Analisi statistica di dati multivariati

La crescita delle capacità dei computer

La disponibilità di tali mezzi a costi contenuti

L'efficienza ed efficacia del software per utilizzarli

Hanno reso strumenti sofisticati alla portata di tutti.

Tanti sfrecciano veloci sul mare dei dati raggiungendo risultati facili e immediati.

Per arrivare però a mete prestigiose è necessaria la professionalità e la pazienza dei sub.

La multidimensionalità

E' infrequente che una indagine statistica esamini un solo fattore. Quasi sempre le elaborazioni tabellari e grafiche sono multidimensionali

Numero di abitazioni nei comuni turistici lombardi. Indagine comunale ICI (anno1999)

Area	*Numero abitazioni 1999*			*% abitazioni secondarie sul totale 1999*	*% di risposta alla scheda di rilevamento*
	Abitazioni principali	*Abitazioni secondarie*	*Totale*		
Area Varesina	6004	2640	8644	*30,54*	*17,40*
Area Alto Lario occidentale, Alpi Lepontine	3189	4074	7263	*56,09*	*44,20*
Area Triangolo Lariano, Lario Intelvese	2601	2470	5071	*48,71*	*36,20*
Area Valchiavenna, Valtellina di Morbegno e Sondrio	2586	6525	9111	*71,62*	*23,80*
Area Valtellina di Tirano e Alta Valtellina	5616	8069	16887	*47,78*	*25,00*
Area Val Brembana, Valle Imagna	4563	15335	22191	*69,10*	*53,00*
Area Val Seriana, Val Seriana Superiore, Valle di Scalve	15652	29083	46285	*62,83*	*63,90*
Area Monte Bronzone, Valle Cavallina, Alto Sebino	9016	9352	18368	*50,91*	*61,90*
Area Val Camonica	3907	7816	11723	*66,67*	*68,20*
Area Lago d'Iseo orientale	5261	2209	7470	*29,57*	*60,00*
Area Garda bresciano, Benaco	5629	7661	13290	*57,64*	*44,50*
Area Val Trompia, Valle Sabbia	929	962	3235	*29,74*	*46,20*
Area Oltrepo' pavese e Lomellina	1668	3160	4828	*65,45*	*26,70*
Area Valsassina, Valvarrone	3960	17904	21864	*81,89*	*44,50*
Area Lario orientale, Valle San Martino	0	0	0	–	*18,20*
Area Pianura Padana	**55170**	**46530**	**101700**	*45,75*	*31,60*
Totale aree comuni turistici	125751	163790	297930	*54,98*	*40,90*

Fonte: nostre elaborazioni su scheda di rilevamento ICI inviate ai comuni turistici lombardi.

Raccogliere e classificare i dati non significa molto: è necessario elaborarli ed analizzarli (con uno scopo).

L'analisi statistica non si limita a raccogliere i mattoni con cui costruire, ma fornisce anche la calce con cui renderli coesi

Osservazioni multidimensionali

Ha varie ragioni:

- **Perché si vuole limitare il rischio di omettere aspetti importanti (si pensi ad un/una paziente ammesso/a in un istituto di cura).**
- **Perché talune caratteristiche si articolano su indicatori diversi (ad esempio la "solvibilità" di un'impresa che richiede un prestito, la percezione di un prodotto da dei consumatorie, le condizioni di vita di una popolazione).**
- **Perché si vogliono estrarre tutti i dati utili da una osservazione che potrebbe non essere ripetibile o del quale è sconsigliabile la ripetizione (l'esplosione di una supernova o il fallimento di un progetto)**

Ne consegue che, in genere, sulle unità non si osserva una sola variabile, ma una m-tupla ordinata di variabili

$$\mathbf{X} = \left(x_1, x_2, \cdots, x_m\right)$$

Esempio

Dove vanno gli studenti: flussi interregionali, A.A. 1989-90.

	Stessa regione							
	Nord		Centro		Sud		Totale	
	numero	%	numero	%	numero	%	numero	%
Nord Ovest	286555	83.6	178692	90.7	253887	74.7	719.124	81.8
Nord Est	18783	5.5	1526	0.8	8378	2.5	28687	3.3
Centro	27308	8.0	4749	2.4	11312	3.3	43369	4.9
Sud	9149	2.7	9396	4.8	38800	11.4	57345	6.5
Totale	929	0.3	2756	1.4	27296	8.0	30981	3.5
Italia	56169	16.4	18427	9.3	85786	25.3	160382	18.2
	342724	100.0	197119	100.0	339663	100.0	879506	100.0

Altre regioni

L'analisi della tabella non è difficile: si vede subito che gli studenti del Sud sono quelli che più si recano a studiare in altre regioni (25.3%) e che più spesso va in altre regioni meridionali (8%).

Esempio/continua

La tabella contiene variabili (regione, residenza, numero, percentuali) la cui trattazione congiunta richiede tecniche adeguate.

Inoltre, la Statistica non si può fermare alla mera descrizione di fatti numerici, ma deve contribuire a rispondere a domande più generali.

Ad esempio, in che modo le percentuali riscontrate nell'anno accademico 1989-90 si applicano agli anni accademici successivi?

Lo studio univariato ha come ipotesi implicita è che si riesca ad avere l'idea di un concetto pluriforme studiandone separatamente gli aspetti singoli.

Questa è una evidente forzatura: sarebbe come parlare delle piramidi d'Egitto descrivendone le facce una ad una.

Modello relazionale dei dati

Tale modello deriva dall'idea matematica di relazione e può essere così formulato: noti gli insiemi (o domini), non necessariamente distinti $\{S_1, S_2, \cdots, S_m\}$

Una relazione "d" si presenta come un insieme di m-tuple ordinate $(d_1, d_2, \cdots, d_m)$

Tali che $d_1 \in S_1,\ d_2 \in S_2, \cdots, d_m \in S_m$

cioè "d" è un sottoinsieme del prodotto cartesiano: $D = S_1 \otimes S_2 \otimes \cdots \otimes S_m$

D è invece lo spazio dei dati cioè l'insieme di tutte le possibili osservazioni.

La rilevazione dei dati -totale o campionaria- consiste nella sistematica annotazione degli elementi di D che si presentano nel processo di acquisizione su "n" unità che porta alla costruzione della relazione.

Quindi $d \subseteq D$

Osservazioni multidimensionali/2

Resta la difficoltà di individuare quali siano le variabile più importanti.

Ecco come Sherlock Holmes risponde a Watson che gli rimprovera le scarse conoscenze di astronomia:

Lei dice che noi giriamo attorno al Sole. Se girassimo attorno alla Luna non cambierebbe nulla per me o per il mio lavoro".

Annota Watson: "...la sua ignoranza era notevole come la sua cultura. Egli non impara nulla che non abbia attinenza coi suoi fini. Quindi, quasi tutte le cognizioni che possedeva avevano per lui una specifica utilità..."

La nostra sensibilità ed esperienza dettano i principi con cui decidiamo sulla preminenza dei vari aspetti di un problema ed è questa "mentalità" che ci frena nel pensarne altre: l'economista prediligerà le variabili economiche, un biologo quelle biologiche, l'urbanista quelle territoriali, la psicologa quelle comportamentali.

Esempio

Lo staff di una organizzazione prevede 6 persone che possono essere donne (D) oppure uomini (U), in possesso di laurea (L) oppure no (N), residenti nel comune dell'organizzazione (R), in comuni vicini (V) o fuori sede (F).

I domini delle tre variabili sono:

$S_1 = \{D, U\};\ S_2 = \{L, N\};\ S_3 = \{R, V, F\}$

In questo caso lo spazio dei dati è

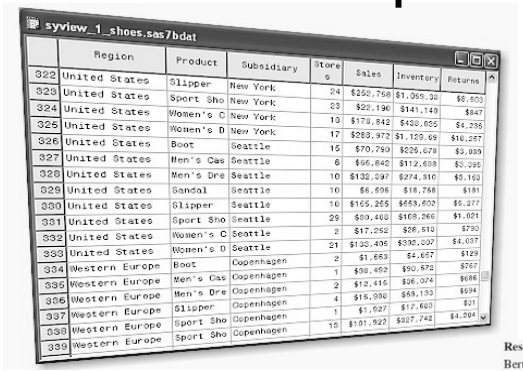
$$S_1 \otimes S_2 \otimes S_3$$

	Maschi		Femmine	
Settore	V.A.	%	V.A.	%
Servizi alle imprese	11	20.4%	21	24.1%
Piccola-media industria	14	25.9%	13	14.9%
Banche, Assicurazioni	7	13.0%	18	20.7%
Media-grande industria	8	14.8%	9	10.3%
Commercio	5	9.3%	6	6.9%
Pubb. Amm.	2	3.7%	6	6.9%
Sanità e servizi pers.	1	1.9%	6	6.9%
Altri servizi	1	1.9%	4	4.6%
Agric. e allev.	1	1.9%	0	0.0%
Altro	4	7.4%	4	4.6%
	54	100.0%	87	

Lo spazio dei dati –teorico- include tutte le configurazioni che si possono ottenere scegliendo una singola modalità da ognuno dei domini delle variabili coinvolte per ognuna delle unità potenziali.

La rilevazione -empirica- include solo quelle modalità delle variabili che si riscontrano nelle unità effettivamente coinvolte.

