoluta MARGINALE per Colonna $f_{\bullet j}$ (*notazione Cicchitelli* n_{0j})

ESEMPIO

Si analizzano 1000 famiglie secondo la variabile *X = numero di auto possedute dalla famiglia* e *Y = numero di componenti della famiglia*. I risultati di tale rilevazione sono raccolti nella seguente Tabella a doppia entrata:

	Y				
X	1	2	3	4	5
0	10	20	20	150	50
1	85	85	330	50	50
2	5	85	10	0	0
3	0	10	40	0	0

X	Y					
	1	2	3	4	5	
0	10	20	20	150	50	250
1	85	85	330	50	50	600
2	5	85	10	0	0	100
3	0	10	40	0	0	50
	100	200	400	200	100	1000

Distribuzioni univariate:

X	f(x)
0	250
1	600
2	100
3	50
	1000

Y	f(y)
1	100
2	200
3	400
4	200
5	100
	1000

Tabella a doppia entrata di Frequenze congiunte **relative**:

X	Y					
	1	2	3	4	5	
0	0,01	0,02	0,02	0,15	0,05	0,25
1	0,085	0,085	0,33	0,05	0,05	0,6
2	0,005	0,085	0,01	0	0	0,1
3	0	0,01	0,04	0	0	0,05
	0,1	0,2	0,4	0,2	0,1	1

Distribuzioni univariate relative:

X	$f^R(x)$
0	0,25
1	0,60
2	0,10
3	0,05
	1,00

Y	$f^R(y)$
1	0,1
2	0,2
3	0,4
4	0,2
5	0,1
	1,0

DISTRIBUZIONI CONDIZIONATE

La Distribuzione CONDIZIONATA di X dato y_j (carattere CONDIZIONATO ***X/Y=y_j***) si ottiene fissando una modalità y_j per il carattere Y ed esaminando la distribuzione di X limitatamente alle unità statistiche che possiedono quella modalità y_j per il carattere Y:

$X/Y=y_j$	frequenze	
	assolute	relative
X_1	f_{1j}	$f_{1j}/f_{\bullet j}$
…	…	…
X_i	f_{ij}	$f_{ij}/f_{\bullet j}$
…	…	…
X_k	f_{kj}	$f_{kj}/f_{\bullet j}$
	$f_{\bullet j}$	1

(per ogni j, con j=1,…,h)

La Distribuzione CONDIZIONATA di Y dato x_i (carattere CONDIZIONATO ***Y/X=x_i***) si ottiene fissando una modalità x_i per il carattere X ed esaminando la distribuzione di Y limitatamente alle unità statistiche che possiedono quella modalità x_i per il carattere X:

$Y/X=x_i$	frequenze	
	assolute	relative
Y_1	f_{i1}	$f_{i1}/f_{i\bullet}$
…	…	…
Y_j	f_{ij}	$f_{ij}/f_{i\bullet}$
…	…	…
Y_h	f_{ih}	$f_{ih}/f_{i\bullet}$
	$f_{i\bullet}$	1

(per ogni i, con i=1,…,k)

Esempio di distribuzione condizionata (v. esempio precedente di tabella a doppia entrata):

X/Y=2	frequenze	
	assolute	relative
0	20	0,100
1	85	0,425
2	85	0,425
3	10	0,050
	200	1,000

Y/X=1	frequenze	
	assolute	relative
1	85	0,14167
2	85	0,14167
3	330	0,55000
4	50	0,08333
5	50	0,08333
	600	1

*Media aritmetica della **SOMMA** di 2 o più variabili statistiche*

Siano X_1, X_2,…,X_s delle variabili statistiche costituite ognuna da n determinazioni, di media, rispettivamente m_1, m_2,…,m_s, e sia

$$Z= X_1+X_2+\ldots+X_s$$

la loro somma.

La media della SOMMA risulta uguale alla somma delle medie dei singoli addendi:

$$M(z)=M(x_1)+M(x_2)+\ldots+M(x_s)= m_1+m_2+\ldots+m_s$$

ESERCIZIO

Sui seguenti dati:

	Y			
X	2	3	4	5
0	100	40	20	0
1	50	200	870	30
2	10	10	500	170

Determinare la variabile **Z=X+Y**. Di essa:

a) presentare la distribuzione di probabilità;
b) calcolare la media di Z e verificare le relazione che lega la media della somma alle medie degli addendi;

SOLUZIONI

a)

z	f(z)	p(z)	z*f(z)
2	100	0,050	200
3	90	0,045	270
4	230	0,115	920
5	880	0,440	4400
6	530	0,265	3180
7	170	0,085	1190
	2000	1,000	10160

b) M(z)=10160/2000=5,08

x	f(x)	x*f(x)	y	f(y)	y*f(y)
0	160	0	2	160	320
1	1150	1150	3	250	750
2	690	1380	4	1390	5560
	2000	2530	5	200	1000
				2000	7630

M(x)=2530/2000=1,265

M(y)=7630/2000=3,815

M(z)=5,08=M(x)+M(y)= 1,265+3,815=5,08

Varianza della SOMMA di due o più variabili statistiche

Siano X_1, X_2,…,X_s delle variabili statistiche costituite ognuna da n determinazioni, di media e varianza, rispettivamente m_1, m_2,…,m_s, e σ_1^2, σ_2^2, …, σ_s^2, sia

$$Z= X_1+X_2+\ldots+X_s$$

la loro somma.

