Capitolo 1

Matrici statistiche

1.1 Ouverture

Mi capita spesso di domandare che nell'insegnamento si faccia un maggiore uso dei metodi che favoriscono l'intuizione e di sentirmi rispondere che innanzitutto bisogna sviluppare il senso del rigore. Questa risposta implica un malinteso che mi stupisce sempre; a mio avviso è un errore contrapporre intuizione e rigore. L'intuizione favorisce la scoperta ed aiuta a comprenderla; una volta che ci ha suggerito un enunciato, si tratta di dimostrarlo rigorosamente. L'intuizione è utile per arrivare ad un risultato come gli occhi lo sono al viandante che si serve dei suoi piedi per arrivare alla meta, ed è così assurdo il combatterla come lo sarebbe il raccomandare al viandante di chiudere gli occhi per arrivare più sicuramente alla meta. L'intuizione ed il rigore scientifico sono due qualità che si completano ed il ruolo dell'educatore deve essere quello di sviluppare il senso di rigore in coloro in cui è carente, senza per questo prendere l'abitudine di combattere il senso dell'intuizione ovunque lo si incontri.

Paul Levy, *Théorie de l'addition des variables aléatoires*, Gauthier-Villars, Paris, 1954

Per quelli che non conoscono la matematica è difficile arrivare al vero apprezzamento della bellezza, la grandissima bellezza della natura. (...) E' un peccato che debba essere la matematica, e che questa sia difficile per certe persone. Si dice - non so se sia vero - che un certo re, il quale stava cercando di imparare la geometria da Euclide, si lamentasse che era difficile. E questi gli rispose: "Non c'è una via regia per la geometria". E in effetti non c'è nessuna via regia: i fisici non possono tradurre in nessun'altra lingua. Se volete conoscere e apprezzare la natura è necessario capire la lingua che parla. Essa offre la sua informazione solo in una forma; noi non dobbiamo essere così superbi da pretendere che essa cambi prima di prestarle attenzione. Tutte le disquisizioni intellettuali che potete fare non riusciranno a comunicare a orecchie sorde quella che veramente è l'esperienza della musica. Allo stesso modo tutte le disquisizioni intellettuali del mondo non daranno una comprensione della natura a quelli dell'"altra cultura". I filosofi possono cercare di insegnarvela parlandovi qualitativamente della natura, io invece cerco di descriverla. Tuttavia non ci riesco perché è impossibile. Forse è perché i loro orizzonti sono così limitati, che alcuni possono immaginare che il centro dell'universo sia l'uomo.

Richard P. Feynman, *La legge fisica*, Boringhieri, 1971

1.2 Matrici di dati

Sia $\Omega = \{\omega_1, \dots, \omega_n\}$ una popolazione finita di n unità statistiche e si supponga di considerarne lo studio attraverso p variabili statistiche $X_1, \dots, X_p$ a valori rispettivamente in $\mathcal{X}_1, \dots, \mathcal{X}_p$. Dopo aver completato tutte le fasi della rilevazione, viene costruita la *matrice dei dati*:

$$\begin{array}{c} \omega_1 \\ \vdots \\ \omega_h \\ \vdots \\ \omega_n \end{array} \begin{pmatrix} X_1 & \cdots & X_i & \cdots & X_p \\ x_{11} & \cdots & x_{1i} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{h1} & \cdots & x_{hi} & \cdots & x_{hp} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{ni} & \cdots & x_{np} \end{pmatrix} .$$

in cui la colonna i -esima fornisce la distribuzione della variabile X_i fra le n unità statistiche rilevate; la riga j -esima fornisce le modalità rilevate dalle p variabili $X_1, \dots, X_p$ nella j -esima unità statistica ω_j . Matrici di dati di questo tipo vengono anche chiamate matrici *unità per variabili*.

Esempio 1.1 La tabella seguente riporta i dati relativi ad alcuni alberghi nella provincia di Catanzaro rilevati nel 2001¹. Si tratta di una matrice avente $n = 15$ righe (unità statistiche = alberghi) e $p = 5$ variabili:

X_1 = Comune di residenza (carattere qualitativo sconnesso);

X_2 = Ragione sociale (carattere qualitativo sconnesso);

X_3 = Categoria (carattere qualitativo ordinale);

X_4 = Posti letto (carattere quantitativo);

X_5 = Numero stanze (carattere quantitativo).

Si noti che nell'ultima riga, che riguarda il *Camping Residence Costa Blu*, non sono stati registrati né la categoria, né il numero di posti letti, né il numero di stanze. Si tratta di dati mancanti su cui torneremo in seguito.

La matrice di dati completa è contenuta nel file *alberghi.xls* insieme al tracciato record che contiene le informazioni sulle variabili rilevate e sulle fonti da cui provengono i dati, che in questo caso sono: annuario alberghi, elenchi Telecom, Pagine gialle Seat, Pagine utili.

¹Fonte: CCIAA Catanzaro, Dati anno 2001, file: *alberghi.xls*; disponibile alla pagina: www.economia.unical.it/statistica/didattica/MSA2/datiSA2.htm

unità	Comune	Ragione sociale	Categoria	P.letto	N. stanze
1	Sellia Marina	106	2	28	18
2	Cropani Marina	4 Lampioni	1	24	12
3	Platania	A Giurranda	2	40	20
4	Montepaone	A Lumera	3	68	35
5	Lamezia Terme	Aerhotel Phelipe	3	26	16
6	Montepaone	Alba Chiara Hotel Residence	3	80	40
7	Tiriolo	Autostello Chiarella	2	15	7
8	Stalettì	Baia Dell'Est	3	60	30
9	Catanzaro	Bellamena residence hotel	3	70	37
10	Badolato	Bell'orizzonte	2	80	40
11	Catanzaro	Belvedere	2	24	16
12	Catanzaro	Benny Hotel	3	152	76
13	Montepaone	Calaghena Hotel Villaggio	4	273	145
14	Decollatura	Caligiuri	3	18	10
15	Sellia Marina	Camping Residence Costa Blu	NA	NA	NA

In generale, una matrice di dati $n \times p$ verrà denotata col simbolo $\tilde{\mathbf{X}}$ o $\mathbf{X}_{n \times p}$, l'elemento della i -esima riga e j -esima colonna è x_{ij} e denota il valore osservato per la j -esima variabile nella i -esima unità statistica. Nel caso precedente si tratta di una matrice di dati 15×5 . Il generico vettore di $\mathbb{R}^n$ viene indicato con $\mathbf{x}$; usualmente $\mathbf{x}$ denota un vettore colonna, cioè:

$$\mathbf{x} := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Pertanto le righe di $\tilde{\mathbf{X}}$ possono essere scritte come $\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n$ ² e le colonne di $\tilde{\mathbf{X}}$ verranno scritte con l'indice fra parentesi, cioè $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots, \mathbf{x}_{(p)}$. Riassumendo, possiamo scrivere:

$$\tilde{\mathbf{X}} := \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots, \mathbf{x}_{(p)}) = (x_{hi})$$

²In genere con $\mathbf{x}$ si intende un vettore colonna e con $\mathbf{x}'$ o $\mathbf{X}'$ denota rispettivamente vettore trasposto o matrice trasposta.

dove

$$\mathbf{x}_h := \begin{pmatrix} x_{h1} \\ x_{h2} \\ \vdots \\ x_{hp} \end{pmatrix} \quad \text{e} \quad \mathbf{x}_{(i)} := \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix}$$

per $h = 1, 2, \dots, n$ e $i = 1, 2, \dots, p$.

Si noti che $\mathbf{x}_i$ è il vettore p -dimensionale che denota le p osservazioni sulla i -esima unità statistica, mentre $\mathbf{x}_{(i)}$ è il vettore n dimensionale che denota le n osservazioni sulla i -esima variabile, cioè su X_i . Ad esempio, con riferimento alla precedente matrice di dati si ha:

$$\begin{aligned} \mathbf{x}_7 &= (\text{Tiriolo, Autostello Chiarella, 2, 15, 7})' \\ \mathbf{x}_{12} &= (\text{Catanzaro, Benny Hotel, 3, 152, 76})' \end{aligned}$$

e

$$\begin{aligned} \mathbf{x}_{(1)} &= (\text{Sellia Marina, Cropani Marina, Platania, Montepaone, Lamezia Terme,} \\ &\quad \text{Montepaone, Tiriolo, Stalettì, Catanzaro, Badolato, Catanzaro, Catanzaro,} \\ &\quad \text{Montepaone, Decollatura, Sellia Marina})' \\ \mathbf{x}_{(3)} &= (2, 1, 2, 3, 3, 3, 2, 3, 3, 2, 2, 3, 4, 3, \text{NA})' \end{aligned}$$

dove NA denota "dato mancante" (not available).