Esempio di data set


syview_1_shoes.sas7bdat

	Region	Product	Subsidiary	Stores	Sales	Inventory	Returns
322	United States	Slipper	New York	24	$262,758	$1,069,38	$8,653
323	United States	Sport Sho	New York	23	$22,196	$141,140	$847
324	United States	Women's C	New York	10	$178,842	$438,695	$4,286
325	United States	Women's D	New York	17	$283,972	$1,129,69	$10,267
326	United States	Boot	Seattle	15	$70,759	$226,678	$3,099
327	United States	Men's Cas	Seattle	6	$66,842	$112,608	$3,395
328	United States	Men's Dre	Seattle	10	$132,297	$274,910	$3,163
329	United States	Sandal	Seattle	10	$6,696	$18,768	$181
330	United States	Slipper	Seattle	10	$165,255	$653,602	$6,277
331	United States	Sport Sho	Seattle	29	$89,400	$168,266	$1,021
332	United States	Women's C	Seattle	2	$17,252	$26,510	$790
333	United States	Women's D	Seattle	21	$133,406	$392,307	$4,037
334	Western Europe	Boot	Copenhagen	2	$1,663	$4,657	$129
335	Western Europe	Men's Cas	Copenhagen	1	$98,452	$90,672	$787
336	Western Europe	Men's Dre	Copenhagen	2	$12,415	$26,074	$686
337	Western Europe	Slipper	Copenhagen	4	$15,990	$69,133	$594
338	Western Europe	Sport Sho	Copenhagen	1	$1,527	$17,603	$21
339	Western Europe	Sport Sho	Copenhagen	15	$101,923	$327,742	$4,204

Questo è pure un data set, per quanto derivato da elaborazioni su un altro data set

Variable	N	Mean	Std Dev	Minimum	Maximum
Age	20	46.7000000	12.8107685	15.0000000	63.0000000
Height	20	66.9500000	4.1482400	61.0000000	75.0000000
Weight	20	174.6500000	36.3119163	102.0000000	240.0000000
Pulse	20	75.1000000	7.9795792	65.0000000	100.0000000
FastGluc	20	299.1000000	125.6146572	152.0000000	568.0000000
PostGluc	20	355.1000000	125.5153167	206.0000000	625.0000000

Restaurant	Type	Price ($)	Score
Bertucci's	Italian	16	77
Black Angus Steakhouse	Seafood/Steakhouse	24	79
Bonefish Grill	Seafood/Steakhouse	26	85
Bravo! Cucina Italiana	Italian	18	84
Buca di Beppo	Italian	17	81
Bugaboo Creek Steak House	Seafood/Steakhouse	18	77
Carrabba's Italian Grill	Italian	23	86
Charlie Brown's Steakhouse	Seafood/Steakhouse	17	75
Il Fornaio	Italian	28	83
Joe's Crab Shack	Seafood/Steakhouse	15	71
Johnny Carino's Italian	Italian	17	81
Lone Star Steakhouse & Saloon	Seafood/Steakhouse	17	76
LongHorn Steakhouse	Seafood/Steakhouse	19	81
Maggiano's Little Italy	Italian	22	83
McGrath's Fish House	Seafood/Steakhouse	16	81
Olive Garden	Italian	19	81
Outback Steakhouse	Seafood/Steakhouse	20	80
Red Lobster	Seafood/Steakhouse	18	78
Romano's Macaroni Grill	Italian	18	82
The Old Spaghetti Factory	Italian	12	79
Uno Chicago Grill	Italian	16	76

La matrice dei dati

Il risultato di molti percorsi di ricerca è l'individuazione della matrice dei dati, X, che costituirà poi l'oggetto dell'analisi.

In genere è una matrice rettangolare di dimensioni (n x m) , in cui le n righe saranno costituite dalle unità oggetto di indagine, e le m colonne rappresenteranno le variabili che sono stati rilevate per ciascuna unità.

Vettori colonna Per le variabili

Vettori riga per le unità

	Indicatori					
Unità	X_1	X_2	.....	X_j	.....	X_m
I_1	x_{11}	x_{12}	.....	x_{1j}	.....	x_{1m}
I_2	x_{21}	x_{22}	.....	x_{2j}	.....	x_{2m}
.....	.....	.....	.....	.....	.....	.....
I_i	x_{i1}	x_{i2}	.....	x_{ij}	.....	x_{im}
.....	.....	.....	.....	.....	.....	.....
I_n	x_{n1}	x_{n2}	.....	x_{nj}	.....	x_{nm}

Con X_{ij} possiamo indicare quindi il valore che il j-esimo indicatore di base assume nell'unità i-esima

Esempio

DATA APPENDIX

Country	y_{1965}	y_{2000}	n	s_k	s_h
Argentina	11747	17494	1.4	17.6	5.7
Australia	17245	38079	1.8	24.7	11.1
Austria	11263	34566	0.4	26.1	8.3
Burundi	1073	1158	1.9	5.0	0.5
Belgium	13065	35625	0.3	24.0	10.2
Benin	2140	2248	2.5	6.5	1.9
Burkin Faso	1281	1898	1.9	8.4	0.7
Bangladesh	1964	2855	2.5	10.0	3.0
Bolivia	3975	4854	2.4	10.1	5.0
Brazil	4644	10901	2.7	20.6	4.8
Canada	17247	39512	1.7	21.9	10.5
Switzerland	21154	39476	0.8	27.8	6.0
Chile	7103	15097	2.1	16.0	7.3
Cote d'Ivoire	3686	3464	3.7	8.2	2.6
Cameroon	3158	4051	2.4	6.9	3.5
Congo	1191	4394	2.6	23.0	5.1
Colombia	4840	8634	2.8	11.5	6.0
Costa Rica	7158	8753	3.5	14.2	6.3
Denmark	16786	40343	0.5	23.4	10.8
Dominican Rep.	3354	8472	2.9	12.4	5.6
Algeria	4729	9247	3.0	17.9	5.8
Ecuador	4113	5650	3.0	20.1	7.0
Egypt	3006	6813	2.5	7.0	8.1
Spain	7222	26556	0.8	24.5	9.6
Ethiopia	1026	1294	2.5	4.4	1.3
Finland	11850	34318	0.6	26.4	11.4
France	12903	34931	0.8	24.7	9.6
United Kingdom	14819	35009	0.3	18.3	9.8
Ghana	3273	2297	2.8	10.1	4.8
Greece	6178	21747	0.7	25.8	8.6
Guatemala	5090	7656	2.7	8.0	2.9
Hong Kong	4946	36058	2.6	25.8	6.6
Honduras	3214	3678	3.2	12.3	4.0
Indonesia	1354	5811	2.3	12.2	4.7
India	1521	4028	2.3	11.6	5.2
Ireland	9029	38195	1.1	17.9	12.8
Israel	10285	28838	2.8	28.2	9.2
Italy	10328	31764	0.4	24.9	7.3
Jamaica	4343	5249	1.5	19.1	10.2
Jordan	3911	6940	4.7	13.1	9.7
Japan	7386	35982	0.9	31.0	10.6
Kenya	1764	2277	3.4	11.1	2.9
Korea	2685	19552	2.3	27.3	9.7

Variabili

Unità

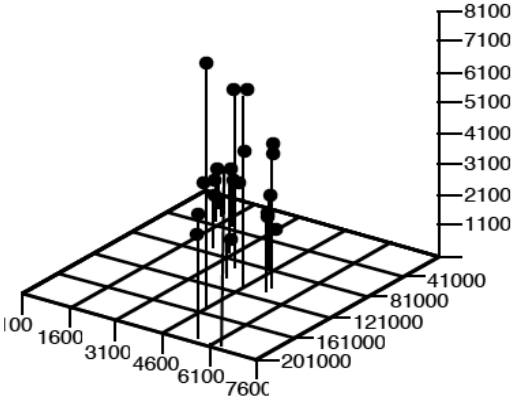
La matrice ha dimensioni nxm (" n per m") da intendersi come "n" righe corrispondenti alle "n" unità ed "m" colonne, una per ogni variabile.

Ogni elemento della matrice è un numero (non necessariamente usato come tale)

Altro esempio

Distribuzione per regioni di studenti, docenti e personale non docente delle università nell'anno accademico 1989/1990.

Da notare la Campania in isolamento rispetto alle altre soprattutto per il numero di unità di personale non docente; la Lombardia è in primo piano per il numero di iscritti.



Regioni	Iscritti	Docenti	Per.Tec.
Piemonte	70592	2901	1543
Lombardia	195049	5464	3927
Trentino A.A.	7179	486	314
Veneto	89390	3593	5664
Friuli V.G.	21329	1591	1051
Liguria	38837	2124	1400
Emilia-Rom.	107597	5185	2406
Toscana	100171	5154	4167
Umbria	20634	1252	1512
Marche	34213	1351	1124
Lazio	194591	6235	5607
Abruzzo	27768	1327	729
Molise	1658	177	112
Campania	151073	4457	7996
Puglia	70986	2380	1804
Basilicata	2548	343	240
Calabria	16031	951	1142
Sicilia	110501	4449	6198
Sardegna	34270	1810	1516

Rappresentazioni grafiche

L'interpretazione dei dati risulta più agevole se il loro contenuto è espresso con grafici che diano un'idea chiara e accurata dei risultati ottenuti

Il messaggio grafico giunge alla mente più rapidamente e per vie diverse rispetto all'informazione numerica o verbale: una nozione di grado più elementare che si assimila con maggiore facilità grazie alla straordinaria capacità della percezione visiva umana.

Provate ad aprire la pagina di un giornale; noterete prima le figure che il testo perché l'immagine richiama i concetti senza bisogno del processo supplementare di ricostruire la parola dalle lettere e dalla posizione nella

"*... Ciò grazie ad una congenita caratteristica della mente umana, per cui questa, mentre deve esercitare uno sforzo più o meno grave per rendersi conto delle variazioni delle grandezze espresse in cifre, quasi intuisce e percepisce senza alcuna riflessione le differenze esistenti tra entità geometriche rappresentatrici delle grandezze di cui trattasi*"
(A. Costanzo, 1969)

Rappresentazioni grafiche/2

Il grafico deve essere redatto in base ad alcune regole che l'esperienza ha mostrato valide:

L'aspetto informativo prevale rispetto a quello descrittivo.

Il grafico contiene dati completi e precisi, richiamati nel testo in cui è collocato

La raffigurazione è improntata a semplicità, chiarezza, efficienza e non ha nulla di superfluo, astratto o misterioso. Se un grafico necessita di troppe spiegazioni per poter essere compreso è meglio non inserirlo. Joan Mirò afferma:
"*Io sento il bisogno di ottenere il massimo dell'intensità con il minimo di mezzi. E' stato questo ad indurmi a dare alla mia pittura un carattere sempre più spoglio*"

Il grafico seduce e informa gli osservatori convincendoli della attendibilità di ciò che raffigura.

