8 Indipendenza in media

8.1 Medie e varianze condizionate e marginali

Consideriamo la distribuzione dei dipendenti di un ente di ricerca per *posizione professionale* e *numero di ore di lavoro effettuate* in un mese:

A/B	b_1 160-| 180	b_2 180-| 200	b_3 200-| 220	b_4 220-| 240	TOTALE
a_1 ricercatore	6	15	14	8	**43**
a_2 1° ricercatore	2	5	4	7	**18**
a_3 dirigente di ricerca	0	3	2	3	**8**
TOTALE	**8**	**23**	**20**	**18**	**69**

Consideriamo le distribuzioni condizionate di B rispetto ad A e calcoliamo le medie aritmetiche e le varianze sulle distribuzioni condizionate:

Media aritmetica e varianza condizionata di B rispetto alla modalità a_1 di A

$_cb_j$	n_{1j}	$_cb_jn_{1j}$	$_cb_j^2$	$_cb_j^2n_{1j}$
170	6	1020	28900	173400
190	15	2850	36100	541500
210	14	2940	44100	617400
230	8	1840	52900	423200
TOTALE	$n_{1.}$=43	8650		1755500

$$M_{B|A=a_1} = \frac{\sum_{j=1}^{4} {}_c b_j n_{1j}}{n_{1.}} = \frac{8650}{43} = 201{,}16$$

$$\sigma^2_{B|A=a_1} = \frac{\sum_{j=1}^{4} \left[{}_c b_j - M_{B|A=a_1} \right]^2 n_{1j}}{n_{1.}} = \sum_{j=1}^{4} \frac{{}_c b_j^2 n_{1j}}{n_{1.}} - M^2_{B|A=a_1} = \frac{1755500}{43} - (201{,}16)^2 = 359{,}11$$

Media aritmetica e varianza condizionata di B rispetto alla modalità a_2 di A

$_cb_j$	n_{2j}	$_cb_jn_{2j}$	$_cb_j^2$	$_cb_j^2n_{2j}$
170	2	340	28900	57800
190	5	950	36100	180500
210	4	840	44100	176400
230	7	1610	52900	370300
TOTALE	$n_{2.}$=18	3740		785000

$$M_{B|A=a_2} = \frac{\sum_{j=1}^{4} {}_c b_j n_{2j}}{n_{2.}} = \frac{3740}{18} = 207{,}78$$

$$\sigma^2_{B|A=a_2} = \frac{\sum_{j=1}^{4} \left[{}_c b_j - M_{B|A=a_2} \right]^2 n_{2j}}{n_{2.}} = \sum_{j=1}^{4} \frac{{}_c b_j^2 n_{2j}}{n_{2.}} - M^2_{B|A=a_2} = \frac{785000}{18} - (207{,}78)^2 = 439{,}51$$

Media aritmetica e varianza condizionata di B rispetto alla modalità a_3 di A

$_cb_j$	n_{3j}	$_cb_jn_{3j}$	$_cb_j^2$	$_cb_j^2n_{3j}$
170	0	0	28900	0
190	3	570	36100	108300
210	2	420	44100	88200
230	3	690	52900	158700
TOTALE	$n_{3.}$=8	1680		355200

$$M_{B|A=a_3} = \frac{\sum_{j=1}^{4} {}_c b_j n_{3j}}{n_{3.}} = \frac{1680}{8} = 210$$

$$\sigma^2_{B|A=a_3} = \frac{\sum_{j=1}^{4} \left[{}_c b_j - M_{B|A=a_3} \right]^2 n_{3j}}{n_{3.}} = \sum_{j=1}^{4} \frac{{}_c b_j^2 n_{3j}}{n_{3.}} - M^2_{B|A=a_3} = \frac{355200}{8} - (210)^2 = 300$$

Calcoliamo, adesso, media aritmetica e varianza sulla distribuzione marginale di B:

Media aritmetica e varianza della distribuzione marginale di B

$_cb_j$	$n_{.j}$	$_cb_jn_{.j}$	$_cb_j^2$	$_cb_j^2n_{.j}$
170	8	1360	28900	231200
190	23	4370	36100	830300
210	20	4200	44100	882000
230	18	4140	52900	952200
TOTALE	N=69	14070		2895700

$$M_B = \frac{\sum_{j=1}^{4} {}_c b_j n_{.j}}{N} = \frac{14070}{69} = 203{,}91$$

$$\sigma_B^2 = \frac{\sum_{j=1}^{4} \left[{}_c b_j - M_B \right]^2 n_{.j}}{N} = \sum_{j=1}^{4} \frac{{}_c b_j^2 n_{.j}}{N} - M_B^2 = \frac{2895700}{69} - (203{,}91)^2 = 386{,}14$$

8.2 Rapporto di correlazione

La variabile B è indipendente in media dalla variabile A se ciascuna media condizionata è uguale alla media calcolata sulla distribuzione marginale di B, quindi se tutte le medie condizionate sono uguali fra loro. L'indipendenza in media non è, quindi, simmetrica come l'indipendenza in distribuzione; è ovvio che, nell'esempio suddetto, non è possibile calcolare la dipendenza in media di A da B, essendo A una variabile qualitativa.

