7 L'interdipendenza fra due variabili

7.1 Tabelle doppie di frequenza

Finora abbiamo supposto di aver rilevato una sola variabile su un collettivo di n unità statistiche. Supponiamo, adesso, di aver rilevato su *N* soggetti/oggetti due variabili A e B; disponiamo, dunque, non più di una singola serie di osservazioni, ma di una serie doppia.

Il primo processo di sintesi per una variabile doppia consiste nella costruzione di una distribuzione di frequenza congiunta; tale distribuzione prende il nome di "tabella a doppia entrata". Una tabella a doppia entrata si presenta nella seguente forma:

A/B	$\mathbf{b_1}$	$\mathbf{b_2}$	**…**	$\mathbf{b_j}$	**…**	$\mathbf{b_c}$	**totale**
$\mathbf{a_1}$	n_{11}	n_{12}	…	n_{1j}	…	n_{1c}	$\boldsymbol{n_{1.}}$
$\mathbf{a_2}$	n_{21}	n_{22}	…	n_{2j}	…	n_{2c}	$\boldsymbol{n_{2.}}$
…	…	…	…	…	…	…	**…**
$\mathbf{a_i}$	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ic}	$\boldsymbol{n_{i.}}$
…	…	…	…	…	…	…	**…**
$\mathbf{a_r}$	n_{r1}	n_{r2}	...	n_{rj}	…	n_{rc}	$\boldsymbol{n_{r.}}$
totale	$\boldsymbol{n_{.1}}$	$\boldsymbol{n_{.2}}$	**…**	$\boldsymbol{n_{.j}}$	**…**	$\boldsymbol{n_{.c}}$	$\boldsymbol{N}$

dove

- a_i rappresenta la generica modalità di A , con $i = 1, 2,\ldots, r$;
- b_j rappresenta la generica modalità di B , con $j = 1, 2,\ldots, c$;
- le n_{ij} sono le cosiddette "*frequenze congiunte*", che stanno ad indicare quante volte si presentano congiuntamente le modalità a_i e b_j.

Inoltre:

$n_{i.}$ sono i totali di riga: $n_{i.} = \sum_j n_{ij}$;

$n_{.j}$ sono i totali di colonna: $n_{.j} = \sum_i n_{ij}$;

N è il totale generale, cioè il totale delle osservazioni: $N = \sum_j \sum_i n_{ij} = \sum_i n_{i.} = \sum_j n_{.j}$.

Ciascuna riga della tabella rappresenta la distribuzione di B condizionata alla modalità a_i di A, mentre ciascuna colonna rappresenta la distribuzione di A condizionata alla modalità b_j di B.

In particolare, l'ultima riga e l'ultima colonna rappresentano, rispettivamente, la distribuzione marginale di B e la distribuzione marginale di A.

Da una tabella a doppia entrata, dunque possiamo ricavare $r+c+2$ distribuzioni di frequenza semplici.

Se A e B sono due variabili qualitative, la tabella a doppia entrata prende il nome di "*tavola di contingenza*", se invece entrambe le variabili sono quantitative la tabella a doppia entrata prende il nome di "*tavola di correlazione*". Se, le variabili sono una qualitativa e l'altra quantitativa, la tavola viene definita "*mista*".

Su una tabella doppia di frequenza possono essere calcolate:

- le frequenze relative rispetto al totale;
- le frequenze relative rispetto ai totali di riga;
- le frequenze relative rispetto ai totali di colonna.

Le tabelle che seguono mostrano i tre casi suddetti:

Frequenze relative rispetto al totale

A/B	$\mathbf{b_1}$	$\mathbf{b_2}$	**…**	$\mathbf{b_j}$	**…**	$\mathbf{b_c}$	**totale**
$\mathbf{a_1}$	n_{11}/N	n_{12}/N	…	n_{1j}/N	…	n_{1c}/N	$\boldsymbol{n_{1.}/N}$
$\mathbf{a_2}$	n_{21}/N	n_{22}/N	…	n_{2j}/N	…	n_{2c}/N	$\boldsymbol{n_{2.}/N}$
…	…	…	…	…	…	…	**…**
$\mathbf{a_i}$	n_{i1}/N	n_{i2}/N	…	n_{ij}/N	…	n_{ic}/N	$\boldsymbol{n_{i.}/N}$
…	…	…	…	…	…	…	**…**
$\mathbf{a_r}$	n_{r1}/N	n_{r2}/N	...	n_{rj}/N	…	n_{rc}/N	$\boldsymbol{n_{r.}/N}$
totale	$\boldsymbol{n_{.1}/N}$	$\boldsymbol{n_{.2}/N}$	**…**	$\boldsymbol{n_{.j}/N}$	**…**	$\boldsymbol{n_{.c}/N}$	$\boldsymbol{N/N=1}$

Frequenze relative rispetto ai totali di riga

(ciascuna riga rappresenta la distribuzione relativa condizionata di B rispetto alla modalità a_i di A)