Rappresentazioni grafiche/3

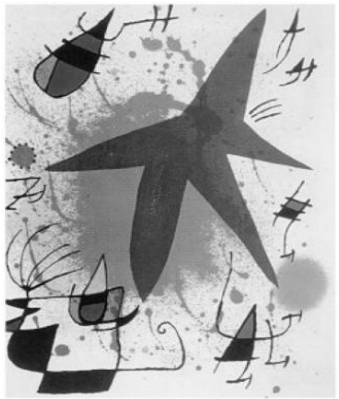
Il grafico non dovrebbe essere più complicato dei dati su cui è basato e dovrebbe essere realizzato correttamente.

Il lettore perde la consapevolezza che tra sé ed il messaggio si interpongono le scelte del disegnatore: ritiene di guardare direttamente ad un fatto e non agli aspetti del fatto che si vuole siano guardati.

Tukey (1977): look at the data and see what it seems to say.

Devono emergere i valori insoliti, tendenze, relazioni.

La trasformazione dei numeri in elementi pittorici è essenziale alla analisi statistica.



Joan Mirò Stella Blu

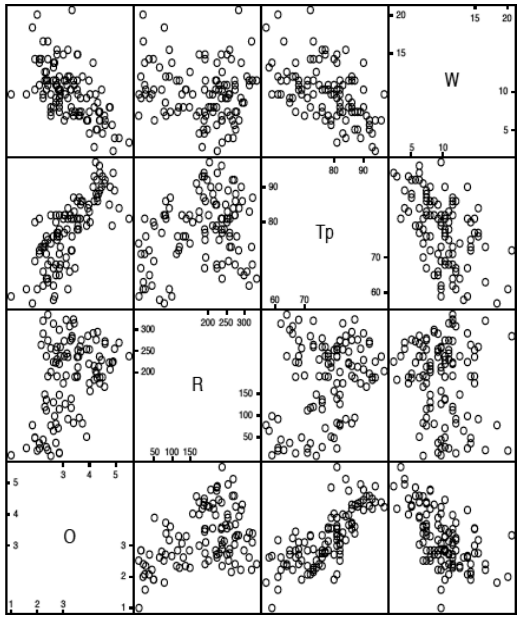
Scatterplot matrices

Consideriamo una matrice di dati con n righe ed m colonne

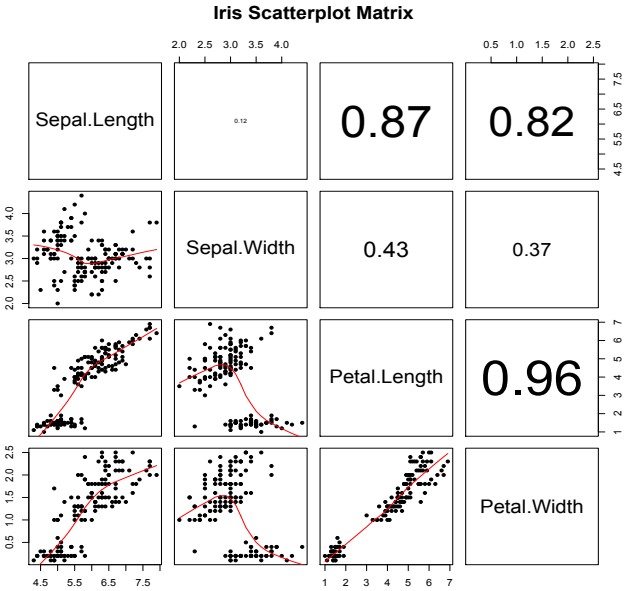
La matrice di scatterplot dispone per righe e per colonne il diagramma di scatter per ciascuna coppia di variabili.

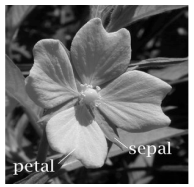
Ne risulta una tabella a doppia entrata dove ogni cella delinea la relazione che sussiste tra le due variabili in riga e colonna.

Poiché la matrice di scatterplo è simmetrica I grafici compaiono due volte, ma con gli assi scambiati.



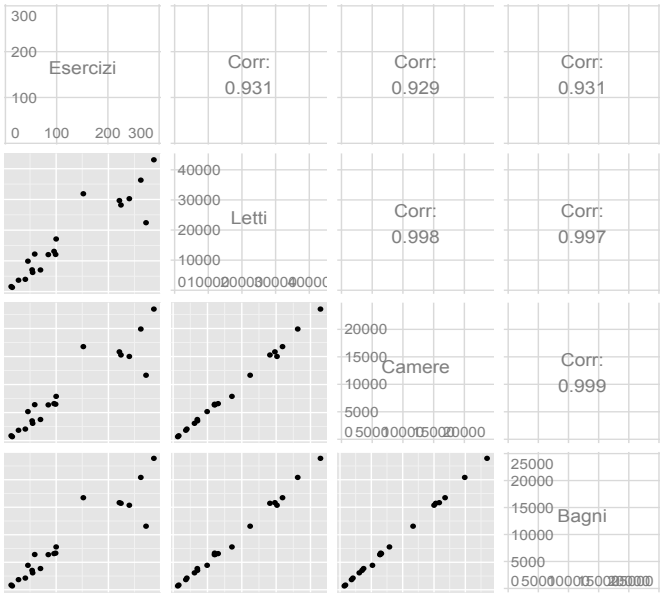
Esempio: Iris dat set





Esempio: hotel 4s

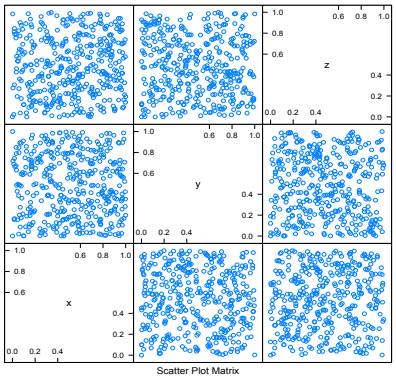
Qui si è usato ggpairs nel pacchetto GGally

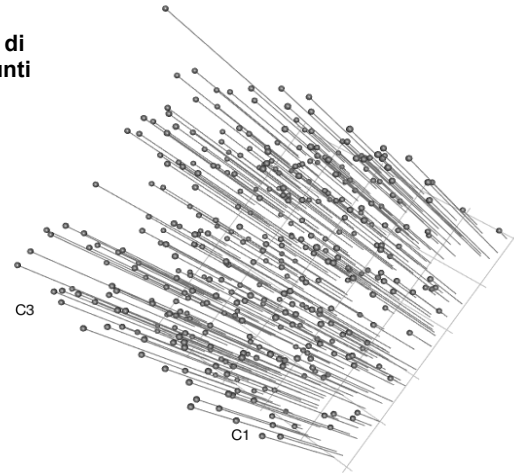


Limiti delle relazioni bivariate

<u>Non è sempre possibile percepire la relazione globale tra le variabili aggregando le relazioni bivariate.</u>

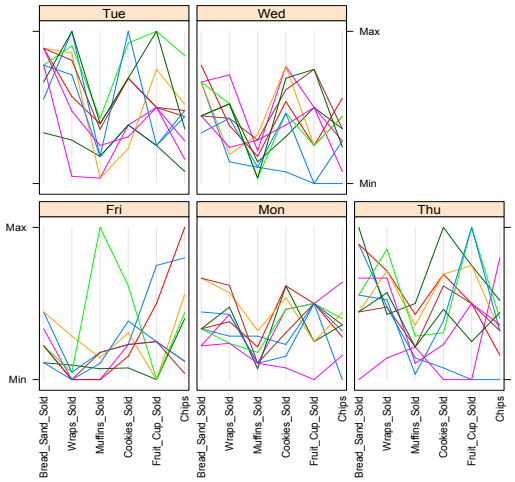
Un esempio è la generazione di numeri pseudo-casuali con un meccanismo inefficiente. Le bivariate non danno idea di alcun legame, ma in 3d si vede che in punti giaggiono su iperpiani a paralleli

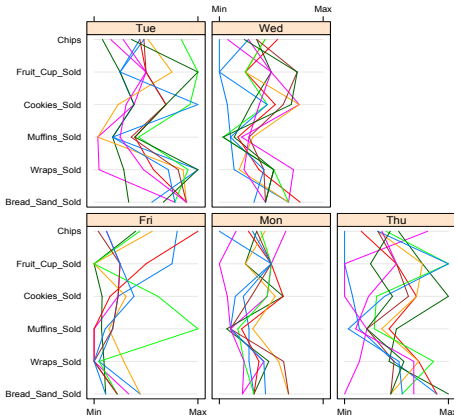




Parallels plot

Draws Parallel Coordinate Plots. (Difficult to describe, see example.)

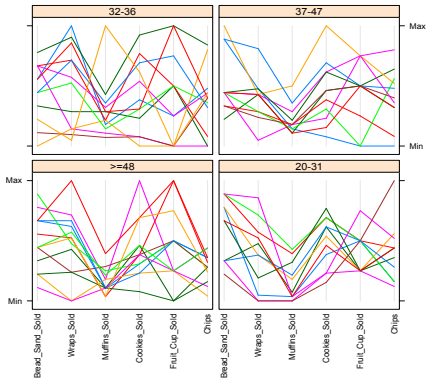


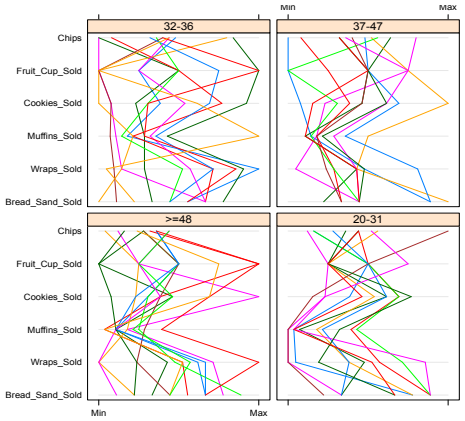


The coordinate axes are represented as parallel lines and a line segment joins each value. Thus, as the sample size increases, the plot tends to become more blurred.