Nell'analisi multivariata, le righe $\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n$ costituiscono usualmente un campione della popolazione, al contrario ciò non si verifica per le colonne $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots, \mathbf{x}_{(p)}$. Chiaramente quando il numero n di unità rilevate ed il numero p di caratteri da analizzare è moderatamente grande, il numero totale di informazioni np può risultare molto grande, per cui risulta essenziale sintetizzare in maniera opportuna i dati rilevati.

Infine rileviamo che le matrici di dati possono essere studiate secondo due diversi punti di vista: l'analisi delle colonne conduce allo studio della relazione fra le variabili; l'analisi delle righe conduce allo studio delle relazioni fra unità statistiche differenti.

Matrici di frequenza (tabelle a doppia entrata) Consideriamo il caso in cui la popolazione Ω viene studiata in base a due caratteri X e Y aventi rispettivamente h e k modalità, cioè $x_1, \dots, x_k$ e $y_1, \dots, y_h$. I dati possono essere riassunti in una *matrice di frequenza* o *tabella a doppia entrata* in cui a ciascuna coppia di modalità (x_i, y_j) di (X, Y) si fa corrispondere la sua frequenza assoluta n_{ij} $i = 1, 2, \dots, k, j = 1, 2, \dots, h$, cioè il numero di volte in cui la coppia di modalità si è presentata:

	y_1	$\dots$	y_j	$\dots$	y_h	
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1h}	n_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ih}	n_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	n_{k1}	$\dots$	n_{kj}	$\dots$	n_{kh}	n_{k0}
	n_{01}	$\dots$	n_{0j}	$\dots$	n_{0h}	N

dove:

$$n_{i0} = \sum_{j=1}^h n_{ij}, \quad n_{0j} = \sum_{i=1}^k n_{ij} \quad \text{e} \quad N = \sum_{i=1}^k \sum_{j=1}^h n_{ij}$$

Le frequenze N_{i0} , $i = 1, 2, \dots, k$ vengono chiamate *frequenze marginali assolute* di X ; le frequenze N_{0j} , $j = 1, 2, \dots, h$ vengono chiamate *frequenze marginali assolute* di Y ; N è ovviamente il numero di elementi della popolazione.

Ad esempio, la seguente matrice di frequenza fornisce il numero di posti di lavoro creati dalle imprese italiane nel 1998 per ramo di attività produttiva e zona geografica³:

Ramo	Nord	Centro	Sud e Isole	TOTALE
Energia, gas, acqua	198	45	79	322
Industrie estrattive e chimiche	4248	1417	1893	7558
Industrie manifatturiere dei metalli	27683	5688	4991	38362
Altre industrie manifatturiere	23063	9898	8958	41919
Industrie per l'edilizia	19885	8321	16101	44307
Commercio, pubblici esercizi e alberghi	41208	16088	18289	75585
Trasporto e comunicazioni	4764	1748	2635	9147
Credito e assicurazione	13265	4423	3836	21524
Servizi pubblici e privati	11238	4407	4749	20394
TOTALE	134314	47628	56782	238724

Ovviamente possono essere introdotte anche in questo caso le frequenze relative percentuali:

$$f_{ij} = \frac{n_{ij}}{N} \quad i = 1, 2, \dots, k, \quad j = 1, 2, \dots, h \quad (1.1)$$

e quindi riscrivere la tabella a doppia entrata in funzione di tali frequenze:

	y_1	$\dots$	y_j	$\dots$	y_h	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1h}	f_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{ih}	f_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	f_{k1}	$\dots$	f_{kj}	$\dots$	f_{kh}	f_{k0}
	f_{01}	$\dots$	f_{0j}	$\dots$	f_{0h}	100

dove:

$$f_{i0} = \sum_{j=1}^h f_{ij} \quad \text{e} \quad f_{0j} = \sum_{i=1}^k f_{ij}$$

Le frequenze f_{i0} , $i = 1, 2, \dots, k$ vengono chiamate *frequenze marginali relative* di X ; le frequenze f_{0j} , $j = 1, 2, \dots, h$ vengono chiamate *frequenze marginali relative* di Y .

Considerando le frequenze relative percentuali, la precedente tabella si scrive:

³Fonte: Banca dati INPS, www.inps.it.

<i>Ramo</i>	<i>Nord</i>	<i>Centro</i>	<i>Sud e Isole</i>	TOTALE
Energia, gas, acqua	0.08	0.02	0.03	0.13
Industrie estrattive e chimiche	1.78	0.59	0.79	3.17
Industrie manifatturiere dei metalli	11.60	2.38	2.09	16.07
Altre industrie manifatturiere	9.66	4.15	3.75	17.56
Industrie per l'edilizia	8.33	3.49	6.74	18.56
Commercio, pubblici esercizi e alberghi	17.26	6.74	7.66	31.66
Trasporto e comunicazioni	2.00	0.73	1.10	3.83
Credito e assicurazione	5.56	1.85	1.61	9.02
Servizi pubblici e privati	4.71	1.85	1.99	8.54
TOTALE	56.26	19.95	23.79	100.00

Matrice di origine-destinazione La matrice di origine-destinazione è una matrice di frequenze in cui i valori n_{ij} sono le frequenze degli spostamenti dalla posizione sulla riga i -esima alla posizione sulla colonna j -esima.

Ad esempio, la seguente tabella fornisce il numero di arrivi degli italiani negli esercizi ricettivi per le regioni di provenienza e destinazione nell'area sud e isole per l'anno 2000⁴

<i>Regione di provenienza</i>	<i>Regione di arrivo</i>					
	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
Campania	775337	116371	41356	120367	136728	39322
Puglia	134421	208380	60197	88092	88739	16796
Basilicata	48034	17188	26866	13677	12835	2810
Calabria	69534	50355	13401	109307	87739	5696
Sicilia	90594	113284	15286	116679	870498	22023
Sardegna	21449	14077	1367	3128	21860	221201

La matrice completa è riportata nel file *turisti.xls*⁵.

Altri esempi di matrici origine-destinazione riguardano: il movimento di pazienti che vengono curati in strutture sanitarie della stessa regione di residenza oppure diversa; clienti di telefonia mobile che continuano ad usufruire del servizio di uno stesso gestore di rete oppure lo cambiano.

1.3 Stastistiche descrittive per matrici di dati quantitativi

Quando la matrice di dati contiene unicamente variabili quantitative, è possibile effettuare opportune sintesi, che sono essenzialmente generalizzazioni delle statistiche nel caso univariato e bivariato.

⁴Fonte: ISTAT, *Statistiche del Turismo*, anno 2000, tav. 2.46.

⁵disponibile alla pagina: www.economia.unical.it/statistica/didattica/MSA2/datiSA2.htm

Consideriamo un matrice di dati che concerne lo studio dei vini di di Bordeaux in base alla loro qualità. La tabella riporta 34 osservazioni climatiche dei mesi da aprile a settembre nelle annate dal 1924 al 1957, dove:

- X_1 : somma delle temperature medie giornaliere in gradi;
- X_2 : numero di ore di insolazione;
- X_3 : numero di giorni di grande caldo;
- X_4 : quantità di pioggia in mm.

Di seguito riportiamo solo i valori concernenti le prime 10 unità statistiche.

Annata	X_1	X_2	X_3	X_4	Qualità
24	3064	1201	10	361	+
25	3000	1053	11	338	-
26	3155	1133	19	393	+
27	3085	970	4	467	--
28	3245	1258	36	294	++
29	3267	1386	35	225	+++
30	3080	966	13	417	--
31	2974	1185	12	488	--
32	3038	1103	14	677	---
33	3318	1310	29	427	+

L'ultima variabile fornisce una valutazione di qualità, secondo 7 categorie da +++ a --- ed è stata inserita successivamente in base al giudizio espresso da alcuni *sommeliers*.

Il vettore delle medie e la matrice di covarianza Un'ovvia estensione delle nozioni univariate di media (aritmetica) e varianza conduce alle definizioni seguenti. Se si considera la generica variabile X_i , $i = 1, 2, \dots, p$, essa avrà media e varianza data rispettivamente da:

$$\bar{x}_i := \frac{1}{n} \sum_{h=1}^n x_{hi} \quad (1.2)$$

$$\sigma_{ii} := \frac{1}{n} \sum_{h=1}^n (x_{hi} - \bar{x}_i)^2 = \frac{1}{n} \sum_{h=1}^n x_{hi}^2 - \bar{x}_i^2 = \sigma_i^2 \quad (1.3)$$

per $i = 1, 2, \dots, p$.