A/B	$\mathbf{b_1}$	$\mathbf{b_2}$	**…**	$\mathbf{b_j}$	**…**	$\mathbf{b_c}$	**totale**
$\mathbf{a_1}$	$n_{11}/n_{1.}$	$n_{12}/n_{1.}$	…	$n_{1j}/n_{1.}$	…	$n_{1c}/n_{1.}$	$\boldsymbol{n_{1.}/n_{1.}=1}$
$\mathbf{a_2}$	$n_{21}/n_{2.}$	$n_{22}/n_{2.}$	…	$n_{2j}/n_{2.}$	…	$n_{2c}/n_{2.}$	$\boldsymbol{n_{2.}/n_{2.}=1}$
…	…	…	…	…	…	…	**…**
$\mathbf{a_i}$	$n_{i1}/n_{i.}$	$n_{i2}/n_{i.}$	…	$n_{ij}/n_{i.}$	…	$n_{ic}/n_{i.}$	$\boldsymbol{n_{i.}/n_{i.}=1}$
…	…	…	…	…	…	…	**…**
$\mathbf{a_r}$	$n_{r1}/n_{r.}$	$n_{r2}/n_{r.}$	...	$n_{rj}/n_{r.}$	…	$n_{rc}/n_{r.}$	$\boldsymbol{n_{r.}/n_{r.}=1}$
totale	$\boldsymbol{n_{.1}/N}$	$\boldsymbol{n_{.2}/N}$	**…**	$\boldsymbol{n_{.j}/N}$	**…**	$\boldsymbol{n_{.c}/N}$	$\boldsymbol{N/N=1}$

Frequenze relative rispetto ai totali di colonna

(ciascuna colonna rappresenta la distribuzione relativa condizionata di A rispetto alla modalità b_j di B)

A/B	$\mathbf{b_1}$	$\mathbf{b_2}$	**…**	$\mathbf{b_j}$	**…**	$\mathbf{b_c}$	**totale**
$\mathbf{a_1}$	$n_{11}/n_{.1}$	$n_{12}/n_{.2}$	…	$n_{1j}/n_{.j}$	…	$n_{1c}/n_{.c}$	$\boldsymbol{n_{1.}/N}$
$\mathbf{a_2}$	$n_{21}/n_{.1}$	$n_{22}/n_{.2}$	…	$n_{2j}/n_{.j}$	…	$n_{2c}/n_{.c}$	$\boldsymbol{n_{2.}/N}$
…	…	…	…	…	…	…	**…**
$\mathbf{a_i}$	$n_{i1}/n_{.1}$	$n_{i2}/n_{.2}$	…	$n_{ij}/n_{.j}$	…	$n_{ic}/n_{.c}$	$\boldsymbol{n_{i.}/N}$
…	…	…	…	…	…	…	**…**
$\mathbf{a_r}$	$n_{r1}/n_{.1}$	$n_{r2}/n_{.2}$	...	$n_{rj}/n_{.j}$	…	$n_{rc}/n_{.c}$	$\boldsymbol{n_{r.}/N}$
totale	$\boldsymbol{n_{.1}/n_{.1}=1}$	$\boldsymbol{n_{.2}/n_{.2}=1}$	**…**	$\boldsymbol{n_{.j}/n_{.j}=1}$	**…**	$\boldsymbol{n_{.c}/n_{.c}=1}$	$\boldsymbol{N/N=1}$

La seguente tabella riporta la distribuzione di un collettivo di 219 studenti secondo il sesso e l'attitudine per determinate discipline:

SESSO/ ATTITUDINE	**Discipline artistiche**	**Discipline umanistiche**	**Discipline scientifiche**	**TOTALE**
M	35	40	44	119
F	22	27	51	100
TOTALE	57	67	95	219

Determiniamo le tre tabelle che contengono rispettivamente:

- le frequenze relative rispetto al totale;

- le frequenze relative rispetto ai totali di riga;
- le frequenze relative rispetto ai totali di colonna

SESSO/ ATTITUDINE	Discipline artistiche	Discipline umanistiche	Discipline scientifiche	TOTALE
M	0,16	0,18	0,20	0,54
F	0,10	0,12	0,23	0,46
TOTALE	0,26	0,31	0,43	1,00

SESSO/ ATTITUDINE	Discipline artistiche	Discipline umanistiche	Discipline scientifiche	TOTALE
M	0,29	0,34	0,37	1,00
F	0,22	0,27	0,51	1,00
TOTALE	0,26	0,31	0,43	1,00

SESSO/ ATTITUDINE	Discipline artistiche	Discipline umanistiche	Discipline scientifiche	TOTALE
M	0,61	0,60	0,46	0,54
F	0,39	0,40	0,54	0,46
TOTALE	1,00	1,00	1,00	1,00

7.2 Indipendenza in distribuzione

Spesso è interessante sapere se tra i due caratteri A e B esiste una relazione di dipendenza.