Esempio: cafedata

Molto utile per evidenziare un comportamento particolare in alcune unità e/o alcune variabili





Parallel plot is useful to quickly identify interactions between variables:
Clusters of units with the similar lines across all axes.
Direct relationship between a pair of variables appears in the plot as two axes connected by a series of parallel lines.
Inverse relationship between two variables should be displayed as a series of lines, which cross each other.

Stars (or diamonds)

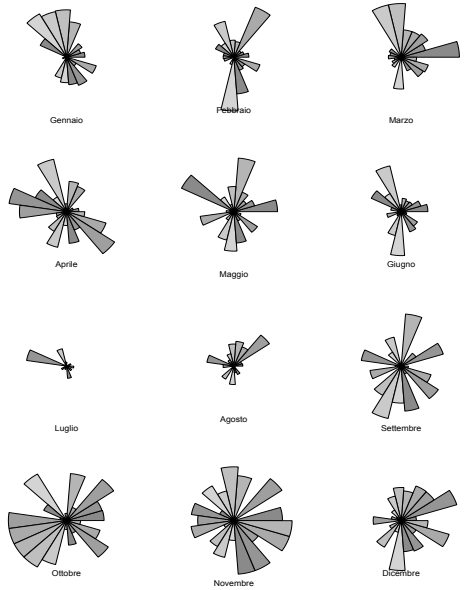
Le variabili costituiscono la lunghezza di segmenti disposti a raggiera e le cui estremità sono unite da linee.

Il risultato è una successione di figure poliedriche, ognuna associata ad una diversa unità che consentono di evidenziare visivamente il loro grado di Similarità.

Quali variabili sono dominanti?

Quale unità hanno struttura simile?

Esistono valori remoti?



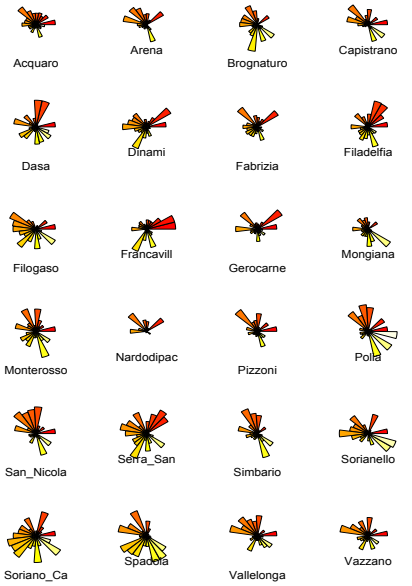
Esempio:

Comuni della calabria.
Diverse unità.

Esempio (continua)

Comuni della calabria.
Diverse variabili.

<u>La percezione delle strutture eventiualmente presentinei dati dipende molto dalla sequenza con cui le variabili compaiono nei disegni</u>



Curve di Andrews

Nel 1972 Andrews propose di rappresentare i dati multivariati utilizzando delle funzioni trigonometriche molto efficaci.

Sia dato il vettore colonna m-dimensionale: $X_i = \begin{bmatrix} Xi_1 \\ X_{i2} \\ \vdots \\ X_{im} \end{bmatrix}$

formato dalle informazioni quantitative rilevate sulla i-esima unità del campione o della popolazione.

Ad esempio per la tabella di dati seguenti relative al ripiano del disavanzo delle Asl per regioni.

Regioni	Disavanzo	Importo max	Importo dato	Fondi propri
Veneto	668.1	471.2	296.6	78.1
Calabria	413.9	292.0	183.7	48.4
Toscana	842.2	594.1	373.9	98.4
Lazio	2602.9	183.6	1155.6	304.3
Liguria	652.0	459.9	289.4	76.2
Puglia	416.5	293.8	184.9	48.7
Sicilia	930.0	655.9	273.1	178.6
Lombardia	718.3	506.6	318.9	83.9
Campania	2034.9	1435.3	903.4	237.9
Marche	386.3	279.5	175.9	46.3
Mdia	966.5	517.2	415.5	120.1
C oef.Var.	0.7338	0.6534	0.7654	0.7085

il dato della Calabria sarebbe espresso dal vettore colonna

$$X_i = \begin{bmatrix} 413.9 \\ 292.0 \\ 183.7 \\ 48.4 \end{bmatrix}$$

Curve di Andrews/2

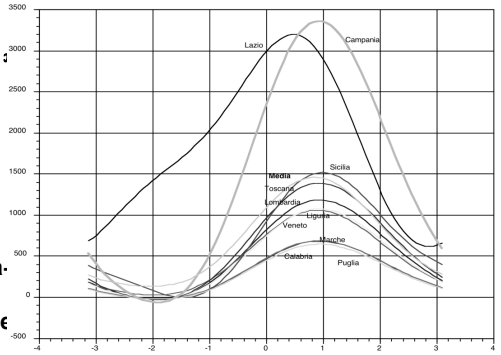
L'idea di Andrews è di esprimere ogni punto nello spazio ad m-dimensioni come una curva armonica (serie di Furier finite) nel piano cartesiano.

$$f_i(t) = \frac{X_{i1}}{\sqrt{2}} + X_{i2}\,sen(t) + X_{i3}\cos(t) + X_{i4}\,sen(2t) + X_{i5}\cos(2t) + \ldots +$$

in cui è coinvolto un numero di addendi per coprire tutte le m variabili

La variabile di rappresentazione t varia in un dominio limitato: $-\pi \le t \le \pi$

il grafico di Andrews si realizza con curve in un unico piano che ne facilita i confronti.
Si coglie la posizione anomala (rispetto a ciò che appare nel novero dei casi presentati) della Campania e del Lazio nonché tre distinti gruppi: Calabria-Puglia-Marche in basso; Veneto e Liguria al Centro e, più in alto, Toscana, Lombardia e Sicilia



Curve di Andrews/3

La chiarezza del grafico è legata:

- **Al numero di dati da rappresentare: non dovrebbe essere superiore a 10 ovvero non bisogna inserire un numero di unità tale da confonderne la leggibilità.**
- **Alla scelta dell'ordine di presentazione delle variabili. Infatti il tipo di curva non è invariante rispetto a quale variabile viene posta in prima posizione o in ultima posizione.**

Poiché le frequenze più basse sono più percepibili rispetto a quelle alte (cioè con coefficiente elevato negli argomenti del seno e del coseno) è opportuno che le variabili con maggiore dispersione (ad esempio quelle con maggiore coefficiente di variazione) siano inserite per prime

Curve di Andrews/4

Le curve di Andrews hanno due proprietà importanti:

- **La curva che rappresenta l'unità media si ottiene applicando la funzione armonica sul vettore delle medie:**

$$f_{\bar{x}}(t) = \frac{\bar{x}_1}{\sqrt{2}} + \bar{x}_2\,sen(t) + \bar{x}_3\cos(t) + \bar{x}_4\,sen(2t) + \bar{x}_5\cos(2t) + \ldots + \qquad -\pi \le t \le \pi$$

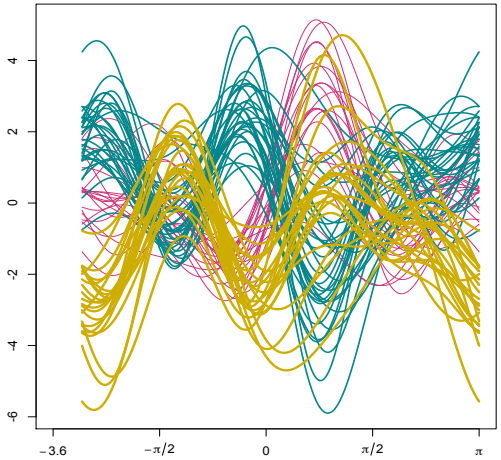
- **La distanza tra due curve ovvero tra due unità misurata come:**

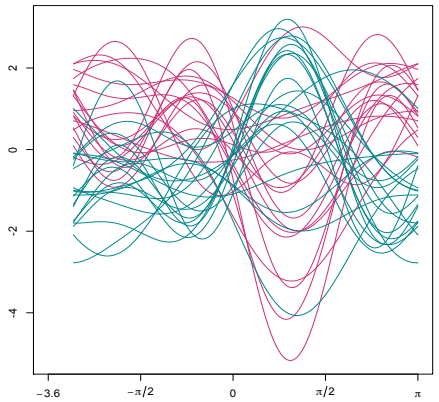
$$d(U_i, U_j) = \sqrt{\int_{-\pi}^{\pi} \left[f_i(t) - f_j(t) \right]^2 dt} \quad \propto \left[\sum_{k=1}^{m} \left[X_{ik} - X_{jk} \right]^2 \right]$$

è proporzionale alla distanza euclidea tra le due unità: ovvero giudicare "vicine" due unità che hanno curve nel piano molto prossime significa giudicarle vicine nel senso più proprio della distanza nello spazio ad "m" dimensioni.

Esempio: Lubishew data sets

Due data sets su alcuni tipi di insetti





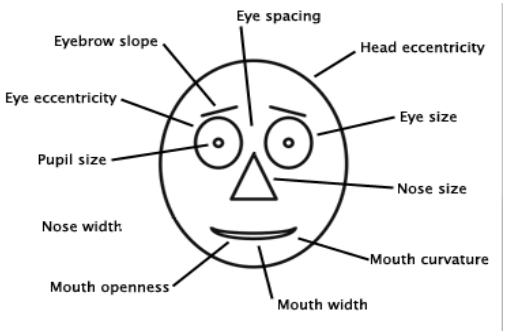
Cernoff faces

Nel 1973, Herman Chernoff ha introdotto una tecnica di visualizzazione per illustrare le tendenze nei dati multidimensionali.

I diversi valori dei dati sono abbinati alle caratteristiche del volto, per esempio la larghezza della faccia, il livello delle orecchie, la lunghezza o la curvatura della bocca, la lunghezza del naso, ecc.

Si usano le caratteristiche facciali per rappresentare le tendenze dei dati, non i valori stessi.