Assegnata una coppia di variabili X_i, X_j , con $i, j = 1, 2, \dots, p$, la covarianza fra X_i e X_j è definita da:

$$\sigma_{ij} := \frac{1}{n} \sum_{h=1}^n (x_{hi} - \bar{x}_i)(x_{hj} - \bar{x}_j) = \frac{1}{n} \sum_{h=1}^n x_{hi}x_{hj} - \bar{x}_i\bar{x}_j. \quad (1.4)$$

I valori delle medie di tutte le p variabili $X_1, X_2, \dots, X_p$ possono essere descritti sinteticamente mediante un vettore p -dimensionale, chiamato *vettore delle medie*, avente per coordinate

le medie delle singole variabili:

$$\bar{\mathbf{x}} := \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = (\bar{x}_1, \dots, \bar{x}_p)' . \quad (1.5)$$

Per esempio, con riferimento all'intera matrice di dati per il caso precedente, si ha:

$$\mathbf{x} = (3158.76, 1248.71, 18.82, 360.44)'$$

Analogamente la covarianza fra tutte le possibili coppie di variabili X_i, X_j , con $i, j = 1, 2, \dots, p$, può essere sinteticamente data mediante una matrice quadrata $\mathbf{S}$ di ordine $p \times p$ chiamata *matrice di varianze e covarianze* o più sinteticamente, *matrice di covarianze*:

$$\mathbf{S}_{p \times p} := (\sigma_{ij}) ,$$

dove, in particolare, $\sigma_{ii} = s_i^2$ è la varianza della i -esima variabile. Essendo $\sigma_{ij} = \sigma_{ji}$, segue che $\mathbf{S}$ è una matrice simmetrica. Nel caso dei vini di Bordeaux si ha:

$$\mathbf{S} = \begin{pmatrix} 20209.00 & 12929.69 & 1229.02 & -5361.47 \\ 12929.69 & 16130.34 & 821.70 & -5514.47 \\ 1229.02 & 821.70 & 100.33 & -367.25 \\ -5361.47 & -5514.47 & -367.25 & 8548.25 \end{pmatrix} .$$

Le statistiche precedenti possono essere espresse in notazione matriciale come segue. In base a (1.2) e (1.5) si ha:

$$\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_p)' = \frac{1}{n} \sum_{h=1}^n \mathbf{x}_h = \frac{1}{n} \left(\sum_{h=1}^n x_{h1}, \sum_{h=1}^n x_{h2}, \dots, \sum_{h=1}^n x_{hp} \right)' . \quad (1.6)$$

Denotato con $\mathbf{1} = (1, \dots, 1)' \in \mathbb{R}^n$ il vettore unitario n dimensionale, in termini matriciali, possiamo scrivere:

$$\frac{1}{n} \mathbf{X}' \mathbf{1} \underset{\sim}{=} \frac{1}{n} \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \cdots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{n} \left(\sum_{h=1}^n x_{h1}, \sum_{h=1}^n x_{h2}, \dots, \sum_{h=1}^n x_{hp} \right)' = \bar{\mathbf{x}} . \quad (1.7)$$

Inoltre dalla (1.4) si ha:

$$\sigma_{ij} = \frac{1}{n} \sum_{h=1}^n x_{hi} x_{hj} - \bar{x}_i \bar{x}_j . \quad (1.8)$$

Per quanto riguarda la matrice di varianze e covarianze $\mathbf{S}$, preliminarmente notiamo che il termine $(x_{hi} - \bar{x}_i)(x_{hj} - \bar{x}_j)$, per $h = 1, \dots, n$, può essere visto come l'elemento ij della matrice

$(\mathbf{x}_h - \bar{\mathbf{x}})(\mathbf{x}_h - \bar{\mathbf{x}})'$; analogamente il termine $x_{hi}x_{hj}$ può essere visto come l'elemento ij della matrice $\mathbf{x}_h\mathbf{x}_h'$. Si ha infatti, ad esempio:

$$\mathbf{x}_h\mathbf{x}_h' = \begin{pmatrix} x_{h1} \\ x_{h2} \\ \vdots \\ x_{hp} \end{pmatrix} (x_{h1}, x_{h2}, \dots, x_{hp}) = \begin{pmatrix} x_{h1}^2 & x_{h1}x_{h2} & \cdots & x_{h1}x_{hp} \\ x_{h2}x_{h1} & x_{h2}^2 & \cdots & x_{h2}x_{hp} \\ \cdots & \cdots & \cdots & \cdots \\ x_{hp}x_{h1} & x_{hp}x_{h2} & \cdots & x_{hp}^2 \end{pmatrix} = (x_{hi}x_{hj}).$$

Allora, per le (1.4) e (1.8), possiamo scrivere:

$$\mathbf{S}_{p \times p} := (\sigma_{ij}) = \frac{1}{n} \sum_{h=1}^n (\mathbf{x}_h - \bar{\mathbf{x}})(\mathbf{x}_h - \bar{\mathbf{x}})' = \frac{1}{n} \sum_{h=1}^n \mathbf{x}_h\mathbf{x}_h' - \bar{\mathbf{x}}\bar{\mathbf{x}}'. \quad (1.9)$$

Tenendo conto che $\sum_{h=1}^n \mathbf{x}_h\mathbf{x}_h' = \underset{\sim}{\mathbf{X}}' \underset{\sim}{\mathbf{X}}$, sostituendo, in base alla (1.7), segue:

$$\mathbf{S} = \frac{1}{n} \underset{\sim}{\mathbf{X}}' \underset{\sim}{\mathbf{X}} - \bar{\mathbf{x}}\bar{\mathbf{x}}' = \frac{1}{n} \left(\underset{\sim}{\mathbf{X}}' \underset{\sim}{\mathbf{X}} - \frac{1}{n} \underset{\sim}{\mathbf{X}}' \mathbf{1} \mathbf{1}' \underset{\sim}{\mathbf{X}} \right) = \frac{1}{n} \underset{\sim}{\mathbf{X}}' \left(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}' \right) \underset{\sim}{\mathbf{X}}, \quad (1.10)$$

dove $\mathbf{I}$ è la matrice identità $n \times n$. Se introduciamo la *matrice centrante* $\mathbf{H}$:

$$\mathbf{H}_{n \times n} := \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}' = \mathbf{I} - \frac{1}{n} \mathbf{J} \quad (1.11)$$

dove $\mathbf{I}$ è la matrice identità e $\mathbf{J} = \mathbf{1} \mathbf{1}'$, si ottiene:

$$\mathbf{S} = \frac{1}{n} \underset{\sim}{\mathbf{X}}' \mathbf{H} \underset{\sim}{\mathbf{X}}, \quad (1.12)$$

che costituisce un'utile rappresentazione della matrice di varianze e covarianze.

La matrice $\mathbf{H}$ è una matrice simmetrica e idempotente. Infatti dalla (1.11) si vede immediatamente che è combinazione lineare di matrici simmetriche; inoltre, tenendo conto che $\mathbf{1}'\mathbf{1} = n$, si ha:

$$\begin{aligned} \mathbf{H}^2 &= \left(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}' \right) \left(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}' \right) = \mathbf{I} - \frac{2}{n} \mathbf{1} \mathbf{1}' + \frac{1}{n^2} \mathbf{1} \mathbf{1}' \mathbf{1} \mathbf{1}' = \mathbf{I} - \frac{2}{n} \mathbf{1} \mathbf{1}' + \frac{1}{n^2} n \mathbf{1} \mathbf{1}' \\ &= \mathbf{I} - \frac{2}{n} \mathbf{1} \mathbf{1}' + \frac{1}{n} \mathbf{1} \mathbf{1}' = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}' = \mathbf{H}. \end{aligned}$$

Proposizione 1.2 Sia $\underset{\sim}{\mathbf{X}}$ una matrice di dati $n \times p$ di rango p , con $n \geq p + 1$. Allora la matrice $\mathbf{S}$ di varianze e covarianze di $\underset{\sim}{\mathbf{X}}$ è semidefinita positiva.

Dimostrazione. Dobbiamo dimostrare che per ogni vettore $\mathbf{a} \in \mathbb{R}^p$ risulta $\mathbf{a}'\mathbf{S}\mathbf{a} \geq 0$. In effetti si ha:

$$\mathbf{a}'\mathbf{S}\mathbf{a} = \frac{1}{n} \mathbf{a}' \underset{\sim}{\mathbf{X}}' \mathbf{H} \underset{\sim}{\mathbf{X}} \mathbf{a}$$

e quindi posto $\mathbf{y} = \underset{\sim}{\mathbf{H}} \mathbf{X} \mathbf{a}$, segue:

$$\mathbf{a}'\mathbf{S}\mathbf{a} = \frac{1}{n} \mathbf{y}' \mathbf{y} \geq 0,$$

cioè la tesi. ♣

Nota 1.3 Nel caso in cui la matrice di dati costituisce la realizzazione di un campione casuale estratto dalla popolazione (secondo la distribuzione di $\mathbf{X}$), si considera la seguente matrice di varianze e covarianze:

$$\mathbf{S}_c = \frac{1}{n-1} \underset{\sim}{\mathbf{X}}' \underset{\sim}{\mathbf{H}} \underset{\sim}{\mathbf{X}} = \frac{n}{n-1} \mathbf{S}$$

che costituisce uno stimatore non distorto della matrice di varianze e covarianza della popolazione.