Supponiamo di aver osservato la seguente tabella:

A/B	b_1	b_2	b_3	TOTALE
a_1	1	5	4	10
a_2	4	20	16	40
a_3	5	25	20	50
TOTALE	10	50	40	100

Calcoliamo le frequenze relative rispetto ai totali di riga:

A/B	b_1	b_2	b_3	TOTALE
a_1	0,1	0,5	0,4	1
a_2	0,1	0,5	0,4	1
a_3	0,1	0,5	0,4	1
TOTALE	0,1	0,5	0,4	1

Notiamo che le righe sono tutte uguali; ovvero, le distribuzioni relative condizionate di B rispetto ad A sono uguali fra loro. Ciò vuol dire che B è indipendente da A, poiché la sua distribuzione non varia al variare delle modalità di A.

Calcoliamo adesso le frequenze relative rispetto ai totali di colonna:

A/B	**b_1**	**b_2**	**b_3**	**TOTALE**
a_1	0,1	0,1	0,1	0,1
a_2	0,4	0,4	0,4	0,4
a_3	0,5	0,5	0,5	0,5
TOTALE	1	1	1	1

Notiamo, in quest'altro caso, che le colonne sono tutte uguali; ovvero le distribuzioni relative condizionate di A rispetto a B sono uguali fra loro. Ciò vuol dire che A è indipendente da B, poiché la sua distribuzione non varia al variare delle modalità di B.

Concludiamo, dunque, che se B è indipendente da A, è anche A indipendente da B e viceversa.

Formalizziamo quanto detto:

se $\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N} \qquad \forall\,(i,j)$

è anche vero che

$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N} \qquad \forall\,(i,j)$

Allora, il generico valore della frequenza congiunta, nell'ipotesi di indipendenza, può essere indicato con:

$\hat{n}_{ij} = \frac{n_{i.} n_{.j}}{N} \qquad \forall\,(i,j)$

e prende il nome di *frequenza teorica di indipendenza*.

Le differenze fra le frequenze osservate e le frequenze teoriche di indipendenza sono definite "*contingenze*":

$c_{ij} = n_{ij} - \hat{n}_{ij}$

Ovviamente, in caso di indipendenza le contingenze saranno tutte nulle.

E' facile dimostrare che $\Sigma_i c_{ij} = \Sigma_j c_{ij} = \Sigma_{i,j} c_{ij} = 0$. Dimostriamo che $\Sigma_i c_{ij} = 0$:

$$\Sigma_i c_{ij} = \Sigma_i (n_{ij} - \hat{n}_{ij}) = \Sigma_i n_{ij} - \Sigma_i \hat{n}_{ij} = n_{.j} - \Sigma_i \frac{n_{i.} n_{.j}}{N} = n_{.j} - \frac{n_{.j}}{N} \Sigma_i n_{i.} = n_{.j} - \frac{n_{.j}}{N} N_{.} = 0.$$

Analogamente, si dimostra che $\Sigma_j c_{ij} = 0$ e che $\Sigma_{i,j} c_{ij} = 0$.

La maggior parte degli indici proposti in letteratura per lo studio dell'associazione si basano proprio sulle contingenze. In particolare, l'indice proposto da Pearson è dato dalla seguente espressione:

$$X^2 = \sum_{i,j} \frac{c_{ij}^2}{\hat{n}_{ij}} = \sum_{i,j} \frac{n_{ij}^2}{\hat{n}_{ij}} - N = N\left(\sum_{i,j} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right).$$

Tale indice assume valore zero in caso di indipendenza, ma cresce indefinitamente all'aumentare delle osservazioni.

Successivamente, per eliminare l'influenza di *N*, Pearson propose il seguente indice:

$$\Phi^2 = \frac{X^2}{N}.$$

Nel tentativo di normalizzare l'indice, nel tentativo cioè di limitare i suoi valori nel range [0,1], ancora Pearson propose il cosiddetto "coefficiente di contingenza":

$$\mathrm{P} = \left(\frac{X^2}{X^2 + N} \right)^{1/2} = \left(\frac{\Phi^2}{\Phi^2 + 1} \right)^{1/2},$$

ma P non raggiunge mai il valore 1, neanche in caso di perfetta dipendenza fra i due caratteri.

Un indice che assume valori nell'intervallo [0, 1] è stato proposto da Cramer:

$$C = \left(\frac{\Phi^2}{\min[(r-1),(c-1)]} \right)^{1/2} .$$

Tale indice assume valore zero in caso di indipendenza e valore 1 in caso di dipendenza perfetta.

Calcoliamo gli indici X^2, Φ^2 e C sulla distribuzione del collettivo di 219 studenti secondo il sesso e l'attitudine:

SESSO/ ATTITUDINE	Discipline artistiche	Discipline umanistiche	Discipline scientifiche	TOTALE
M	35	40	44	119
F	22	27	51	100
TOTALE	57	67	95	219

Frequenze teoriche $\hat{n}_{ij}$

SESSO/ ATTITUDINE	Discipline artistiche	Discipline umanistiche	Discipline scientifiche	TOTALE
M	30,97	36,41	51,62	119,00
F	26,03	30,59	43,38	100,00
TOTALE	57,00	67,00	95,00	219,00