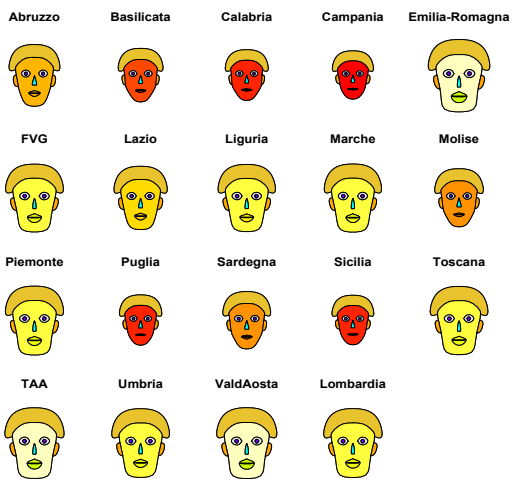
Mentre questa è chiaramente una limitazione, la conoscenza delle tendenze nei dati potrebbe contribuire a determinare quali sezioni dei dati sono di particolare interesse.



Cernoff faces/2

La tecnica è stata migliorata da Flury-Riedwyl (1988) con l'aggiunta di elementi somatici e del colore

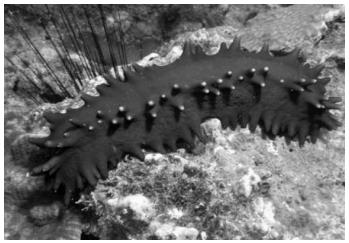
Occupazione femminile regionale 2005-2008



1 right eye size
2 right pupil size
3 position of right pupil
4 right eye slant
5 horizontal position of right eye
6 vertical position of right eye
7 curvature of right eyebrow
8 density of right eyebrow
9 horizontal position of right eyebrow
10 vertical position of right eyebrow
11 right upper hair line
12 right lower hair line
13 right face line
14 darkness of right hair
15 right hair slant
16 right nose line
17 right size of mouth
18 right curvature of mouth
19-36 like 1-18, only for the left side

Esempio: sea cucumber

Data on sea cucumbers (Edwards, 1908-9)



La soggettività della mappatura variabili/tratto somatico è una debolezza del metodo.
Ciò ouò comportare una variazione media del 25% tra una mappatura ed un'altra.

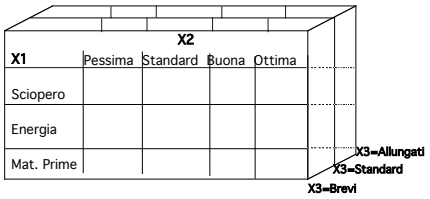
Un miglioramento si può avere proponendo le facce in 3d e rendendole dinamiche.

Distribuzioni discrete multidimensionali

Le indagini statistiche possono riguardare una molteplicità di aspetti e generare più variabili casuali

ESEMPIO

Il ciclo di produzione può avere 3 tipi di interruzione: X_1=(Sciopero, Energia,Materie prime). I prodotti sono di qualità X_2=(Pessima, Standard, Buona, Ottima) ed i tempi di produzione: possono essere X_3=(Standard, Ridotti, Allungati)



Per rappresentare l'esperimento usiamo la distribuzione congiunta trivariata

$$P(X_1, X_2, X_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \geq 0$$

$$\sum_{x_1}\sum_{x_2}\sum_{x_3} P(_1 = x_1, X_2 = x_2, X_3 = x_3) = 1$$

La P(.) associa ad ogni possibile terna una probabilità non negativa con il vincolo di somma unitaria

Distribuzioni marginali

Se è $X = (X_1, X_2, \ldots, X_n)$ **è una v.c. n-dimensionale la i-esima distribuzione marginale**

$$P(X_i = x_i) = \sum_{x_1}\sum_{x_2}\cdots\sum_{\substack{x_n \\ Escluso\ x_i}} P(X_1, X_2, \ldots, X_n)$$

dove la somma è estesa a tutte le variabili casuali tranne la i-esima

ESEMPIO
la distribuzione congiunta è presentata a strati

X_3=1

$X1/X2$	-1	0	1
-1	$4/24$	$1/24$	0
0	$2/24$	$1/24$	$2/24$
1	0	$1/24$	$2/24$

X_3=2

$X1/X2$	-1	0	1
-1	$1/24$	0	$2/24$
0	$1/24$	$1/24$	$2/24$
1	$3/24$	$1/24$	0

Calcoliamo $P(X_1)$

Occorre sommare sia rispetto alla X_2 che rispetto alla X_3

$X1$		$P(X_1 = x_1)$
-1	$4/24 + 1/24 + 0 + 1/24 + 0 + 2/24 =$	$8/24$
0	$2/24 + 1/24 + 2/24 + 1/24 + 1/24 + 2/24$	$9/24$
1	$0 + 1/24 + 2/24 + 3/24 + 1/24 + 0$	$7/24$

Esempio

In questo esperimento definiamo

E_1=Risultato del primo dadoE_1=5
E_2: Eisultato del secondo dado: E_2=3

Gli aspetti che ci interessano sono:

X_1= Somma dei punti: E_1+E_2
X_2= Valore massimo: max{E_1, E_2}
X_3= Differenza: l E_1-E_2 l

Ogni variabile casuale ha la sua distribuzione di probabilità marginale.

Ma ci sono anche le bivariate e la congiunta trivariata

X_1	$P(X_1 = x_1)$	X_2	$P(X_2 = x_2)$	X_3	$P(X_3 = x_3)$
2	1/36	1	1/36	0	6/36
3	2/36	2	3/36	1	10/36
4	3/36	3	5/36	2	8/36
5	4/36	4	7/36	3	6/36
6	5/36	5	9/36	4	4/36
7	6/36	6	11/36	5	2/36
8	5/36		1		1
9	4/36				
10	3/36				
11	2/36				
12	1/36				
	1				

Distribuzioni congiunte condizionali

Solo una estensione delle definizioni del caso bivariato

La distribuzione congiunta delle variabili $\dot{X}_1, \dot{X}_2, \ldots, \dot{X}_m$ **condizionate dalle altre variabili** $X_1^*, X_2^*, \ldots, X_k^*$ **è data dal rapporto:**

$$P(\dot{X}_1, \dot{X}_2, \ldots, \dot{X}_m | X_1^*, X_2^*, \ldots, X_k^*) = \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P(X_1^* = x_1^*, X_2^* = x_2^*, \ldots, X_k^* = x_k^*)}$$

ESEMPIO

$$P(X_1, X_2 | X_3 = 1) = \frac{P(X_1, X_2, X_3)}{P(X_3 = 1)} =$$

X_1/X_2	-1	0	1
-1	$4/13$	$1/13$	0
0	$2/13$	$1/13$	$2/13$
1	0	$1/13$	$2/13$

$$P(X_1, X_2 | X_3 = 2) = \frac{P(X_1, X_2, X_3)}{P(X_3 = 2)} =$$

X_1/X_2	-1	0	1
-1	$1/11$	0	$2/11$
0	$1/11$	$1/11$	$2/11$
1	$3/11$	$1/11$	0

Distribuzioni congiunte marginali

Le variabili casuali multidimensionali consentono di definire le distribuzioni marginali di ogni sottoinsieme.

Dividiamo le "n" v.c. in due gruppi distinti:

VARIABILI CHE INTERESSANO: $\dot{X}_1,\ \dot{X}_2,\ldots,\dot{X}_m$

con k+m=n

VARIABILI CHE NON INTERESSANO: $X_1^*,\ X_2^*,\ldots,X_k^*$

Per ottenere la congiunta delle variabili che interessano (marginali) rispetto alle altre occorre sommare per le variabili che non interessano

$$P(\dot{X}_1,\dot{X}_2,\ldots,\dot{X}_m)=\sum_{x_1^*}\sum_{x_2^*}\ldots\sum_{x_k^*}P(X_1,X_2,\ldots,X_n)$$

Sommando si elimina l'influenza degli aspetti dell'esperimento che si vogliono tenere fuori.

Esempio

Per determinare $P(X_1,X_2)$ dobbiamo eliminare l'influenza di X_3 sommando - cella per cella- le due tabelle precedenti

$X2/X3$	-1	0	1
-1	$5/24$	$1/24$	$2/24$
0	$3/24$	$2/24$	$4/24$
1	$3/24$	$2/24$	$2/24$

Possiamo determinare le altre due distribuzioni congiunte-marginali eliminando di volta in volta l'influenza della terza variabile

$X1/X3$	1	2
-1	$5/24$	$3/24$
0	$5/24$	$4/24$
1	$3/24$	$4/24$

$X1/X3$	1	2
-1	$5/24$	$3/24$
0	$5/24$	$4/24$
1	$3/24$	$4/24$

E' possibile ottenere la marginale singola di ognuna delle altre v.c. usando una qualsiasi delle congiunte che la coinvolgono.

La distribuzione multinomiale

Un esperimento consiste di n prove indipendenti svolte in condizioni identiche. In ciascuna prova sono possibili k modalità distinte, anche qualitative:

$$(X_1,X_2,\ldots,X_k)$$

Le probabilità dei singoli risultati sono costanti di prova in prova

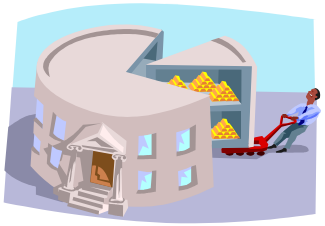
$$p_1,p_2,\ldots,p_k\geq 0;\qquad \sum_{i=1}^{k}p_i=1$$

Un modello adatto a tale esperimento è quello multinomiale con la seguente funzione di distribuzione:

$$P(X_1=x_1,X_2=x_2,\ldots,X_k=x_k)=\frac{n!}{x_1!*x_2!*\ldots*x_k!}p^{X_1}p^{X_2}*\ldots*p^{X_k}$$

$$x_1+x_2+\ldots+x_k=n$$

Esempio



Gli affidati di una banca sono:
X_1=solvibili con p1=0.60,
X_2=insolventi con p2=0.05,
X_3=incerti, ma positivi con p3=0.30,
X_4=incerti, ma negativi con p4=0.15.