Nei casi pratici la matrice $\mathbf{S}$ risulta essere definita positiva se risulta $n \geq p + 1$ e scriveremo $\mathbf{S} > 0^6$

La matrice di correlazione. Il coefficiente di correlazione lineare semplice fra due variabili X_i e X_j è definito da

$$r_{ij} := \frac{\sigma_{ij}}{\sigma_i \sigma_j} . \quad (1.13)$$

dove, come ben noto, si ha $-1 \leq r_{ij} \leq 1$, e $r_{ii} = 1$.

Analogamente a quanto visto in precedenza, si può quindi introdurre la *matrice di correlazione* $\mathbf{R}$:

$$\mathbf{R} := (r_{ij}) . \quad (1.14)$$

La matrice di correlazione, in base alla (1.13), è quindi simmetrica. Nel caso dei vini di Bordeaux si ha:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.7147 & 0.8614 & -0.4118 \\ 0.7147 & 1.0000 & 0.6459 & -0.4750 \\ 0.8614 & 0.6459 & 1.0000 & -0.4011 \\ -0.4118 & -0.4750 & -0.4011 & 1.0000 \end{pmatrix} . \quad (1.15)$$

Posto $\mathbf{D} = \text{diag}(\sigma_i)$, dove $\sigma_i = \sqrt{\sigma_{ii}}$, in funzione della matrice di covarianza si ha la seguente relazione che generalizza la definizione di coefficiente di correlazione lineare al caso matriciale:

$$\mathbf{R} = \mathbf{D}^{-1} \mathbf{S} \mathbf{D}^{-1} , \quad (1.16)$$

e pertanto $\mathbf{R}$ è semidefinita positiva (per un noto risultato dell'algebra lineare)⁷ In particolare se $\mathbf{R} = \mathbf{I}$, diremo che le variabili sono non correlate.

Misure di dispersione multivariata. Come già osservato, la matrice di covarianza $\mathbf{S}$ è una generalizzazione della nozione di varianza introdotta nel caso univariato. Spesso comunque, anche nel caso multivariato, è conveniente avere un singolo numero che esprima una misura della dispersione dei dati.

Due misure che usualmente vengono considerate a tal fine sono

1. la *variazione totale* VAR_T : la traccia della matrice di covarianze $\text{tr}\mathbf{S}$;

⁶Assegnata una matrice quadrata $\mathbf{A}$, nel seguito adottiamo la seguente notazione: se $\mathbf{A}$ è definita positiva, scriveremo $\mathbf{A} > 0$; se $\mathbf{A}$ è semidefinita positiva, scriveremo $\mathbf{A} \geq 0$; analogamente per le matrici definite negative e semidefinite negative scriveremo rispettivamente $\mathbf{A} < 0$ e $\mathbf{A} \leq 0$. In particolare se una matrice $\mathbf{A}$ è definita positiva (semidefinita positiva) allora $\det \mathbf{A} > 0$ ($\det \mathbf{A} \geq 0$).

⁷**Teorema.** Sia $\mathbf{A} \geq 0$ una matrice $p \times p$. Allora per ogni matrice $\mathbf{C} p \times n$, la matrice $\mathbf{C}' \mathbf{A} \mathbf{C}$ è semidefinita positiva. In particolare, se $\mathbf{A} > 0$ e $\mathbf{C}$ è una matrice non singolare (e quindi $p = n$), allora $\mathbf{C}' \mathbf{A} \mathbf{C} > 0$.

2. la *varianza generalizzata* VAR_G : il determinante della matrice di covarianza $|\mathbf{S}|$.

Per entrambe le misure, grandi valori indicano una grande variabilità intorno al punto medio $\bar{x}$ e, al contrario, piccoli valori indicano che i dati sono abbastanza concentrati intorno al punto medio $\bar{x}$. Ciascuna di tali misure, inoltre, riflette aspetti differenti della variabilità dei dati:

La traccia della matrice di covarianze corrisponde alla somma delle varianze della singole variabili $X_1, \dots, X_p$:

$$\text{VAR}_T := \sum_{i=1}^p \sigma_{ii} = \sum_{i=1}^p \sigma_i^2 .$$

La variazione totale presenta problemi interpretativi se le variabili hanno diversa unità di misura ed inoltre ha il difetto di non tenere in conto della correlazione eventualmente presente fra le variabili.

La varianza generalizzata di Wilks è data dal determinante della matrice di covarianze:

$$|\mathbf{S}| := \det \mathbf{S}$$

ed è una misura di variabilità che tiene conto della correlazione fra le variabili. In particolare la varianza generalizzata risulta uguale a zero se il rango⁸ della matrice di covarianza è minore di p e tale caso si manifesta quando almeno una variabile assume sempre il medesimo valore nelle n unità statistiche (cioè risulta costante), oppure almeno una variabile è perfettamente correlata con un'altra, oppure se una variabile è combinazione lineare di altre variabili. In queste circostanze, la variabilità non è a p dimensioni ma a $(p - 1)$, $(p - 2)$, ... dimensioni.

Il valore massimo che può assumere la varianza generalizzata è uguale al prodotto delle varianze delle p variabili, per cui si può definire la seguente misura relativa di variabilità multi-dimensionale:

$$V_R := \frac{|\mathbf{S}|}{\prod_{i=1}^p \sigma_i^2} .$$

Infine si può dimostrare che V_R è uguale al determinante della matrice di correlazione:

$$V_R = |\mathbf{R}| .$$

Ovviamente V_R risulta uguale a zero nelle medesime circostanze in cui vale zero la varianza generalizzata ed assume valore massimo unitario quando le variabili sono fra loro incorrelate a due a due. Da ciò si deduce che, per valori fissi delle varianze delle p variabili, la varianza generalizzata aumenta quando diminuisce la correlazione fra le variabili.

⁸Sia $\mathbf{A}$ una matrice di ordine $n \times p$. Si definisce *rango* di $\mathbf{A}$, denotato con $r(\mathbf{A})$, il massimo numero di righe (colonne) linearmente indipendenti di $\mathbf{A}$.

Alcune proprietà del rango di una matrice sono qui di seguito riassunte: Sia $\mathbf{A}$ una matrice di ordine $n \times p$. Allora si ha:

1. $0 \leq r(\mathbf{A}) \leq \min(n, p)$;
2. $r(\mathbf{A}) = r(\mathbf{A}')$;
3. $r(\mathbf{A}'\mathbf{A}) = r(\mathbf{A}\mathbf{A}') = r(\mathbf{A})$;
4. $r(\mathbf{A} + \mathbf{B}) \leq r(\mathbf{A}) + r(\mathbf{B})$ per ogni matrice $\mathbf{B}(n \times p)$;
5. $r(\mathbf{A}\mathbf{B}) \leq \min\{r(\mathbf{A}), r(\mathbf{B})\}$ per ogni matrice $\mathbf{B}(p \times q)$;
6. $r(\mathbf{B}\mathbf{A}\mathbf{C}) = r(\mathbf{A})$ per tutte le matrici non singolari $\mathbf{B}(n \times n)$ e $\mathbf{C}(p \times p)$;
7. se $n = p$ allora $r(\mathbf{A}) = n$ se e solo se $\mathbf{A}$ è non singolare.

1.4 Codifica dei dati

In molti casi, prima di applicare tecniche di analisi dei dati si ricorre ad una ricodifica dei dati rilevati per una o più variabili.

Ad esempio, con riferimento ad applicazioni in ambito bancario, una variabile X del tipo "Saldo attuale del conto

- A_1 : $X < 0$ euro
- A_2 : $0 \leq X < 5000$ euro
- A_3 : $5000 < X \leq 10000$ euro
- A_4 : $X > 10000$ euro

è una variabile categoriale ordinale a 4 modalità che possono essere codificate come: 1, 2, 3, 4. Si noti che la variabile originaria è quantitativa ed è stata ricodificata come categoriale ordinale. Ovviamente il trattamento delle variabili dipende dalla loro natura e non dalla codifica; non ha quindi senso fare la media di X_4 , anche se viene rappresentata in forma numerica.