Contingenze $c_{ij} = n_{ij} - \hat{n}_{ij}$

SESSO/ ATTITUDINE	Discipline artistiche	Discipline umanistiche	Discipline scientifiche	TOTALE
M	4,03	3,59	-7,62	0,00
F	-4,03	-3,59	7,62	0,00
TOTALE	0,00	0,00	0,00	0,00

Contingenze al quadrato c_{ij}^2

SESSO/ ATTITUDINE	Discipline artistiche	Discipline umanistiche	Discipline scientifiche
M	16,22	12,91	58,08
F	16,22	12,91	58,08

Contingenze al quadrato / Frequenze teoriche $c_{ij}^2 / \hat{n}_{ij}$

SESSO/ ATTITUDINE	Discipline artistiche	Discipline umanistiche	Discipline scientifiche	TOTALE
M	0,52	0,35	1,13	2,00
F	0,62	0,42	1,34	2,38
TOTALE	1,15	0,78	2,46	4,39

X^2=4,39 Φ^2=0,02 C=0,14.

Dal valore di quest'ultimo indice, molto più vicino a 0 che ad 1, si evince che i due caratteri non sono associati, ovvero non sembra che l'attitudine verso determinate discipline possa dipendere dal sesso.

7.3 Dipendenza perfetta

La situazione di dipendenza non è univocamente caratterizzata; può essere unilaterale, se $r \neq c$, o bilaterale, se $r=c$. I seguenti tre esempi mostrano, rispettivamente, come:

a) il carattere B dipende perfettamente da A, ma il carattere A non dipende da B ($r > c$): ad ogni modalità di A corrisponde sempre una sola modalità di B, ma non è vero il contrario (in ogni riga c'è solo una frequenza congiunta non nulla);

b) il carattere A dipende perfettamente da B ($r < c$). Infatti, ad ogni modalità di B corrisponde sempre una sola modalità di A, ma non è vero il contrario (in ogni colonna c'è solo una frequenza congiunta non nulla);

c) i due caratteri A e B sono perfettamente associati ($r = c$): in ogni riga e in ogni colonna c'è solo una frequenza congiunta non nulla.

Il carattere B dipende perfettamente da A

A/B	b_1	b_2	b_3	TOTALE
a_1	10	0	0	10
a_2	0	0	30	30
a_3	0	0	15	15
a_4	0	5	0	5
TOTALE	10	5	45	60

Il carattere A dipende perfettamente da B

A/B	b_1	b_2	b_3	b_4	TOTALE
a_1	10	0	0	0	10
a_2	0	0	0	5	5
a_3	0	30	15	0	45
TOTALE	10	30	15	5	60

I due caratteri sono perfettamente associati

A/B	b_1	b_2	b_3	TOTALE
a_1	0	5	0	5
a_2	0	0	10	10
a_3	30	0	0	30
TOTALE	30	5	10	45

In quest'ultimo caso le frequenze potrebbero disporsi sulla diagonale principale, indicando una "*perfetta associazione positiva*" o sulla diagonale secondaria, indicando una "*perfetta associazione negativa*" o "*perfetta dissociazione*".

Gli indici X^2 e C assumono nelle tre situazioni suddette a), b) e c) il loro massimo valore ma, poiché possono assumere solo valori positivi, non distinguono l'associazione dalla dissociazione.

7.4 Indici di associazione per tabelle 2×2

Si consideri una tabella dicotomica, ossia una tabella in cui entrambe le variabili possono assumere solo due modalità:

A/B	b_1	b_2	totale
a_1	n_{11}	n_{12}	$n_{1.}$
a_2	n_{21}	n_{22}	$n_{2.}$
totale	$n_{.1}$	$n_{.2}$	N

La particolarità di una tabella 2×2 sta nel fatto che, fissati i totali marginali, la conoscenza di una sola frequenza congiunta n_{ij} è sufficiente per determinare le altre tre. Un'analisi sull'associazione può essere condotta dunque basandosi su una sola frequenza. In particolare, confrontando la frequenza osservata n_{11} con la corrispondente frequenza teorica $\hat{n}_{11} = \frac{n_{1.} n_{.1}}{n}$, si può affermare che:

1. se $n_{11} = \hat{n}_{11}$, A e B sono indipendenti;
2. se $n_{11} > \hat{n}_{11}$, tra A e B c'è associazione positiva;
3. se $n_{11} < \hat{n}_{11}$, tra A e B c'è associazione negativa.

Sono stati proposti diversi coefficienti per misurare l'associazione fra variabili dicotomiche; il più importante è l'indice *V* di Pearson:

$$V = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11} + n_{12})(n_{11} + n_{21})(n_{12} + n_{22})(n_{21} + n_{22})}}.$$

L'indice *V* varia nel range [-1,+1]; in particolare, assume valore:

1. 0 in caso di indipendenza;
2. 1 in caso di perfetta associazione ($n_{12} = n_{21} = 0$);
3. -1 in caso di perfetta dissociazione ($n_{11} = n_{22} = 0$).