La varianza della somma risulta uguale a

$$V(z) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_s^2 + 2\sum_{i=1}^{s-1}\sum_{j=i+1}^{s} Cov\left(x_i, x_j\right)$$

Dove il simbolo $\sum_{i=1}^{s-1}\sum_{j=i+1}^{s} Cov\left(x_i, x_j\right)$ indica la somma di tutte le

$$\binom{s}{2} = \frac{s!}{2!(s-2)!}$$

Covarianze ottenibili associando fra di loro ***s*** variabili prese 2 a 2.

ESERCIZIO (Variabile SOMMA)

Le famiglie di un Comune sono state classificate secondo i caratteri **X** n. di viaggi all'estero e **Y** n. di viaggi in Italia. I risultati sono i seguenti:

	Y		
X	0	2	4
0	30	20	10
1	30	100	20
3	0	10	180

a) definire la variabile somma $\boldsymbol{Z = X + Y}$ e dare la distribuzione di frequenze di $\boldsymbol{Z}$.
b) Verificare la relazione esistente fra la media della somma e le medie delle variabili;
c) Verificare la relazione esistente fra la varianza della somma e le varianze delle variabili.

SOLUZIONI
a)

z	f(z)
0	30
1	30
2	20
3	100
4	10
5	30
7	180
	400

b)

z	f(z)	z*f(z)
0	30	0
1	30	30
2	20	40
3	100	300
4	10	40
5	30	150
7	180	1260
	400	1820

M(z)=1820/400=4,55

X	f(x)	x*f(x)
0	60	0
1	150	150
3	190	570
	400	720

y	f(y)	y*f(y)
0	60	0
2	130	260
4	210	840
	400	1100

M(x)=720/400=1,8

M(y)=1100/400=2,75

M(z)= M(x)+M(y)=1,8+2,75=4,55

c)

z	f(z)	z*f(z)	z^2*f(z)
0	30	0	0
1	30	30	30
2	20	40	80
3	100	300	900
4	10	40	160
5	30	150	750
7	180	1260	8820
	400	1820	10740

Var(z)=10740/400-$(4,55)^2$=26,85-20,7025=6,1475

Var(z)= V(x)+V(y)+2*Cov(x,y)

X	f(x)	x*f(x)	x^2*f(x)
0	60	0	0
1	150	150	150
3	190	570	1710
	400	720	1860

Var(x)=1860-$(1,8)^2$=1,41

y	f(y)	y*f(y)	y^2*f(y)
0	60	0	0
2	130	260	520
4	210	840	3360
	400	1100	3880

Var(y)=3880/400-$(2,75)^2$=2,1375

Cov(x,y)=M(x*y)-M(x)*M(y)

x*y	fxy	x*y*fxy
0	90	0
2	100	200
4	20	80
6	10	60
12	180	2160
	400	2500

M(x*y)=2500/400=6,25

Cov(x,y)=6,25-(1,8)*(2,75)=1,3

Var(z)= V(x)+V(y)+2*Cov(x,y)=1,41+2,1375+2*1,3=**6,1475** (c.v.d)

*Media aritmetica del **PRODOTTO** di 2 variabili statistiche*

Sia X una variabile di media $M(x)=m_x$ e Y una variabile di media $M(y)=m_y$. Sia $Z=X\cdot Y$ la *variabile PRODOTTO* delle due variabili.

La media aritmetica di Z risulta:

$$M(Z) = M(X \cdot Y) = m_x \cdot m_y + Cov(x, y)$$

ESERCIZIO (Variabile PRODOTTO)

Date le variabili X e Y, calcolare la media del prodotto $Z=X\cdot Y$.

X	Y
1	8
2	7
3	6
4	5
5	4

Metodo diretto di calcolo:

X	Y	Z
1	8	8
2	7	14
3	6	18
4	5	20
5	4	20
		80

M(Z)=80/5=16

Metodo indiretto di calcolo:

X	Y	$x-m_x$	$y-m_y$	$(x-m_x)(y-m_y)$
1	8	-2	2	-4
2	7	-1	1	-1
3	6	0	0	0
4	5	1	-1	-1
5	4	2	-2	-4
15	30			-10

M(x)=15/5=3

M(y)=30/5=6

M(z)=3*6+(-10/5)=18-2=16