Calcolare la probabilità che su n=10 ne risultino X_1=4,X_2=2,X_3=1, X_4=3

$$P(4,2,1,3)=\frac{10!}{4!2!1!3!}0.40^4*0.05^2*0.30*0.15^3=0.00081$$

Calcolare la probabilità che su n=20 ne risultino 5 di ciascun tipo:

$$P(5,5,5,5)=\frac{20!}{5!*5!*5!*5!}*0.6^5*0.05^5*0.30^5*0.15^5=0.000053$$

Ancora sulla multinomiale

Le distribuzioni marginali sono delle binomiali. Infatti, in ogni prova si verifica $X=X_i$ oppure non si verifica e p_i rimane costante nelle prove indipendenti

$$P(X_i = x_i) = \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n-x_i} \quad \textbf{con} \quad E(X_i) = np_i;\ \sigma^2(X_i) = np_i(1-p_i)$$

Note (k-1) variabili casuali componenti la variabile casuale multinomiale è nota anche la n-esima dato il vincolo di somma ad n dei risultati. Quindi la multinomiale ha (k-1) dimensioni

Le variabili componenti la multinomiale sono necessariamente correlate (e quindi dipendenti)

$$Cov(X_i, X_j) = -np_ip_j \quad \forall\ i \neq j$$

L'aumento dei successi in X_i non può che avvenire a danno di X_j

V.C. continue multidimensionali

La capacità rappresentativa del modello rispetto all'esperimento aumenta se aumentano gli aspetti di cui riesce a tenere conto. Ciò vale anche per i fenomeni continui.

ESEMPIO
La valutazione del carico di lavoro di una unità di personale che svolge 4 diversi compiti tiene conto dei tempi di svolgimento di ciascuno. Se i compiti sono tra loro indipendenti un modello adatto è:

$$f(X_1, X_2, X_3, X_4) = \lambda_1 * \lambda_2 * \lambda_3 * \lambda_4 * e^{-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_3 x_3 - \lambda_4 x_4} \qquad X_i \geq 0$$

La definizione della funzione di densità ricalca quella bivariata

$$f(X_1, X_2, \ldots, X_n) \geq 0$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(X_1, X_2, \ldots, X_n)dx_1dx_2\ldots dx_n = 1$$

$$P(X_1, X_2, \ldots X_n \in A) = \int\int\ldots\int_A f(X_1, X_2, \ldots, X_n)dx_1dx_2\ldots dx_n$$

Gli eventi di cui si calcola la probabiiltà sono degli ipervolumi

Esempio: uniforme sul tetraedro

$$f(X_1, X_2, X_3) = \begin{cases} 6 & per \quad x_1 + x_2 + x_3 \leq 1; \qquad x_1, x_2, x_3 \geq 0 \\ 0 \end{cases}$$

Ad ogni porzione del volume del tetraedro unitario la "f" assegna una densità di probabilità.

In questo modello la densità è costante

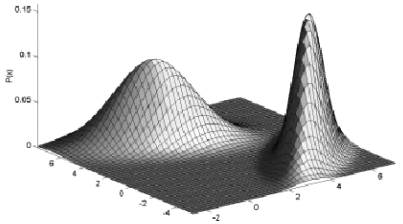
$$\int_0^1\int_0^{1-x_1}\int_0^{1-x_1-x_2} 6dx_3dx_2dx_1 =$$

$$6\int_0^1\int_0^{1-x_1}(1-x_1-x_2)dx_2dx_1 = 6\int_0^1 x_1 - x_1^2 dx_1$$

$$6\left(\frac{1}{2} - \frac{1}{3}\right) = 1$$

La gaussiana multivariata

Un vettore di variabili casuali ha distribuzione gaussiana multivariata con media μ e matrice di varianze-covarianze W



$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \mathbf{W})$$

se la sua densità congiunta è data da

$$f(x_1, x_2, \ldots, x_m) = \frac{e^{-0.5(\mathbf{x}-\mu)^t \mathbf{W}^{-1}(\mathbf{x}-\mu)}}{(2\pi)^{0.5n}|\mathbf{W}|^{0.5}}$$

<u>Proprietà importanti</u>

1)Se z=Ax è una trasformazione del vettore delle x allora anche z avrà distribuzione gaussiana

$$z \sim N(\mathbf{A}\mu, \mathbf{A}^t\mathbf{W}\mathbf{A})$$

2) Per la forma quadratica basata sulle variabili centrate si ha

$$(\mathbf{x} - \boldsymbol{\mu})^t \mathbf{W}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_m \ (chi\ quadrato)$$

Densità marginali e condizionate

Analogamente al caso discreto occorre integrare la funzione di densità congiunta rispetto alle variabili che non interessano

$$f_i(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_m) \prod_{j=1, j \neq i}^{m} dx_j$$

Per le distribuzioni condizionate le possibilità sono ora molte di più.

Infatti è possibile studiare la distribuzione di un gruppo di variabili G_1 condizionata da un altro gruppo di variabili G_2 con

$$G_1 \cap G_2 = \phi.$$

Ad esempio, supponendo che $G_1 = \{x_1, x_2, \ldots, x_k\}, \quad G_2 = \{x_{k+1}, x_{k+2}, \ldots, x_m\}$

La funzione di densità condizionata è

$$g(x_1, x_2, \ldots, x_k | x_{k+1}, x_{k+2}, \ldots, x_m) = \frac{f(x_1, x_2, \ldots, x_m)}{h(x_{k+1}, x_{k+2}, \ldots, x_m)}$$

Indipendenza di variabili casuali multiple

Siano $\{X_1, X_2, \ldots, X_n\}$ **n variabili casuali. Esse sono considerate indipendenti se per tutti gli eventi** $A_1, A_2, \ldots, A_n$

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

Che è la naturale estensione della definizione data per il caso n=2

L'indipendenza può anche essere formulata in base alla funzione di ripartizione:

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} F(x_i)$$

Accomunando così le v.c. continue e discrete in una unica formulazione della indipendenza

N.B. $g(G_1|G_2) = g(G_1) \;\; se \;\; G_1 \cap G_2 = \varnothing$

Riflessioni sull'indipendenza

L'indipendenza è una condizione molto forte a cui conseguono diversi risultati

- **Se** $\{X_1, X_2, \ldots, X_n\}$ **è un insieme di v.c. indipendenti allora lo è qualsiasi loro sottoinsieme.**
- **Se** $\{X_1, X_2, \ldots, X_n\}$ **è un insieme di v.c. indipendenti allora lo sono le rispettive trasformate**

$$\{g_1(x_1), g_2(x_2), \ldots, g_n(x_n)\}$$

- **Se** $\{X_1, X_2, \ldots, X_n\}$ **è un insieme di v.c. indipendenti allora lo è qualsiasi combinazione di loro funzioni.**
- **Se "n" variabili casuali sono MUTUALMENTE INDIPENDENTI, cioè indipendenti due a due, cioè indipendenti a coppie, questo non implica l'indipendenza delle terne, quaterne, etc.**

Esempio

Supponiamo che le tre variabili casuali discrete X,Y,Z abbiano distribuzione

$$P(X = x, Y = y, Z = z) = \frac{1}{4} \quad se \; (x,y,z) = \{(1,0,0);(0,1,0);(0,0,1);(1,1,1)\}$$

Come si vede, le coppie di v.c. sono indipendenti

$$P(X=0) = \frac{1}{2};\; P(X=1) = \frac{1}{2}; \qquad P(Y=0) = \frac{1}{2};\; P(Y=1) = \frac{1}{2}; \qquad P(Z=0) = \frac{1}{2};\; P(Z=1) = \frac{1}{2};$$

$$P(X=x, Y=y) = \frac{1}{4}; \quad P(X=x, Z=z) = \frac{1}{4}; \quad P(Y=x, Z=z) = \frac{1}{4}$$

La terna non è però indipendente

$$P(X=1, Y=0, Z=0) = \frac{1}{4} \neq P(X=1) * P(Y=0) * P(Z=0) = \frac{1}{8}$$

Relazioni tra variabili

Se tra le variabili sussistessero delle relazioni di dipendenza, allora la conoscenza di una o più potrebbe rendere superflue delle altre.

Tuttavia, non conosciamo quali relazioni legano le variabili

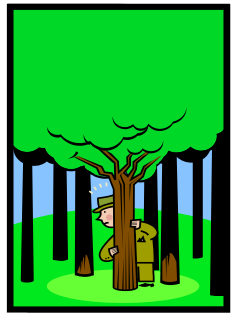
Se anche esistessero non sappiamo se sono esatte o approssimate e non sappiamo se sono stocastiche o deterministiche.

Le variabili casuali multidimensionali forniscono i modelli con I quali decriere analizzare ed interpretare le informazioni contenute nella matrice dei dati osservati.

Vedremo i pregi (poci) ed I difetti (molti) delle analisi multivariate basate su modelli). Al momento ci impegniamo nella ricerca di schemi elementari più gstiili.

La semplificazione suggerisce di pensare a relazioni facili da riconoscere e agevoli da negare

Perché i legami lineari



il rasoio di Occam

Se è necessario dare una soluzione ad un problema di cui si sa poco, la risposta più semplice comporta meno rischi in caso di errore ed è spesso quella giusta.

Smarriti in una foresta se ne esce spesso procedendo in linea retta.

L'uovo di Colombo

Principio di semplicità di Galilei

La natura procede per vie semplici ed offre così la sicura scelta tra le varie spiegazioni possibili dei suoi fenomeni

Una ragione formale

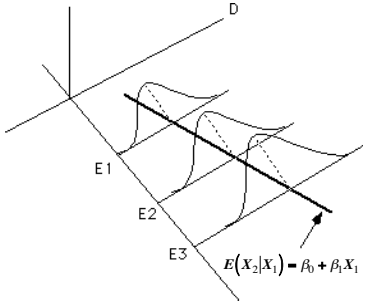
Supponiamo di scegliere un campione di unità e di rilevare su ogni unità due variabili : X_1 e X_2

La casualità di tale esperimento è descritta da una densità bivariata $f(X_1,X_2)$. Ipotizziamo che sia valido il modello gaussiano.