Un importante esempio di codifica è la *codifica disgiuntiva completa*. Consideriamo un carattere qualitativo che presenti r modalità (usualmente è $r \leq n$, ma solitamente $r \ll n$). La matrice di dati può prevedere che la corrispondente colonna fornisca i valori codificati di tali modalità. In alternativa le osservazioni possono essere codificate in maniera disgiuntiva sostituendo alla colonna in esame r colonne relative a *variabili indicatrici* X_v , $v = 1, \dots, r$ ciascuna delle quali corrisponde ad una modalità di X e tali che:

$$x_{hi} = \begin{cases} 1 & \text{se la modalità } i \text{ è presente nell}'h\text{-esima unità} \\ 0 & \text{se la modalità } i \text{ è assente nell}'h\text{-esima unità.} \end{cases}$$

Consideriamo il saldo del conto corrente di un insieme di clienti di un istituto bancario:

Unità	Codifica	Codifica disgiuntiva completa			
		A_1	A_2	A_3	A_4
1	A_3	0	0	1	0
2	A_2	0	1	0	0
3	A_3	0	0	1	0
4	A_1	1	0	0	0
5	A_3	0	0	1	0
6	A_4	0	0	0	1
7	A_3	0	0	1	0
8	A_3	0	0	1	0
9	A_2	0	1	0	0
10	A_3	0	0	1	0
11	A_2	0	1	0	0
12	A_3	0	0	1	0
13	A_1	1	0	0	0
14	A_2	0	1	0	0

Questa codificazione viene chiamata *disgiuntiva*. Essa può essere proposta anche per i caratteri quantitativi. Se la variabile è discreta ed il numero r di modalità che essa può assumere è piccolo, si procede come nel caso qualitativo. Se la variabile è continua oppure discreta ma presenta un grande numero di modalità, si suddivide l'intervallo in r classi a ciascuna delle quali si fa corrispondere una variabile indicatrice che risulta uguale a 1 se la generica h -esima unità presenta un valore compreso in tale classe e 0 altrimenti.

Se la codificazione disgiuntiva viene effettuata per tutti i p caratteri, la matrice dei dati originaria viene sostituita da una matrice $n \times P$ dove $P = \sum_{j=1}^p r_j$, dove r_j è il numero di modalità (o classi) della j -esima variabile.

1.5 Matrici a tre vie

Un'importante generalizzazione delle matrici di dati è costituita dalle cosiddette *matrici a tre vie*. In termini generali, esse sono caratterizzate dal fatto che ciascun dato elementare che vi compare x_{hij} presenta tre indici che corrispondono ad una classificazione dello stesso in base a tre criteri:

$$\underset{n \times p \times q}{\mathbf{X}} = (x_{hij})$$

Esempi tipici sono i seguenti:

1. una matrice di dati del tipo "unità $\times$ variabili $\times$ occasioni" (ove il termine "occasioni" può indicare: diversi tempi, differenti situazioni sperimentali, luoghi diversi, etc.);
2. una successione di matrici di indici di prossimità del tipo "oggetti $\times$ oggetti" rilevate in differenti "occasioni" (ad esempio, la similarità fra n prodotti di marche differenti, valutata dai consumatori in anni successivi, oppure in regioni diverse).

Un importante caso delle matrici a tre vie è costituito da quelle del tipo: "unità $\times$ variabili $\times$ tempi". In questo caso sulle stesse unità vengono rilevate le stesse variabili in più tempi successivi. La matrice di dati a tre vie può allora essere pensata come una successione temporale di nq matrici dei dati $\underset{\sim}{\mathbf{X}}_t = (x_{hit})$ del tipo consueto "unità $\times$ variabili". Alcuni esempi sono i seguenti:

1. la rilevazione in n aziende di p variabili desunte dai dati di bilancio in q anni successivi;
2. la misurazione in n pazienti di p sintomi, esprimibili quantitativamente, in q giorni;
3. i valori di un insieme di indicatori demografici, economici e sociali per ciascuna delle provincie italiane, rilevati dai Censimenti Istat nei vari anni.

I dati di questo tipo vengono chiamati *longitudinali* e la loro caratteristica essenziale è che le medesime unità statistiche sono "misurate" ripetutamente nel corso del tempo. Le analisi longitudinali estendono gli studi basati sulla matrice dei dati unità $\times$ variabili (spesso chiamati dati *sezionati* o di tipo *cross-section*), in cui per ogni unità statistica si dispone di un solo vettore di dati riferito ad un istante (o ad un intervallo temporale) prefissato. L'interesse delle matrici di dati a tre vie, in cui una delle dimensioni è il tempo, risiede nel fatto che esse permettono di misurare il cambiamento, con riguardo alle singole unità ed ai differenti fenomeni.

Un esempio economico di matrici a tre vie è fornito dalle azioni quotate in Borsa (le unità statistiche), dalle seguenti variabili: prezzo ufficiale, prezzo di riferimento (cioè calcolato sull'ultimo 10% delle contrattazioni giornaliere), numero di contratti effettuati, controvalore degli scambi, considerando più giorni successivi di Borsa aperta.

Se le unità statistiche delle matrici a tre vie formano un campione estratto da un certo universo, esse costituiscono un *panel*. Le indagini su dati provenienti da un *panel* costituiscono pertanto un caso particolare delle analisi longitudinali.

1.6 Matrici di preferenza

Consideriamo il caso in cui ad ogni unità statistica si richiede di classificare in ordine di preferenza p elementi di un insieme, quali ad esempio le p diverse marche di un prodotto, p spettacoli televisivi, p diversi sport, etc. In tal caso x_{hi} costituisce il grado di preferenza espresso dalla h -esima unità statistica con riferimento all' i -esima variabile. La matrice $\tilde{\mathbf{X}}$ che si ottiene viene chiamata *matrice di preferenze*. Le variabili $X_1, \dots, X_p$ sono variabili categoriali ordinali; lo studio della dipendenza fra due unità statistiche di tale matrice viene effettuato mediante *indici di cograduazione*.

1.6.1 La cograduazione (caso senza unità alla pari)

Consideriamo prima il caso in cui in ognuna delle due graduatorie non vi siano unità alla pari, cioè non vi siano unità che occupino lo stesso posto in una graduatoria (per esempio due distinte unità non possono occupare il terzo posto). La tabella seguente riporta le valutazioni fornite da due consumatori X e Y in merito a sette marche diverse di detersivi per piatti (da "A: migliore" a "G: peggiore"):

<i>unità</i>	X	Y
ω_1	C	D
ω_2	G	E
ω_3	D	F
ω_4	B	A
ω_5	E	G
ω_6	F	C
ω_7	A	B

Ci poniamo il problema di valutare la concordanza fra i giudizi espressi per le due variabili in esame. La matrice sopra riportata è un semplice esempio di *matrice di punteggi* dove il generico elemento x_{hi} rappresenta il punteggio attribuito dalla h -esima unità statistica per la i -esima variabile.

Siano X e Y due variabili aventi ciascuna n modalità. Ci si pone il problema di misurare la relazione fra X e Y analizzando i rispettivi *posti d'ordine* o *gradi* o *ranghi* (*rank*). Si assegna pertanto grado "1" alla modalità inferiore, grado "2" a quella immediatamente successiva, e così via grado n alla modalità maggiore. La metodologia statistica che studia le relazioni tra i

posti d'ordine delle modalità di caratteri ordinali o quantitativi si chiama *cograduazione* (*rank correlation*).

Assumiamo preliminarmente che le n modalità di X e di Y siano distinte, così che non vi siano *ex-aequo*. Consideriamo una funzione $g(\cdot)$ che associa a ciascuna modalità il suo rango, cioè $g : \mathcal{X} \rightarrow \{1, 2, \dots, n\}$. Indichiamo con $x_{(1)}$ la più piccola modalità assunta da X in Ω , con $x_{(2)}$ quella immediatamente successiva e così via:

$$x_{(1)} < x_{(2)} < \dots < x_{(n)}$$

e quindi scriveremo:

$$g_X(x_i) = j \quad i = 1, \dots, n .$$

se l'osservazione x_i ha posto j nella graduatoria in esame; in altre parole $g(x_i)$ fornisce il posto nella graduatoria delle modalità del valore osservato x_i . Con riferimento all'esempio precedente, il prodotto 1 ha ottenuto una valutazione pari a C da X (corrispondente al quinto posto in graduatoria) ed una valutazione pari a D da Y (pari al quarto posto in graduatoria):

unità	consumatore X		consumatore Y	
	valutazione	graduatoria	valutazione	graduatoria
ω_1	C	5	D	4
ω_2	G	1	E	3
ω_3	D	4	F	2
ω_4	B	6	A	7
ω_5	E	3	G	1
ω_6	F	2	C	5
ω_7	A	7	B	6

Le quantità $g_X(x_i)$ vengono chiamati "ranghi" o "gradi". Se sostituiamo alle modalità x_i e y_i i rispettivi ranghi (in ordine crescente), lo studio della relazione fra X e Y può essere ricondotto allo studio della relazione fra i vettori:

$$\begin{pmatrix} g_X(x_1) \\ g_X(x_2) \\ \vdots \\ g_X(x_n) \end{pmatrix} \quad \text{e} \quad \begin{pmatrix} g_Y(y_1) \\ g_Y(y_2) \\ \vdots \\ g_Y(y_n) \end{pmatrix}$$

ciascuno dei quali contiene i numeri da 1 a n in un ordine che dipende dalle modalità dei fenomeni. Introduciamo ora i due casi limite.

Definizione 1.4 Diremo che fra i due fenomeni X e Y vi è *perfetta cograduazione* se risulta:

$$g_X(x_i) = g_Y(y_i) \quad i = 1, \dots, n .$$

cioè quando le unità statistiche, con riferimento ai due caratteri X e Y , presentano lo stesso ordinamento nelle due graduatorie (una certa unità statistica assume il primo posto nella graduatoria di X ed il primo posto nella graduatoria di Y ; un'altra unità statistica assume il secondo posto nella graduatoria sia di X che di Y , e così via).