Supponiamo di aver osservato la seguente tabella:

SESSO/ ESAME DI MATEMATICA	N	S	Totale
F	10	2	12
M	16	2	18
Totale	26	4	30

Calcoliamo l'indice *V*:

$$V = \frac{10 \cdot 2 - 2 \cdot 16}{\sqrt{12 \cdot 18 \cdot 26 \cdot 4}} = -0{,}08$$

Il valore di V, molto più vicino a 0 che a -1, indica che non c'è alcuna relazione fra il sesso degli studenti intervistati e il fatto che abbiano sostenuto o meno l'esame di matematica.

7.5 Indici di cograduazione

Se i dati riportati in una tabella di contingenza sono relativi a variabili misurate su scala nominale, gli indici che quantificano la dipendenza tra le due variabili vengono definiti, come si è visto, *misure o indici di associazione*. Se le due variabili sono misurabili su scala ordinale, gli indici preposti prendono il nome di *indici di cograduazione*. Tali indici consentono non solo di misurare l'intensità di un'eventuale associazione, ma anche di individuarne il verso, ovvero consentono di stabilire se tra i due caratteri ordinati c'è concordanza (associazione positiva) o discordanza (associazione negativa). Si ha concordanza quando a modalità basse della prima variabile corrispondono modalità basse della seconda variabile e a modalità alte corrispondono modalità alte. Si ha discordanza quando a modalità basse corrispondono modalità alte e viceversa a modalità alte corrispondono modalità basse.

7.5.1 Concordanza tra graduatorie

Prima di esaminare gli indici che misurano l'intensità della relazione esistente fra due variabili ordinabili espresse sotto forma di tabella a doppia entrata, analizziamo due indici utilizzati per misurare la "concordanza" tra due semplici graduatorie, relative allo stesso insieme di unità statistiche.

La forma più comune di graduatoria è quella che si fonda sull'ipotesi che le modalità siano tutte differenti ed equidistanti, quindi rappresentabili con i numeri naturali da 1 ad n.

Consideriamo il seguente esempio. Supponiamo di aver rilevato i due caratteri "Attività sportiva" e "Autocontrollo" su un insieme di 10 soggetti e supponiamo che tali caratteri siano stati misurati secondo scale di livello ordinale:

Individuo	**Attività sportiva**	**Autocontrollo**
Francesco	20	16
Paolo	17	19
Giovanna	16	15
Stefano	11	18
Carlo	8	6
Piero	8	10
Marco	6	7
Cecilia	5	4
Franco	5	3
Maria	1	2

Si vuol verificare se fra le due variabili esiste una relazione.

Per misurare la concordanza tra le due graduatorie utilizziamo il coefficiente "Rho" proposto da Spearman:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

dove $d_i = j - k$, in cui j e k sono i ranghi delle due graduatorie poste a confronto, ed n è il numero delle osservazioni. Il "rango" indica la posizione che il "grado" o "punteggio" occupa nella serie ordinata in senso crescente o decrescente. In corrispondenza di punteggi uguali (*tied*), si attribuisce un rango dato dalla media dei ranghi:

Individuo	**Attività sportiva**	**Rango**	**Autocontrollo**	**Rango**	d_i	d_i^2
Francesco	20	1	16	3	-2	4
Paolo	17	2	19	1	1	1
Giovanna	16	3	15	4	-1	1
Stefano	11	4	18	2	2	4
Carlo	8	5,5	6	7	-1,5	2,25
Piero	8	5,5	10	5	0,5	0,25
Marco	6	7	7	6	1	1
Cecilia	5	8,5	4	8	0,5	0,25
Franco	5	8,5	3	9	-0,5	0,25
Maria	1	10	2	10	0	0
					totale	14

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} = 1 - \frac{6 \cdot 14}{10\left(100 - 1\right)} = 1 - \frac{84}{990} = 0{,}9\,.$$

Il coefficiente ρ varia nel range [-1,+1] e assume, in particolare:

- valore +1 in caso di perfetta concordanza ($j=k$);
- valore –1 in caso di massima discordanza;
- valore 0 in caso di indipendenza.

Nell'esempio suddetto il valore $\rho=0,9$ esprime una concordanza quasi perfetta fra le due variabili, per cui si può concludere che l'attività sportiva facilita l'autocontrollo.

Il coefficiente ρ è stato ricavato da Spearman come coefficiente di correlazione (cfr.par. 7.6) tra ranghi, come si può facilmente dimostrare.

Un altro indice utilizzato per misurare il grado di corrispondenza fra due graduatorie è il "tau" di Kendall:

$$\tau = \frac{2s}{n\left(n-1\right)}\,.$$

Per calcolare il τ si dispone la prima graduatoria in ordine naturale e si considera la nuova disposizione della seconda graduatoria.