Ne consegue:

$$E\left(X_2|X_1\right) = \beta_0 + \beta_1 X_1;\ dove:\ \beta_0 = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1;\ \beta_1 = \rho\frac{\sigma_2}{\sigma_1}$$

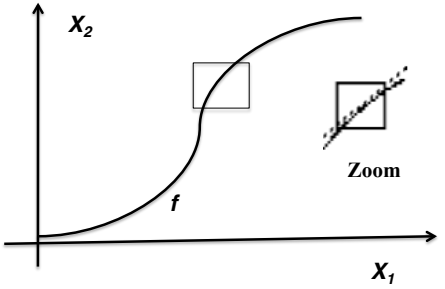
In questo modello il valore atteso di una variabilecasuale condizionato al valore dell'altra, è -necessariamente- una funzione lineare della condizionante.



Che succede se il coefficiente di correlazione "ρ" è nullo?

Teorema di Taylor

Se la funzione "f" che lega X_1 ad X_2 ha derivate prime e seconde continue in un intorno del punto P, in tale intorno la "f" è ben approssimata dalla retta



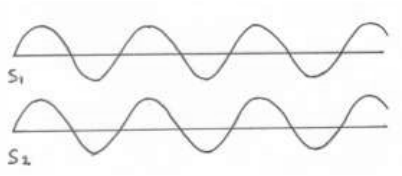
In generale si può dire che la scelta dei legami lineari è motivata da

- **Ragioni di semplicità**
- **Formalismo della gaussiana**
- **Approssimazione funzionale**

La concordanza

Se il numero di variabili è superiore a due o tre ovvero se i valori della matrice dei dati non denotano strutture subito visibili conviene partire da una misura delle intensità dei loro legami.

Un aspetto essenziale della dipendenza tra due variabili su scala almeno intervallare è la concordanza, cioè la ricerca della direzione e della intensità della dipendenza tra Y ed X.



Ci si chiede se valori inferiori (superiori) alla media di una variabile si accompagnino con valori inferiori (superiori) alla media nell'altra

Per ognuna delle combinazione di possibili valori si può averne una indicazione dagli SCARTI MISTI:

$$v_{ij} = \left(x_{h,j} - \mu_i\right)\left(x_{h,j} - \mu_j\right)$$

La codevianza

La sintesi più semplice degli scarti misti è la loro somma che costituisce la codevianza tra X_i ed X_j

$$v_{ij} = \sum_{h=1}^{n}\left(x_{h,j} - \mu_i\right)\left(x_{h,j} - \mu_j\right)$$

Se $v_{ij}>0$; Predominano gli scarti di segno concorde. Ci si aspetta X e Y tendano a cambiare nella stessa direzione

Se Se $v_{ij<}0$; Predominano gli scarti di segno discorde. Ci si aspetta X e Y tendano a cambiare in direzioni opposte

Se Se $v_{ij=}0$; le forze di discordanza e di concordanza sono bilanciate e le due variabili si dicono INCORRELATE

Significato della concordanza

Il segno degli scarti è utile per stabilire se, per la combinazione dei valori "$X_{h,i}$" e "$Y_{h,j}$" l'andamento delle due variabili è concorde oppure discorde:

CONCORDANZA

$$v_{ij} > 0 \Rightarrow \left(x_{h,j} > \mu_i\right)\left(x_{h,j} > \mu_j\right) \;\; ovvero \;\; \left(x_{h,j} < \mu_i\right)\left(x_{h,j} < \mu_j\right)$$

DISCORDANZA

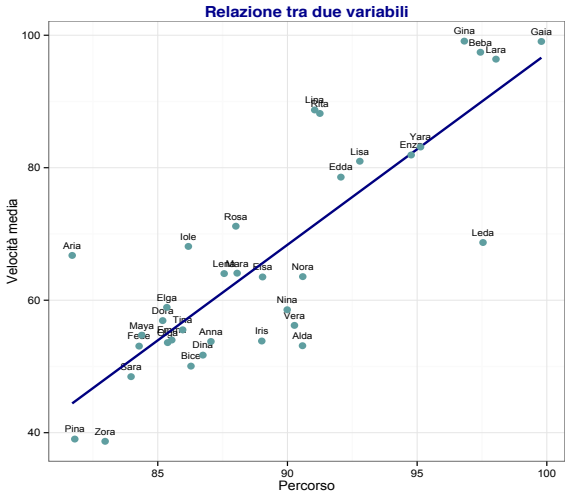
$$v_{ij} < 0 \Rightarrow \left(x_{h,j} > \mu_i\right)\left(x_{h,j} < \mu_j\right) \;\; ovvero \;\; \left(x_{h,j} < \mu_i\right)\left(x_{h,j} > \mu_j\right)$$

E' difficile cogliere il senso della concordanza analizzando uno per uno TUTTI gli scarti misti.

Guidatrice	Percorso	Vel_media
Alda	90.58	53.13
Anna	87.05	53.77
Aria	81.70	66.76
Beba	97.44	97.43
Bice	86.28	50.05
Dina	86.74	51.71
Dora	85.19	56.92
Edda	92.06	78.58
Elga	85.35	58.92
Elsa	89.04	63.51
Emma	85.54	53.99
Enza	94.77	81.92
Fede	84.28	53.07
Gaia	99.79	99.06
Gina	96.82	99.11
Iole	86.18	68.12
Iris	89.01	53.84
Lara	98.04	96.39
Leda	97.54	68.71
Lena	87.56	64.02
Lina	91.05	88.72
Lisa	92.79	80.98
Mara	88.06	64.08
Maya	84.38	54.72
Nina	89.99	58.56
Nora	90.59	63.56
Olga	85.38	53.60
Pina	81.80	39.04
Rosa	88.01	71.17
Rita	91.25	88.19
Sara	83.97	48.46
Tina	85.96	55.53
Vera	90.27	56.20
Yara	95.13	83.13
Zora	82.97	38.70

	Percorso	Vel_media
Percorso	23.944	69.080
Vel_media	69.080	287.787

Scatterplot



Dominano gli scarti concordi

Valore atteso o Centroide

Il concetto di centroide generalizza quello di valore atteso al caso multi-variato.

$$E(x) = \left[E(X_1), E(X_2), \ldots, E(X_m)\right] = (\mu_1, \mu_2, \ldots, \mu_m) = \mu$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}$$

E' costituito dal vettore i cui elementi sono le medie aritmetiche delle singole variabili.

La media globale ponderata ovvero il baricentro è definito dalla media aritmetica degli elementi del centroide.

$$\hat{\mu} = \sum_{j=1}^{m} w_j \mu_j; \qquad w_j \geq 0, \quad \sum_{j=1}^{m} w_j = 1$$

Dove i pesi derivano dall'importanza che ha la singola variabile casuale nella distribuzione

Esempio

$$u_1 = \begin{bmatrix} 4 & 8 & 16 \end{bmatrix} \qquad u_2 = \begin{bmatrix} 6 & 1 & -3 \end{bmatrix} \qquad \mu = \begin{bmatrix} 5.0 \\ 4.5 \\ 6.5 \end{bmatrix} \qquad w_1 = \frac{4}{15},\ w_1 = \frac{9}{15},\ w_1 = \frac{2}{15}$$

$$\hat{\mu} = 5\left(\frac{4}{15}\right) + 4.5\left(\frac{7}{15}\right) + 6.5\left(\frac{1}{15}\right) = \frac{20 + 31.5 + 6.5}{15} = \frac{58}{15} \cong 3.87$$

$$2^{-1} v_2 X = 2^{-1}\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} 4 & 8 & 16 \\ 6 & 1 & -3 \end{bmatrix} = 2^{-1}\begin{bmatrix} 4+6 & 8+1 & 16-3 \end{bmatrix} = 2^{-1}\begin{bmatrix} 10 & 9 & 13 \end{bmatrix} = \begin{bmatrix} 5 & 4.5 & 6.5 \end{bmatrix}$$

$$\hat{\mu} = \begin{bmatrix} \frac{4}{15} & \frac{7}{15} & \frac{1}{15} \end{bmatrix}\begin{bmatrix} 5.0 \\ 4.5 \\ 6.5 \end{bmatrix} = \left(\frac{4}{15}\right)5.0 + \left(\frac{7}{15}\right)4.5 + \left(\frac{1}{15}\right)6.5 \cong 3.87$$

Matrice di varianze-covarianze

Adottiamo la simbologia

$$\sigma_i^2 = E\left[(x_i - \mu_i)^2\right] \quad i = 1,2,\ldots,m$$

$$\sigma_{ij} = \operatorname{cov}(x_i, x_j) = E\left[(x_i - \mu_i)(x_j - \mu_j)\right] \quad i,j = 1,2,\ldots,m$$

Dal prodotto esterno dell vetto degli scarti si ha

$$x - \mu = \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_m - \mu_m \end{bmatrix}$$

$$(x-\mu)(x-\mu)^t = \begin{bmatrix} (x_1 - \mu_1) \\ (x_2 - \mu_2) \\ \vdots \\ (x_m - \mu_m) \end{bmatrix}\left[(x_1 - \mu_1)(x_2 - \mu_2)\ldots(x_m - \mu_m)\right]$$

$$= \begin{bmatrix} (x_1 - \mu_1)^2 & (x_1 - \mu_1)(x_2 - \mu_2) & \ldots & (x_1 - \mu_1)(x_m - \mu_m) \\ (x_1 - \mu_1)(x_2 - \mu_2) & (x_2 - \mu_2) & & (x_2 - \mu_2)(x_m - \mu_m) \\ \vdots & \vdots & \ddots & \vdots \\ (x_m - \mu_m)(x_1 - \mu_1) & (x_m - \mu_m)(x_2 - \mu_2) & \ldots & (x_m - \mu_m)^2 \end{bmatrix}$$

Matrice di devianze-codevianze

E' una matrice in cui ogni elemento è dato dalla somma degli scarti misti dalle medie aritmetiche (prodotto di scarti semplici) per due variabili alla volta

$$v_{ij} = \sum_{h=1}^{n}\left(x_{h,j} - \mu_i\right)\left(x_{h,j} - \mu_j\right) \qquad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_{11} & \boldsymbol{v}_{12} & \cdots & \boldsymbol{v}_{1m} \\ \boldsymbol{v}_{21} & \boldsymbol{v}_{22} & \cdots & \boldsymbol{v}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{v}_{m1} & \boldsymbol{v}_{m2} & \cdots & \boldsymbol{v}_{mm} \end{bmatrix}$$