Diremo che fra i due fenomeni X e Y vi è *perfetta contrograduazione* se risulta

$$g_X(x_i) = n + 1 - g_Y(y_i) \quad i = 1, \dots, n.$$

cioè quando le unità statistiche, con riferimento ai due caratteri X e Y , presentano ordinamento opposto nelle due graduatorie (una certa unità statistica assume il primo posto nella graduatoria di X e l'ultimo posto nella graduatoria di Y ; un'altra unità statistica assume il secondo posto nella graduatoria di X ed il penultimo in quella di Y , e così via)).

La relazione fra i due fenomeni può pertanto essere misurata facendo riferimento alle rispettive graduatorie, cioè costruendo indici che basati sulle graduatorie.

Nota 1.5 Con riferimento ad una matrice di dati di dimensioni $n \times p$, se essa contiene fenomeni sia quantitativi che ordinali, la trasformazione degli stessi in termini delle rispettive graduatorie rappresenta inoltre un criterio per il trattamento simultaneo dei caratteri. Ovviamente per i fenomeni quantitativi ciò comporta una perdita di informazione dovuta al passaggio da una scala di misurazione quantitativa ad una scala di misurazione ordinale.

Si considerino due coppie di gradi riferite alle generiche unità i e j :

$$\{g_X(x_i), g_Y(y_i)\} \quad \text{e} \quad \{g_X(x_j), g_Y(y_j)\}.$$

Tali coppie si dicono *concordanti* se l'unità che presenta grado maggiore su X presenta grado maggiore su Y , cioè se risulta:

$$g_X(x_i) < g_X(x_j) \quad \text{e} \quad g_Y(y_i) < g_Y(y_j),$$

oppure

$$g_X(x_i) > g_X(x_j) \quad \text{e} \quad g_Y(y_i) > g_Y(y_j).$$

Due coppie di gradi si dicono *discordanti* se l'unità che presenta grado maggiore su X presenta grado minore su Y , cioè se risulta:

$$g_X(x_i) < g_X(x_j) \quad \text{e} \quad g_Y(y_i) > g_Y(y_j),$$

oppure

$$g_X(x_i) > g_X(x_j) \quad \text{e} \quad g_Y(y_i) < g_Y(y_j).$$

Ad esempio, con riferimento ai dati presentati all'inizio, le valutazioni per i prodotti 2 e 6 sono concordanti, in quanto $g_X(G) = 1 < g_Y(E) = 3$ e $g_X(F) = 2 < g_Y(C) = 5$; al contrario le valutazioni per i prodotti 2 e 3 sono discordanti in quanto $g_X(G) = 1 < g_Y(E) = 3$ e $g_X(D) = 4 > g_Y(F) = 2$.

Prima di passare alla costruzione di indici di cograduazione, facciamo rilevare che il numero di confronti da effettuare è pari a $n(n-1)/2$.

1.6.2 L'indice di Kendall (caso senza unità alla pari)

Si considerino le graduatorie di n unità statistiche, riferite alle modalità di due caratteri ordinali X e Y . Senza perdere di generalità, possiamo assumere che le unità statistiche siano disposte seguendo l'ordine naturale di X . Ne segue che la misura della concordanza fra X e Y può essere colta in base al modo secondo il quale si succedono i corrispondenti gradi del carattere Y :

- se esiste una perfetta cograduazione, i gradi di Y costituiscono anch'essi una successione ordinata da 1 a n ;
- se esiste una perfetta contrograduazione, allora i gradi di Y si susseguono nell'ordine opposto, cioè da n a 1.

Una misura di cograduazione può allora essere definita in base al numero di inversioni della graduatoria di Y rispetto a quella di X .

Definizione 1.6 Si dice che vi è un'inversione quando, disponendo le unità in modo che i gradi di X seguano l'ordine naturale da 1 a n e confrontando il grado d'un termine generico $g(x_i)$ con il grado di un altro termine che lo segue nella successione, risulta:

$$g(x_i) > g(x_j) \quad j > i; \quad i = 1, \dots, n$$

Riprendiamo l'esempio iniziale e riportiamo i dati nella seguente tabella:

unità	X		Y	
	valutazione	graduatoria	valutazione	graduatoria
ω_2	G	1	E	3
ω_6	F	2	C	5
ω_5	E	3	G	1
ω_3	D	4	F	2
ω_1	C	5	D	4
ω_4	B	6	A	7
ω_7	A	7	B	6

Al fine di misurare la concordanza fra le due graduatorie, con riferimento a due qualunque unità statistiche ω_i e ω_j ($i, j = 1, \dots, n, i \neq j$), consideriamo le seguenti misure di diversità fra X e Y con riferimento alle unità ω_i e ω_j :

$$d_{ij} = \begin{cases} +1 & \text{se } g_X(x_i) < g_X(x_j) \\ -1 & \text{se } g_X(x_i) > g_X(x_j) \end{cases} \quad \text{e} \quad \delta_{ij} = \begin{cases} +1 & \text{se } g_Y(y_i) < g_Y(y_j) \\ -1 & \text{se } g_Y(y_i) > g_Y(y_j) \end{cases} \quad (1.17)$$

e consideriamo l'insieme di tutti i possibili confronti fra coppie di unità statistiche. In altre parole, confrontiamo il rango di ciascuna unità statistica ω_i con quello di tutte le altre unità statistiche ω_j ($j \neq i$). Una misura sintetica di concordanza o discordanza è data dalla somma dei prodotti delle diversità associate a tutte le coppie di unità diverse.

$$\sum_{\substack{i,j \\ i \neq j}} d_{ij} \delta_{ij} = C - D. \quad (1.18)$$

dove C indica il numero di concordanze e D il numero di discordanze. Tale somma risulta positiva in caso di concordanza (prevalgono i termini d_{ij}, δ_{ij} aventi lo stesso segno) e risulta negativa in caso di discordanza (prevalgono i termini d_{ij}, δ_{ij} aventi segno opposto). Ovviamente la (1.18) è un indice assoluto, per avere un indice normalizzato dobbiamo dividere per

il suo valor massimo teoricamente ammissibile. Nel caso di n unità statistiche il numero totale di confronti è $n(n - 1)/2$ e pertanto, nell'ipotesi in cui fra le due variabili vi sia perfetta cograduazione, risulta:

$$\max \sum_{\substack{i,j \\ i \neq j}} d_{ij} \delta_{ij} = n(n - 1)/2 . \quad (1.19)$$

e quindi si costruisce il seguente indice, noto come τ di Kendall:

$$\tau := \frac{\sum_{i < j} d_{ij} \delta_{ij}}{n(n - 1)/2} = \frac{2(C - D)}{n(n - 1)} . \quad (1.20)$$

Ovviamente si tratta di un indice normalizzato che assume valore fra -1 (nel caso in cui $d_{ij} \delta_{ij} = -1$ per tutte le coppie $i < j$) e $+1$ (nel caso in cui $d_{ij} \delta_{ij} = +1$ per tutte le coppie $i < j$). In particolare risulta:

- $\tau = -1$ quando vi è perfetta contrograduazione fra i due fenomeni in esame;
- $\tau = 0$ quando i due fenomeni in esame non presentano né concordanza né discordanza;
- $\tau = +1$ quando vi è perfetta graduazione fra i due fenomeni in esame.

Inoltre l'indice di cograduazione di Kendall è simmetrico: $\tau(X, Y) = \tau(Y, X)$.

Proprietà dell'indice τ di Kendall. L'indice τ di Kendall è invariante per trasformazioni monotone crescenti delle graduatorie dei due caratteri. L'assegnazione dei numeri da 1 a n alle modalità ordinali di un fenomeno è convenzionale. In alternativa, si potrebbero assegnare all'uno o all'altro fenomeno gradi crescenti (o ad entrambi), ad esempio secondo una funzione quadratica: $g(x_{(i)}) = i^2$, oppure con progressione geometrica: $g(x_{(i)}) = a^i$ (per $a > 1$ e $i = 1, \dots, n$). L'indice τ non risulta affatto influenzato da tali assegnazioni in quanto, in base alla (1.17), il confronto nella definizione (1.20) si basa unicamente sulle relazioni di maggiore o minore fra le modalità, che rimangono immutate a seguito di una qualunque trasformazione monotona crescente dei valori numerici assunti come gradi.