Si supponga, ad esempio, di aver rilevato l'ordine di arrivo di 5 atleti in due diverse gare:

Individuo	**Prima gara**	**Seconda gara**
Francesco	3	5
Paolo	1	2
Giovanna	5	3
Stefano	2	1
Carlo	4	4

Ordiniamo la prima graduatoria; la nuova disposizione è:

Individuo	**Prima gara**	**Seconda gara**
Paolo	1	2
Stefano	2	1
Francesco	3	5
Carlo	4	4
Giovanna	5	3

Si consideri la seconda graduatoria e si confronti ciascun punteggio con i successivi; si assegni valore -1 ogniqualvolta tale punteggio risulta superiore al punteggio con cui è confrontato, viceversa si assegni valore +1. La somma di tali valori darà *s*:

Individuo		**Totale**
Paolo	-1 +1 +1 +1	+2
Stefano	+1 +1 +1	+3
Francesco	-1 -1	-2
Carlo	-1	-1
	s	+2

$$\tau = \frac{2s}{n(n-1)} = \frac{2 \cdot 2}{5 \cdot 4} = 0{,}2.$$

Come ρ, anche il coefficiente τ può assumere valori compresi tra –1 (massima discordanza) e +1 (massima concordanza) ed è una misura simmetrica rispetto allo 0. Il risultato ottenuto, dunque, non sembra confermare una concordanza fra i punteggi riportati nelle due gare.

I due coefficienti ρ e τ risultano uguali solo nel caso in cui le graduatorie considerate sono perfettamente concordanti o discordanti, viceversa ρ tende ad assumere valori più alti di τ, poiché tende ad amplificare gli scarti.

7.5.2 Cograduazione per tabelle doppie di frequenza

Si consideri adesso una tabella di contingenza, in cui la variabile di riga A e la variabile di colonna B sono misurate su scala ordinale, entrambe in senso crescente o decrescente. Definiamo la "concordanza" e la "discordanza" in modo più dettagliato.

Due osservazioni che, all'interno della tabella, occupano le posizioni (i, j) e (i', j') sono:

- **concordanti** se (i<i') e (j<j') o se (i>i') e (j>j');

- **discordanti** se (i<i') e (j>j') o se (i>i') e (j<j');

- **tied** se hanno la stessa classificazione rispetto alla variabile A e/o B.

Consideriamo, ad esempio, la seguente tabella di contingenza, in cui:

A: condizione meteorologica;

B: livello di traffico automobilistico

A/B	basso	medio	alto
pioggia	7	26	55
variabile	29	98	29
sereno	84	26	11

Le osservazioni nelle celle di posizione (1,1) e (2,2) sono concordanti. In generale, le osservazioni nella cella (1,1) sono concordanti con tutte le osservazioni che si trovano a *sud-est* della tabella, che hanno livelli maggiori per entrambe le variabili. Tale regola può essere estesa a ciascuna osservazione in ciascuna cella, per cui il numero delle coppie concordanti sarà N_c=4339:

dalla cella		**numero di coppie**	**Totale**
pioggia	basso	7(98+26+29+11)	1148
pioggia	medio	26(29+11)	1040
variabile	basso	29(26+11)	1073
variabile	medio	98·11	1078
		N_c	4339

Le osservazioni nelle celle di posizione (1,2) e (2,1) sono discordanti. In generale, ciascuna osservazione sarà discordante con le osservazioni che si trovano a *sud-ovest* nella tabella, per cui il numero delle coppie discordanti sarà N_d=27395:

dalla cella		**numero di coppie**	**totale**
pioggia	medio	26(29+84)	2938
pioggia	alto	55(29+98+84+26)	13035
variabile	medio	98·84	8232
variabile	alto	29(84+26)	3190
		N_d	27395

Il numero di coppie tied rispetto alla variabile A è T_a =11916:

dalla riga	**numero coppie**	**totale**
pioggia	7(26+55)+26·55	1997
variabile	29(98+29)+98·29	6525
sereno	84(26+11)+26·11	3394
	T_a	11916

Il numero di coppie tied rispetto alla variabile B è T_b =11518:

dalla colonna	**numero coppie**	**totale**
basso	7(29+84)+29·84	3227
medio	26(98+26)+98·26	5772
alto	55(29+11)+29·11	2519
	T_b	11518

Fra gli indici di cograduazione proposti in letteratura per tabelle a doppia entrata, analizziamo il Γ di Goodman e Kruskal e il τ di Kendall, che nell'esempio suddetto assumono i seguenti valori:

$$\Gamma = \frac{N_c - N_d}{N_c + N_d} = -0{,}73$$

$$\tau = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_a)(N_c + N_d + T_b)}} = -0{,}53\,.$$

Entrambi gli indici variano tra –1 e +1, assumendo valori positivi in caso di concordanza e valori negativi in caso di discordanza. In particolare, per tabelle quadrate, l'indice τ assume i valori estremi solo in caso di perfetta concordanza (discordanza), ossia quando tutte le frequenze congiunte si dispongono sulla diagonale principale (secondaria). L'indice Γ, invece, assume valore –1 quando N_c=0 e valore +1 quando N_d=0. L'indice τ pertanto può ritenersi migliore dell'indice Γ.

In caso di indipendenza tali indici sono uguali a 0, ma non è vero il contrario. Infatti, sia Γ che τ valgono 0 se $N_c = N_d$.

Nell'esempio considerato Γ e τ, pur assumendo valori diversi, mostrano una discordanza fra i due caratteri, ossia al peggiorare delle condizioni climatiche, ad esempio in caso di pioggia, il traffico automobilistico tende ad aumentare.