L'eventuale dipendenza in media di B da A si può misurare attraverso il "rapporto di correlazione", dato dal rapporto tra la varianza delle medie condizionate e la varianza di B:

$$\eta^2_{B|A} = \frac{\sigma^2_{M_{B|A}}}{\sigma^2_B}$$

Calcoliamo la varianza delle medie condizionate e la media delle varianze condizionate:

$n_{i.}$	$M_{B\|A=a_i}$	$M_{B\|A=a_i} n_{i.}$	$M^2_{B\|A=a_i}$	$M^2_{B\|A=a_i} n_{i.}$
43	201,16	8.650,00	40.466,47	1.740.058,14
18	207,78	3.740,00	43.171,60	777.088,89
8	210,00	1.680,00	44.100,00	352.800,00
69		**14.070,00**		**2.869.947,03**

Media delle medie condizionate

$$\frac{1}{N}\sum_{i=1}^{3} M_{B|A=a_i} n_{i.} = \frac{14070}{69} = 203{,}91 = M_B$$

Varianza delle medie condizionate

$$\sigma^2_{M_{B|A}} = \frac{\sum_{i=1}^{3}\left[M_{B|A=a_i} - M_B\right]^2 n_{i.}}{N} = \frac{\sum_{i=1}^{3} M^2_{B|A=a_i} n_{i.}}{N} - M^2_B = \frac{2869947{,}03}{69} - (203{,}91)^2 = 12{,}91$$

$n_{i.}$	$\sigma^2_{B\|A=a_i}$	$\sigma^2_{B\|A=a_i} n_{i.}$
43	359,11	15.441,86
18	439,51	7.911,11
8	300,00	2.400,00
69		**25.752,97**

Media delle varianze condizionate

$$M_{\sigma^2_{B|A}} = \frac{\sum_{i=1}^{3} \sigma^2_{B|A=a_i} n_{i.}}{N} = \frac{25752{,}97}{69} = 373{,}23$$

Si dimostra che sommando la varianza delle medie condizionate e la media delle varianze condizionate, si ottiene la varianza di B. Infatti è:

$$\sigma^2_{M_{B|A}} + M_{\sigma^2_{B|A}} = 12{,}91+373{,}23 = 386{,}14 = \sigma^2_B$$

Dunque, il rapporto di correlazione varia tra 0 ed 1, assumendo valore 0 in caso di perfetta indipendenza in media e valore 1 in caso di perfetta dipendenza in media. Nell'esempio, il rapporto di correlazione è molto più vicino a zero che ad 1, per cui B si può ritenere indipendente in media da A.

$$\eta^2_{B|A} = \frac{\sigma^2_{M_{B|A}}}{\sigma^2_B} = \frac{12{,}91}{386{,}14} = 0{,}03\,.$$

Naturalmente, l'indipendenza in distribuzione implica l'indipendenza in media, ma non viceversa. Infatti, vi è indipendenza in distribuzione se tutte le distribuzioni condizionate relative sono uguali fra loro; a maggior ragione, dunque, saranno uguali le medie calcolate su di esse. Ciò si può anche dimostrare analiticamente. Consideriamo la generica media condizionata di B dato A, quando A assume la modalità a_i:

$$M_{B|A=a_i} = \frac{\sum_{j=1}^{c} b_j n_{ij}}{n_{i.}}$$

Se vi è indipendenza in distribuzione, si verifica che

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$$

dunque

$$M_{B|A=a_i} = \frac{\sum_{j=1}^{c} b_j n_{.j}}{N}$$

ma quest'ultima altro non è che la media di B. Allora, se tale uguaglianza vale per ogni *i*, ciò vuol dire che tutte le medie condizionate saranno uguali alla media di B e quindi saranno uguali fra loro.

8.3 Punto medio e punto mediano

Qualora fosse possibile calcolare la media aritmetica su entrambe le variabili A e B, tali medie costituirebbero le coordinate del PUNTO MEDIO (M_A, M_B) della distribuzione di frequenza doppia, mentre le mediane calcolate sulle distribuzioni marginali di A e di B costituirebbero le coordinate del PUNTO MEDIANO (M_{eA}, M_{eB}).

Nell'esempio considerato, non è possibile calcolare il punto medio, poiché le variabili in esame non sono entrambe quantitative. E' possibile però calcolare la mediana anche sulla distribuzione marginale di A, essendo questa una variabile qualitativa ordinabile:

A	$\mathbf{n_{i.}}$	$\mathbf{N_{i.}}$
Ricercatore	43	43
1° ricercatore	18	61
Dirigente	8	69
	69	

Poiché *N=69* è dispari, la mediana sarà quella modalità che occuperà la posizione *(N+1)/2=35*, ossia "ricercatore". In effetti, in tal caso, non sarebbe stato necessario neanche calcolare le frequenze cumulate N_i, essendo il valore "*35*" già compreso nella prima frequenza assoluta.

8.4 Frequenze cumulate per una tabella doppia

E' possibile definire le frequenze cumulate anche su una tabella doppia di frequenza. Le frequenze assolute cumulate rappresentano il numero di unità statistiche che hanno modalità di $A \leq i$ e modalità di $B \leq j$:

$$N_{ij} = \sum_{h=1}^{i} \sum_{k=1}^{j} n_{hk}$$

Supponiamo, ad esempio, di aver osservato la seguente tabella di frequenze congiunte, dove A e B sono almeno ordinabili:

A\B	b_1	b_2	b_3	*totale*
a_1	3	5	7	15
a_2	13	20	4	37
a_3	4	2	6	12
totale	20	27	17	64

La tabella delle frequenze assolute cumulate è:

A\B	b_1	b_2	b_3
a_1	3	8	15
a_2	16	41	52
a_3	20	47	64

dove, ad esempio, è:

$N_{13}=n_{11}+n_{12}+n_{13}=3+5+7=15$

$N_{32}=n_{11}+n_{12}+n_{21}+n_{22}+n_{31}+n_{32}=47$.

Per determinare la tabella delle frequenze relative cumulate, basta dividere ciascuna frequenza assoluta cumulata per il totale delle osservazioni:

$$F_{ij} = \sum_{h=1}^{i} \sum_{k=1}^{j} f_{hk} = \frac{N_{ij}}{N}.$$