Con riferimento al caso in esame, essendo $n = 7$ si hanno 21 confronti, che vengono riportati nella seguente tabella:

ω_i	ω_j	d_{ij}	δ_{ij}	$d_{ij} \delta_{ij}$	ω_i	ω_j	d_{ij}	δ_{ij}	$d_{ij} \delta_{ij}$
1	2	-1	-1	1	3	4	1	1	1
1	3	-1	-1	1	3	5	-1	-1	1
1	4	1	1	1	3	6	-1	1	-1
1	5	-1	-1	1	3	7	1	1	1
1	6	-1	1	-1	4	5	-1	-1	1
1	7	1	1	1	4	6	-1	-1	1
2	3	1	-1	-1	4	7	1	-1	-1
2	4	1	1	1	5	6	-1	1	-1
2	5	1	-1	-1	5	7	1	1	1
2	6	1	1	1	6	7	1	1	1
2	7	1	1	1					

e pertanto segue:

$$\tau = \frac{9}{21} = 0.4286.$$

1.6.3 Indice di Spearman (caso senza unità alla pari)

Un altro indice di cograduazione è il seguente. Consideriamo le seguenti misure di diversità fra X e Y con riferimento alle unità ω_i e ω_j

$$d_{ij} := g_X(x_j) - g_X(x_i) \quad \delta_{ij} := g_Y(y_j) - g_Y(y_i). \quad (1.21)$$

e quindi la seguente quantità:

$$\sum_i \sum_j d_{ij} \delta_{ij} = \sum_i \sum_j [g(x_j) - g(x_i)][g(y_j) - g(y_i)], \quad (1.22)$$

che, come nel caso precedente, risulta positiva in caso di concordanza (prevalgono i termini d_{ij}, δ_{ij} aventi lo stesso segno) e risulta negativa in caso di discordanza (prevalgono i termini d_{ij}, δ_{ij} aventi segno opposto). Ovviamente si tratta di un indice assoluto, per avere un indice normalizzato dobbiamo dividere per il suo valor massimo teoricamente ammissibile che possiamo ottenere tenendo conto della disuguaglianza di Cauchy⁹:

$$\left(\sum_i \sum_j d_{ij} \delta_{ij} \right)^2 \leq \left(\sum_i \sum_j d_{ij}^2 \right) \left(\sum_i \sum_j \delta_{ij}^2 \right).$$

Possiamo quindi costruire il seguente indice, noto come *indice di cograduazione di Spearman*, usualmente denotato con la lettera ρ :

$$\rho := \frac{\sum_i \sum_j d_{ij} \delta_{ij}}{\sqrt{\left(\sum_i \sum_j d_{ij}^2 \right) \left(\sum_i \sum_j \delta_{ij}^2 \right)}} = \frac{\sum_i \sum_j [g(x_j) - g(x_i)][g(y_j) - g(y_i)]}{\sqrt{\sum_i \sum_j [g(x_j) - g(x_i)]^2 \sum_i \sum_j [g(y_j) - g(y_i)]^2}}. \quad (1.24)$$

In particolare risulta:

⁹Siano $(a_1, \dots, a_n)$ e $(b_1, \dots, b_n)$ due vettori di $\mathbb{R}^n$. Si ha allora:

$$\left(\sum_{i=1}^n a_i b_i \right)^2 \leq \left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right)$$

Infatti essendo:

$$\begin{aligned} 0 &\leq \frac{1}{2} \sum_i \sum_j (a_i b_j - a_j b_i)^2 = \frac{1}{2} \sum_i \sum_j a_i^2 b_j^2 + \frac{1}{2} \sum_i \sum_j a_j^2 b_i^2 - \sum_i \sum_j a_i b_j a_j b_i \\ &= \frac{1}{2} \sum_i a_i^2 \sum_j b_j^2 + \frac{1}{2} \sum_j a_j^2 \sum_i b_i^2 - \left(\sum_i a_i b_i \right) \left(\sum_j a_j b_j \right) \\ &= \sum_i a_i^2 \sum_i b_i^2 - \left(\sum_i a_i b_i \right)^2 \end{aligned}$$

da cui segue la tesi:

$$\left(\sum_i a_i b_i \right)^2 \leq \left(\sum_i a_i^2 \right) \left(\sum_i b_i^2 \right). \quad (1.23)$$

□

- $\rho = -1$ quando vi è perfetta contrograduazione fra i due fenomeni in esame;
- $\rho = 0$ quando i due fenomeni in esame non presentano né concordanza né discordanza;
- $\rho = +1$ quando vi è perfetta graduazione fra i due fenomeni in esame.

Inoltre, ovviamente, in base alla (1.24), l'indice di cograduazione di Spearman è simmetrico, cioè $\rho(X, Y) = \rho(Y, X)$.

Ai fini pratici, sviluppando numeratore e denominatore della (1.24), si può dimostrare che il coefficiente di Spearman può essere scritto nella forma seguente:

$$\rho = 1 - \frac{6 \sum_{i=1}^n [g(x_i) - g(y_i)]^2}{n(n^2 - 1)} \quad (1.25)$$

che risulta più agevole nei casi pratici.

Nota 1.7 Si può dimostrare che il coefficiente di cograduazione di Spearman fra due graduatorie (senza unità alla pari) è numericamente uguale al coefficiente di correlazione lineare fra due distribuzioni aventi come modalità corrispondenti i posti occupati nelle due graduatorie da una stessa unità in quanto si dimostra che risulta:

$$\rho = \frac{\sum_i [g(x_i) - \frac{n+1}{2}][g(y_i) - \frac{n+1}{2}]}{\sqrt{\sum_i [g(x_i) - \frac{n+1}{2}]^2 \sum_i [g(y_i) - \frac{n+1}{2}]^2}} \quad (1.26)$$

e la quantità $\frac{n+1}{2}$ si può interpretare come la media aritmetica di $g(x_1), \dots, g(x_n)$ e di $g(y_1), \dots, g(y_n)$ che sono permutazioni dei numeri $1, \dots, n$.

Proprietà dell'indice ρ di Spearman. In base alla definizione (1.24), si vede immediatamente che l'indice di cograduazione ρ di Spearman è invariante per trasformazioni monotone crescenti delle modalità di uno o di entrambi i fenomeni. Infatti, qualunque modificazione delle modalità originarie di un carattere che non alteri i posti d'origine, non altera la graduatoria che conduce al calcolo dell'indice stesso. Si noti che questa proprietà di invarianza è più debole di quella dell'indice τ in quanto non è lecito trasformare monotonicamente i gradi, bensì solo le modalità originarie, sulle quali si determinano sempre le graduatorie, del tipo consueto da 1 a n . Inoltre, in base alla (1.26), poichè ρ è uguale al coefficiente di correlazione calcolato sulle distribuzioni dei posti d'ordine, esso è invariante solo per trasformazioni lineari crescenti delle graduatorie.

Confronto fra l'indice τ di Kendall e l'indice ρ di Spearman. Nell'indice ρ l'invarianza è raggiunta imponendo che la misurazione assegni valore 1 alla modalità minore di tutte le altre, valore 2 a quella immediatamente successiva, etc. La scala di misurazione è rigidamente imposta ed è possibile trasformarla solo linearmente: non si può più parlare allora, legittimamente, di scala ordinale. L'assunzione dei primi n numeri naturali come misure delle n modalità di ciascun vettore implica dunque l'ipotesi addizionale, usualmente non enunciata esplicitamente, di *uguale distanza fra i gradi* che traducono numericamente l'ordinamento delle modalità. Nel caso di fenomeni qualitativi con modalità ordinabili, ciò è in contrasto con la scala di misurazione, che ammette soltanto le relazioni di "uguale", "maggiore" e "minore". Nel caso di fenomeni (quantitativi) misurabili su scala d'intervalli o su scala di rapporti, la trasformazione

dei valori originari in gradi secondo la regola sopra descritta comporta l'arbitraria riduzione a 1 di tutte le differenze effettive fra valori consecutivi. Il ricorso all'indice ρ non significa dunque, in realtà, valutare semplicemente la concordanza, tra gli ordinamenti delle modalità dei due vettori componenti la serie doppia, come spesso si dice, bensì quantificare la relazione fra misurazioni (i gradi) i cui valori costituiscono una progressione aritmetica di ragione 1 e con primo termine anch'esso uguale a 1.

Nell'indice τ di Kendall, invece, l'impiego delle graduatorie ha l'unico scopo di semplificare i calcoli. La relazione fra il valore numerico di ρ e quello di τ calcolati sulle medesime graduatorie non è affatto semplice, se si escludono i casi banali di perfetta cograduazione e perfetta contrograduazione in cui entrambi assumono valore rispettivamente uguale a +1 e -1. Con larga approssimazione, si può dire che il rapporto fra ρ e τ si avvicina a 3/2 per n grande e per valori degli indici non troppo prossimi agli estremi unitari

Relazione fra indici di cograduazione e correlazione. Gli indici di cograduazione, che sono applicabili a caratteri unicamente ordinali, sono calcolabili anche in presenza di fenomeni quantitativi, per i quali è sempre determinabile la corrispondente graduatoria. In quest'ultima circostanza, per lo studio delle relazioni fra due variabili, gli indici di cograduazione rappresentano un criterio alternativo rispetto al consueto coefficiente di correlazione lineare r introdotto in (1.13). Risulta importante mettere quindi in luce le differenze fra i vari indici.