7.6 Interdipendenza fra variabili quantitative

Supponiamo di aver rilevato su *n* unità statistiche due variabili quantitative *X* ed *Y*. Per misurare l'interdipendenza lineare fra due variabili quantitative ci serviamo della *covarianza*, data dalla media del prodotto degli scarti delle due variabili dalla propria media:

$$\sigma_{XY} = \frac{\sum_{i=1}^{n} (x_i - M_X)(y_i - M_Y)}{n}$$

La covarianza assume valori positivi se vi è una prevalenza di scarti concordi; in tal caso le due variabili sono interdipendenti linearmente in modo diretto, dunque al crescere (decrescere) di una di esse, l'altra cresce (decresce). La covarianza assume valori negativi se vi è una prevalenza di scarti discordi; in tal caso, le

variabili sono interdipendenti linearmente in modo inverso e al crescere di una l'altra decresce e viceversa.

In particolare, secondo la disuguaglianza di Cauchy-Schwartz, si può definire un range all'interno del quale la covarianza può variare:

$$-\sigma_X \sigma_Y \leq \sigma_{XY} \leq +\sigma_X \sigma_Y$$

Dividendo ciascun membro della disuguaglianza per $\sigma_X \sigma_Y$, si ottiene il *coefficiente di correlazione lineare di Bravais-Pearson*:

$$-1 \leq \rho \leq +1,$$

che assume i valori estremi, –1 e +1, in caso di perfetta relazione lineare fra le due variabili.

L'indice ρ è un numero adimensionale, poiché numeratore e denominatore sono espressi nella stessa unità di misura:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

mentre la covarianza è espressa nel prodotto delle unità di misura delle due variabili.

Così come per la varianza, esiste una formula ridotta anche per la covarianza. Infatti è:

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - M_X)(y_i - M_Y)}{n} = \frac{\sum_{i=1}^{n}(x_i y_i - x_i M_Y - y_i M_X + M_X M_Y)}{n} =$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - M_Y \sum_{i=1}^{n} x_i - M_X \sum_{i=1}^{n} y_i + n M_X M_Y}{n} =$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{n} - M_Y M_X - M_X M_Y + M_X M_Y = M_{XY} - M_X M_Y.$$

Tale formula consente di calcolare la covarianza, evitando di calcolare tutti gli scarti di X e di Y dalle rispettive medie.

Quando non si dispone di una serie doppia di osservazioni, ma di una tabella doppia di frequenza, per calcolare la covarianza bisogna tener conto delle frequenze congiunte:

$$\sigma_{XY} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(x_i - M_X)(y_j - M_Y)n_{ij}}{N}$$

E' conveniente, anche in questo caso, calcolare la covarianza con la formula ridotta:

$$\sigma_{XY} = M_{XY} - M_X M_Y,$$

dove, però, le medie aritmetiche sono medie aritmetiche ponderate:

$$M_{XY} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} x_i y_j n_{ij}}{N}, \qquad M_X = \frac{\sum_{i=1}^{r} x_i n_{i.}}{N}, \qquad M_Y = \frac{\sum_{j=1}^{c} y_j n_{.j}}{N}.$$

Se le due variabili X ed Y sono indipendenti in distribuzione, la covarianza, è nulla. Infatti, se X ed Y sono indipendenti in distribuzione (cfr.par. 7.2) è $n_{ij} = \frac{n_{i.} n_{.j}}{N}$, quindi è lecito scrivere:

$$\sigma_{XY} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(x_i - M_X)(y_j - M_Y)n_{i.}n_{.j}}{N^2} = \frac{\sum_{i=1}^{r}(x_i - M_X)n_{i.}}{N}\frac{\sum_{j=1}^{c}(y_j - M_Y)n_{.j}}{N} = 0$$

in quanto, per la prima proprietà della media aritmetica, la somma degli scarti di ciascuna variabile dalla propria media è nulla:

$$\sum_{i=1}^{r}(x_i - M_X)n_{i.} = \sum_{j=1}^{c}(y_j - M_Y)n_{.j} = 0.$$

Ovviamente, in tal caso, è nullo anche il coefficiente di correlazione, pertanto due variabili indipendenti sono anche non correlate, ma non è vero il contrario.

7.6.1 Esempi di calcolo della covarianza e di ρ

Nella seguente tabella, sono riportati i Tassi di attività lavorativa (X) della popolazione e il Prodotto interno lordo per abitante (Y), in milioni di lire, di otto regioni italiane nel 1979:

REGIONI	TAL(x_i)	PIL(y_i)
Piemonte	63	6,0
Lombardia	61	6,3
Liguria	55	6,2
Toscana	60	5,3
Emilia	64	5,9
Lazio	53	4,6
Puglia	55	3,3
Sicilia	50	3,2
TOTALE	461	40,8

Si vuol verificare se le due variabili X ed Y sono correlate.