L'ordine della matrice è (m x m) poiché vi sono m variabili che si possono porre in relaziioone con tutte le altre m, se stesse incluse

La matrice di devianze-codevianze si ottiene con un prodotto matriciale che coinvolge la matrice di centramento

$$\hat{X}^t \hat{X} = X^t C^t C X = X^t C X = V$$

Esempio

$$X = \begin{bmatrix} 1 & 1 & -3 \\ 4 & 0 & 10 \\ 2 & 2 & 5 \\ 1 & 1 & 0 \end{bmatrix},\ \hat{X} = CX = \begin{bmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix} \begin{bmatrix} 1 & 1 & -3 \\ 4 & 0 & 10 \\ 2 & 2 & 5 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & -6 \\ 2 & -1 & 7 \\ 0 & 1 & 2 \\ -1 & 0 & -3 \end{bmatrix}$$

$$V = \hat{X}^t \hat{X} = \begin{bmatrix} -1 & 2 & 0 & -1 \\ 0 & -1 & 1 & 0 \\ -6 & 7 & 2 & -3 \end{bmatrix} \begin{bmatrix} -1 & 0 & -6 \\ 2 & -1 & 7 \\ 0 & 1 & 2 \\ -1 & 0 & -3 \end{bmatrix} = \begin{bmatrix} 6 & -2 & 23 \\ -2 & 2 & -5 \\ 23 & -5 & 98 \end{bmatrix}$$

Questi risultati giustificano l'adozione della trasformazione lineare

$$\hat{X} = CX$$

il cui effetto è di spostare l'origine degli assi sul centroide della matrice dei dati. La dispersione dei punti rimane però invariata.

Matrice di varianze-covarianze campionaria

Se invece del totale degli scarti misti consideriamo la loro media otteniamo la covarianza tra le due variabili

$$w_{ij} = \frac{\sum_{h=1}^{n} \left(x_{h,j} - \mu_i \right) \left(x_{h,j} - \mu_j \right)}{n}$$

La matrice di varianze-covarianze deriva dal prodotto di matrici trasformate

$$W = \frac{1}{n} V = \frac{1}{n} X^t C X = \begin{bmatrix} \frac{1}{\sqrt{n}} & X^t & C^t \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n}} & C & X \end{bmatrix}$$

Ovvero dalla trasformazione

$$Y = \frac{1}{\sqrt{n}} CX \quad \Rightarrow \quad Y^t Y = W$$

Esempio

$$X = \begin{bmatrix} 1 & 1 & -3 \\ 4 & 0 & 10 \\ 2 & 2 & 5 \\ 1 & 1 & 0 \end{bmatrix}$$

$$Y = \frac{1}{\sqrt{n}} CX = \frac{1}{\sqrt{4}} \begin{bmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix} \begin{bmatrix} 1 & 1 & -3 \\ 4 & 0 & 10 \\ 2 & 2 & 5 \\ 1 & 1 & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -1 & 0 & -6 \\ 2 & -1 & 7 \\ 0 & 1 & 2 \\ -1 & 0 & -3 \end{bmatrix}$$

$$Y^t Y = \frac{1}{4} \begin{bmatrix} -1 & 2 & 0 & -1 \\ 0 & -1 & 1 & 0 \\ -6 & 7 & 2 & -3 \end{bmatrix} \begin{bmatrix} -1 & 0 & -6 \\ 2 & -1 & 7 \\ 0 & 1 & 2 \\ -1 & 0 & -3 \end{bmatrix} = \begin{bmatrix} 3 & -1 & 11.5 \\ -1 & 1 & -2.5 \\ 11.5 & -2.5 & 49 \end{bmatrix} = S$$

Covarianza e trasformazioni lineari

La covarianza risente delle trasformazioni moltiplicative, ma non di quelle additive. Consideriamo, ad esempio, le trasformazioni lineari

$$Y_{h,i} = a_i + b_i X_{h,i}, \quad h = 1,2,\ldots,n; \quad i = 1,2,\ldots,m$$

Si ha

$$\begin{aligned} n\left[\operatorname{cov}\left(Y_i, Y_j\right)\right] &= \sum_{h=1}^{n} \left(Y_{h,i} - \bar{y}_i \right) \left(Y_{h,j} - \bar{y}_j \right) = \sum_{h=1}^{n} \left(a_i + b_i X_{h,i} - a_i - b_i \bar{x}_i \right) \left(a_j + b_j X_{h,j} - a_j - b_j \bar{x}_j \right) \\ &= \sum_{h=1}^{n} \left(X_{h,i} - \bar{x}_i \right) \left(X_{h,j} - \bar{x}_j \right) = b_i b_j \sum_{h=1}^{n} \left(X_{h,i} - \bar{x}_i \right) \left(X_{h,j} - \bar{x}_j \right) \\ &= b_i b_j \operatorname{cov}\left(X_i, X_j\right) \end{aligned}$$

i parametri additivi sono spariti, quelli moltiplicativi compaiono come fattore

Disuguaglianza Cauchy-Schwartz

Consideriamo una relazione che lega linearmente gli scarti medi della X_i agli scarti medi della X_j

$$v(\theta) = \frac{\sum_{h=1}^{n}\left[\left(X_{h,i} - \bar{x}_i\right) - \theta\left(X_{h,j} - \bar{x}_j\right)\right]^2}{n}$$

Esprime l'errore quadratico medio che si commette ipotizzando un comune fattore proporzionale tra i due tipi di scarto

$$v(\theta) = \frac{\sum_{h=1}^{n}\left[\left(X_{h,i} - \bar{y}_i\right)^2 + \theta^2\left(X_{h,j} - \bar{y}_j\right)^2 - 2\theta\left(X_{h,i} - \bar{y}_i\right)\left(X_{h,j} - \bar{y}_j\right)\right]}{n}$$

$$= \frac{\sum_{h=1}^{n}\left(X_{h,i} - \bar{y}_i\right)^2 + \theta^2\sum_{h=1}^{n}\left(X_{h,j} - \bar{y}_j\right)^2 - 2\theta\sum_{h=1}^{n}\left(X_{h,i} - \bar{y}_i\right)\left(X_{h,j} - \bar{y}_j\right)}{n}$$

$$= \frac{\sum_{h=1}^{n}\left(X_{h,i} - \bar{y}_i\right)^2}{n} + \frac{\theta^2\sum_{h=1}^{n}\left(X_{h,j} - \bar{y}_j\right)^2}{n} - 2\theta\frac{\sum_{h=1}^{n}\left(X_{h,i} - \bar{y}_i\right)\left(X_{h,j} - \bar{y}_j\right)}{n}$$

$$= \mathrm{var}\left(X_i\right) + \theta^2\,\mathrm{var}\left(X_j\right) - 2\theta\,\mathrm{cov}\left(X_i, X_j\right)$$

Poiché la quantità è positiva le radici dell'equazione quadratica sono immaginarie.

$$4\,\mathrm{cov}\left(X_i, X_j\right)^2 - 4\,\mathrm{var}\left(X_i\right)\mathrm{var}\left(X_j\right) < 0 \Rightarrow \mathrm{cov}\left(X_i, X_j\right)^2 < \mathrm{var}\left(X_i\right)\mathrm{var}\left(X_j\right)$$

La covarianza, al quadrato, è inferiore o uguale al prodotto delle varianze delle due variabili

Criticità della covarianza

La covarianza ha tutti i difetti delle misure assolute di variabilità: dipendenza dall'unità di misura, mancanza di limiti predefiniti, etc.

E' però legata alla dispersione delle due variabili nel senso che non può superare, se considerata in valore assoluto, il prodotto degli scarti quadratici medi

Per ottenere un indice normalizzato e standardizzato si considerano le variabili espresse come corretti rispetto alla media e divisi per lo scarto quadratico medio

$$r_{ij} = \frac{\sum_{h=1}^{n}\left(\frac{x_{h,j} - \mu_i}{v_i}\right)\left(\frac{x_{h,j} - \mu_j}{v_j}\right)}{n}, \qquad v_i = \sqrt{\frac{\sum_{h=1}^{b}\left(x_{h,i} - \mu_i\right)^2}{n}}; \quad v_j = \sqrt{\frac{\sum_{h=1}^{b}\left(x_{h,j} - \mu_j\right)^2}{n}}$$

Che esprime il coefficiente di correlazione tra X_i ed X_j.

Coefficiente di correlazione

- **E' compreso tra -1 e +1 perché espresso come rapporto di una quantità (la covarianza) al suo massimo (in valore assoluto)**
- **E' standardizzato. Se una o entrambe le variabili subiscono una trasformazione lineare il coefficiente rimane lo stesso:**
- **E' simmetrico rispetto alle due variabili**
- **E' uguale a zero se c'è equilibrio tra scarti superiori ed inferiori alla media**

Coefficiente di correlazione/2

Assume i valori estremi solo in caso di relazione lineare esatta

$$r\left(X_i, a + bX_i\right) = \frac{\sum_{h=1}^{n}\left(\frac{X_{h,i} - \bar{x}_i}{v_i}\right)\left(\frac{a + bX_{h,j} - a - b\bar{x}_i}{|b|v_i}\right)}{n} = \frac{\sum_{h=1}^{n}\left(\frac{X_{h,i} - \bar{x}_i}{v_i}\right)\left(\frac{bX_{h,j} - b\bar{x}_i}{|b|v_i}\right)}{n}$$

$$= \frac{\frac{b}{|b|}\sum_{h=1}^{n}\left(\frac{X_{h,i} - \bar{x}_i}{v_i}\right)\left(\frac{X_{h,j} - \bar{x}_i}{v_i}\right)}{n} = \frac{b}{|b|}\frac{1}{v_i}\left[\frac{\sum_{h=1}^{n}\left(X_{h,i} - \bar{x}_i\right)^2}{n}\right]$$

$$= \frac{b}{|b|}$$

Ne consegue che r è +1 oppure -1 secondo che la relazione esatta sia di tipo diretto oppure inverso

il coefficiente di correlazione misura, quindi, l'intensità del legame lineare che esiste tra le due variabili.