In primo luogo, si può dimostrare che gli indici τ , ρ e r sono tutti riconducibili ad una stessa formula generale:

$$\gamma := \frac{\sum_i \sum_j \alpha_{ij} \beta_{ij}}{\sqrt{(\sum_i \sum_j \alpha_{ij}^2)(\sum_i \sum_j \beta_{ij}^2)}} \quad \text{con } \alpha_i = \beta_i = 0 \quad (1.27)$$

dove α_{ij} rappresenta un punteggio di un certo tipo, assegnato alla coppia di unità ω_i e ω_j con riferimento al carattere X e β_{ij} ha un significato analogo, con riferimento al carattere Y . Si assume che $\alpha_{ij} = -\alpha_{ji}$ e $\beta_{ij} = -\beta_{ji}$.

1. L'indice τ di Kendall si ottiene dalla (1.27) assegnando i punteggi nella maniera seguente:

$$\begin{aligned} a_{ij} &= +1 && \text{se } g(x_i) < g(x_j) \\ a_{ij} &= -1 && \text{se } g(x_i) > g(x_j), \end{aligned}$$

e similmente per b_{ij} con riferimento ai gradi di Y .

2. L'indice ρ di Spearman si ricava ponendo:

$$\begin{aligned} a_{ij} &= g(x_i) - g(x_j) \\ b_{ij} &= g(y_i) - g(y_j). \end{aligned}$$

3. Il coefficiente di correlazione lineare r si ricava ponendo:

$$\begin{aligned} a_{ij} &= x_i - x_j \\ b_{ij} &= y_i - y_j. \end{aligned}$$

Pertanto i tre indici citati sono tutti esprimibili nella forma di prodotto interno rapportato al suo massimo. Questo approccio unitario fa emergere il diverso modo in base al quale vengono assegnati i "punteggi" nei tre indici:

1. nell'indice τ si adotta il criterio più semplice, cioè per ciascun fenomeno si assegna un punteggio unitario, positivo o negativo, alla disuguaglianza fra due unità, senza tener conto del numero di posti che separano le rispettive modalità;
2. nell'indice ρ si assegna invece, per ciascun fenomeno, un punteggio uguale al numero di posti che separano le unità nella graduatoria;
3. infine, nel coefficiente di correlazione lineare r , il "punteggio" per ciascuna variabile è uguale alla differenza fra i valori nell'unità ω_i e nell'unità ω_j , e questo presuppone ovviamente che i caratteri siano misurati almeno su scala ad intervalli.

Inoltre, il coefficiente di correlazione misura l'entità del legame lineare fra i valori dei due fenomeni, mentre gli indici di cograduazione misurano la relazione monotona fra di essi.

Dato che la funzione lineare è un caso particolare delle funzioni monotone, l'esistenza di perfetta relazione lineare diretta (inversa) implica perfetta cograduazione (contrograduazione), ma non viceversa. In altre parole si ha:

$$\begin{array}{ll} r = +1 \Rightarrow \tau = \rho = +1 & \text{ma } \tau = \rho = +1 \not\Rightarrow r = +1 \\ r = -1 \Rightarrow \tau = \rho = -1 & \text{ma } \tau = \rho = -1 \not\Rightarrow r = -1 . \end{array}$$

1.6.4 Indice di cograduazione in presenza di gradi ripetuti

Può accadere che due o più unità statistiche possono presentare la medesima modalità di un carattere ordinale, per cui nella rispettiva graduatoria vi sono dei posti a pari merito o "gradi ripetuti" (*tied ranks*).

Il criterio abitualmente adottato consiste nell'assegnare convenzionalmente a ciascuno dei posti a pari merito un grado uguale alla media aritmetica dei gradi che essi avrebbero avuto se distinguibili. Pertanto, se in una graduatoria vi sono alla pari le prime n_1 unità, le successive n_2 unità, ..., fino alle ultime n_k unità, possiamo considerare che ognuna delle prime n_1 unità occupi il posto:

$$\frac{1 + 2 + \dots + n_1}{n_1} = \frac{n_1(n_1 + 1)}{2n_1} = \frac{n_1 + 1}{2}$$

che ognuna delle n_2 unità successive occupi il posto:

$$\frac{(n_1 + 1) + (n_1 + 2) + \dots + (n_1 + n_2)}{n_2} = n_1 + \frac{n_2 + 1}{2}$$

e così via fino all'ultimo gruppo in cui ognuna delle n_k unità occupa il posto:

$$n_1 + n_2 + \dots + n_{k-1} + \frac{n_k + 1}{2} .$$

Ad esempio, se nella successione ordinata la terza e la quarta unità presentano modalità identiche del carattere X , si assegna ad entrambe grado 3.5; e le unità ω_5, ω_6 e ω_7 presentano la stessa modalità si attribuisce ad ognuna di esse grado 6, etc. Questo implica una riduzione della variabilità all'interno di ogni graduatoria, poichè vi sono più gradi uguali fra di loro.

Riprendiamo l'esempio precedente supponendo che si possa attribuire la stessa valutazione a marche diverse:

<i>unità</i>	<i>X</i>	<i>Y</i>
ω_1	C	C
ω_2	F	C
ω_3	D	D
ω_4	B	A
ω_5	D	E
ω_6	E	C
ω_7	A	B

ottenendo quindi le seguenti graduatorie:

unità	<i>consumatore X</i>		<i>consumatore Y</i>	
	valutazione	graduatoria	valutazione	graduatoria
ω_1	C	5	C	4
ω_2	F	1	C	4
ω_3	D	3.5	D	2
ω_4	B	6	A	7
ω_5	D	3.5	E	1
ω_6	E	2	C	4
ω_7	A	7	B	6

In tale caso, si può utilizzare l'indice τ di Kendall dove si pone $d_{ij} = 0$ nel caso in cui le unità ω_i e ω_j presentino lo stesso grado. Si tenga inoltre conto che in questo caso τ non può assumere valore $+1$ o -1 .

In alternativa si può utilizzare l'indice ρ di Spearman (1.24) (attenzione a non usare la formula (1.25) che non può essere direttamente applicata in questo caso, ma richiede una rielaborazione che risulta poco pratica, a differenza del caso senza unità alla pari).

Ad esempio, per il calcolo dell'indice di Kendall si ha:

ω_i	ω_j	d_{ij}	δ_{ij}	$d_{ij}\delta_{ij}$
1	2	-1	0	0
1	3	-1	-1	1
1	4	1	1	1
1	5	-1	-1	1
1	6	-1	0	0
1	7	1	1	1
2	3	1	-1	-1
2	4	1	1	1
2	5	1	-1	-1
2	6	1	0	0
2	7	1	1	1
3	4	1	1	1
3	5	0	-1	0
3	6	-1	1	-1
3	7	1	1	1
4	5	-1	-1	1
4	6	-1	-1	1
4	7	1	-1	-1
5	6	-1	1	-1
5	7	1	1	1
6	7	1	1	1

e pertanto segue:

$$\tau = \frac{7}{21} = 0.33.$$

Per ottenere un indice che assuma valori nell'intervallo $[-1, +1]$ anche in presenza di gradi ripetuti, occorre determinare il valore massimo di $\sum_{i < j} d_{ij} \delta_{ij} = C - D$ in tale circostanza, sottraendo dal numero totale dei confronti a coppia il numero di casi in cui il punteggio è nullo. Se per il carattere X vi sono u gradi ripetuti consecutivi, risultano nulli i punteggi di $u(u-1)/2$ coppie, confrontando ciascun termine con quelli consecutivi; analogamente se per il carattere Y vi sono v gradi ripetuti consecutivi, risultano nulli i punteggi di $v(v-1)/2$ coppie, confrontando ciascun termine con quelli consecutivi. Indicato con T_X e con T_Y il numero di coppie di unità statistiche "a pari merito" (cioè che presentano lo stesso grado) rispettivamente su X e su Y , risulta quindi:

$$T_X := \frac{\sum u(u-1)}{2} \quad T_Y := \frac{\sum v(v-1)}{2}.$$

Pertanto una generalizzazione dell'indice di Kendall, in presenza di gradi ripetuti, è la seguente:

$$\tau_b := \frac{C - D}{\sqrt{\left(\frac{n(n-1)}{2} - T_X\right) \left(\frac{n(n-1)}{2} - T_Y\right)}}. \quad (1.28)$$

Con riferimento al caso precedente, si ha: $T_X = 2(2-1)/2 = 1$, $T_Y = 3(3-1)/2 = 3$ e $n(n-1)/2 = 21$; risulta quindi

$$\tau_b = \frac{7}{\sqrt{(21-1)(21-3)}} = \frac{7}{\sqrt{340}} = 0.3796.$$

Si noti che, in assenza di gradi ripetuti si ha $T_X = T_Y = 0$ e quindi la (1.28) si riduce alla (1.20).