Calcoliamo, innanzitutto la covarianza:

x_i-M_X	y_i-M_Y	$(x_i$-$M_X)(y_i$-$M_Y)$	$(x_i$-$M_X)^2$	$(y_i$-$M_Y)^2$
5,4	0,9	4,9	29,2	0,8
3,4	1,2	4,1	11,6	1,4
-2,6	1,1	-2,9	6,8	1,2
2,4	0,2	0,5	5,8	0,0
6,4	0,8	5,1	41,0	0,6
-4,6	-0,5	2,3	21,2	0,3
-2,6	-1,8	4,7	6,8	3,2
-7,6	-1,9	14,4	57,8	3,6
		33,1	179,9	11,2

$$M_X = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{461}{8} = 57{,}6 \qquad M_Y = \frac{\sum_{i=1}^{8} y_i}{8} = \frac{40{,}8}{8} = 5{,}1$$

$$\sigma_{XY} = \frac{\sum_{i=1}^{8} (x_i - M_X)(y_i - M_Y)}{8} = \frac{33{,}1}{8} = 4{,}1$$

quindi le due varianze:

$$\sigma_X^2 = \frac{\sum_{i=1}^{8} (x_i - M_X)^2}{8} = \frac{179{,}9}{8} = 22{,}485 \qquad \sigma_Y^2 = \frac{\sum_{i=1}^{8} (y_i - M_Y)^2}{8} = \frac{11{,}2}{8} = 1{,}405\,.$$

Il coefficiente di correlazione è:

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{4{,}1}{\sqrt{22{,}485 \cdot 1{,}405}} = 0{,}736 .$$

Volendo utilizzare le formule ridotte, sia per il calcolo della covarianza che delle due varianze, non sono necessari gli scarti; basta determinare le seguenti tre colonne:

$x_i y_i$	x_i^2	y_i^2
378,0	3.969	36,0
384,3	3.721	39,7
341,0	3.025	38,4
318,0	3.600	28,1
377,6	4.096	34,8
243,8	2.809	21,2
181,5	3.025	10,9
160,0	2.500	10,2
2384,2	26745	219,3

$$M_{XY} = \frac{\sum_{i=1}^{8} x_i y_i}{8} = \frac{2384{,}2}{8} = 298$$

$$\sigma_{XY} = M_{XY} - M_X M_Y = 298 - 57{,}6 * 5{,}1 = 4{,}1$$

$$\sigma_X^2 = \frac{\sum_{i=1}^{8} x_i^2}{8} - M_X^2 = \frac{26745}{8} - (57{,}6)^2 = 22{,}485$$

$$\sigma_Y^2 = \frac{\sum_{i=1}^{8} y_i^2}{8} - M_Y^2 = \frac{219{,}3}{8} - (5{,}1)^2 = 1{,}405$$

Si perviene, dunque, allo stesso risultato, a meno di approssimazioni:

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{4{,}1}{\sqrt{22{,}485 \cdot 1{,}405}} = 0{,}736 .$$

Tale valore sta ad indicare una correlazione positiva fra le due variabili, ovvero all'aumentare del PIL, aumenta anche il TAL e viceversa.

Supponiamo, adesso di aver osservato la seguente tabella a doppia entrata:

X/Y	*19*	*22,5*	*26*	*29*	*totale*
39	3	4	0	0	*7*
45,5	0	17	6	0	*23*
51,5	0	0	18	0	*18*
57,5	0	0	7	17	*24*
totale	*3*	*21*	*31*	*17*	*72*

Calcoliamo il coefficiente di correlazione:

$y_j \cdot n_{.j}$	$y_j^2 \cdot n_{.j}$
57	1083
472,5	10631,3
806	20956
493	14297
1828,5	46967,3

$x_i \cdot n_{i.}$	$x_i^2 \cdot n_{i.}$
273	10647
1046,5	47615,75
927	47740,5
1380	79350
3626,5	185353,3

$$M_X = \frac{\sum_{i=1}^{4} x_i n_{i.}}{N} = \frac{3626,5}{72} = 50,4 \qquad M_Y = \frac{\sum_{i=1}^{4} y_j n_{.j}}{N} = \frac{1828,5}{72} = 25,4$$

$$\sum_{i=1}^{4}\sum_{j=1}^{4} x_i y_j n_{ij} = 39 \cdot 19 \cdot 3 + 39 \cdot 22,5 \cdot 4 + \ldots + 57,5 \cdot 29 \cdot 17 = 93149,25$$

$$M_{XY} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} x_i y_j n_{ij}}{N} = \frac{93149,25}{72} = 1293,7$$

$$\sigma_{XY} = M_{XY} - M_X M_Y = 1293,7 - 50,4 \cdot 25,4 = 13,5$$

$$\sigma_X^2 = \frac{\sum_{i=1}^{4} x_i^2 n_{i.}}{72} - M_X^2 = \frac{185353,3}{72} - (50,4)^2 = 37,4$$

$$\sigma_Y^2 = \frac{\sum_{j=1}^{4} y_i^2 n_{.j}}{72} - M_Y^2 = \frac{46967,3}{72} - (25,4)^2 = 7,4$$

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{13,5}{\sqrt{37,4 \cdot 7,4}} = 0,82\,.$$