DISTRIBUZIONI DOPPIE
(ANALISI DESCRITTIVE)

Fulvio De Santis

a.a. 2007-2008

Prerequisiti

- Popolazione, unità, carattere
- Come "nascono" i dati: osservazione e sperimentazione
- Popolazione: reale e virtuale
- Classificazione dei caratteri
- Le distribuzioni semplici - Distribuzioni per unità e di frequenze
- Sintesi delle distribuzioni semplici - Indici di posizione (moda, mediana e medie)
- Sintesi delle distribuzioni semplici: misura della variabilità

Distribuzioni Multiple

Quando su ciascuna di n unità statistiche di un collettivo si rilevano 2 o più caratteri, otteniamo una *distribuzione statistica multipla*.

Come nel caso delle distribuzioni semplici (secondo un solo carattere), possiamo avere sia distribuzioni *unitarie* che distribuzioni di *frequenze*

La distribuzione unitaria è rappresentata dalla *matrice dei dati*.

Esempio: vedi la distribuzione unitaria di 50 individui secondo anni di scolarità, condizione professionale, età. Si tratta di una ditribuzione tripla (secondo 3 caratteri).

La Matrice dei Dati

Indipendentemente dallo schema concettuale che ha prodotto i dati, (osservazionale/sperimentale), questi possono essere "organizzati" sotto forma della *matrice dei dati*:

• *Matrice dei Dati*: tabella in cui le righe si riferiscono alle unità statistiche e le colonne ai caratteri (variabili) di interesse.

Indichiamo con

$$X_1, X_2, \ldots, X_j, \ldots, X_m$$

m generiche variabili di interesse, ciascuna rilevata su ognuna di n unità statistiche:

Unità	X_1	$\ldots$	X_j	$\ldots$	X_m
1	x_{11}	$\ldots$	x_{1j}	$\ldots$	x_{1m}
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$
i	x_{i1}	$\ldots$	x_{ij}	$\ldots$	x_{im}
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$
n	x_{n1}	$\ldots$	x_{nj}	$\ldots$	x_{nm}

NOTA BENE

• x_{ij} rappresenta la modalità del carattere X_j presente nella unità i-esima.

• La matrice dei dati rappresenta la *distribuzione multipla per unità* delle variabili $X_1, X_2, \ldots, X_j, \ldots, X_m$.

• Informazione massima: molto dettaglio, poca sintesi!!

Distribuzioni Doppie

Ci concentriamo sul caso delle distribuzioni doppie, ovvero secondo 2 caratteri ($m = 2$). In particolare, introduciamo le seguenti distribuzioni

- distribuzione **congiunta**
- distribuzioni **marginali**
- distribuzioni **condizionate** o **parziali**

In questo caso, la distribuzione di frequenze viene rappresentata attraverso

Distribuzione Congiunta di Due Caratteri

Consideriamo due generici caratteri X e Y:

$X \backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	Totale
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2t}	$n_{2.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{it}	$n_{i.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_s	n_{s1}	n_{s2}	$\cdots$	n_{sj}	$\cdots$	n_{st}	$n_{s.}$
Totale	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.t}$	n

dove:

- n_{ij} : numero di osservazioni che presentano la modalità i di X e j di Y;
- $n_{i\cdot} = \sum_{j=1}^{t} n_{ij}$: numero di unità con la modalità x_i di X $(i = 1, 2, \ldots, s)$ (*frequenze marginali assolute di* X);
- $n_{\cdot j} = \sum_{i=1}^{s} n_{ij}$: numero di unità con la modalità y_j di Y $(j = 1, 2, \ldots, t)$ (*frequenze marginali assolute di* Y); ovvero;
- $n = \sum_{i=1}^{s} \sum_{j=1}^{t} n_{ij} = \sum_{i=1}^{s} n_{i\cdot} = \sum_{j=1}^{t} n_{j\cdot}$: numero totale osservazioni.

Distribuzioni Marginali

Sono due: la distribuzione marginale del carattere X e quella del carattere Y

Distribuzione marginale di X

X	x_1	x_2	$\cdots$	x_i	$\cdots$	x_s	Totale
	$n_{1.}$	$n_{2.}$	$\cdots$	$n_{i.}$	$\cdots$	$n_{s.}$	n

Distribuzione marginale di Y

Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	Totale
	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.t}$	n

- La distribuzione marginale del carattere X permette di studiare il modo in cui si presenta il carattere nel collettivo in esame, indipendentemenete dai valori assunti dal carattere Y;
- la distribuzione marginale di Y permette di studiare il modo in cui si presenta il carattere Y nel collettivo in esame, indipendentemente dai valori assunti dal carattere X.

Distribuzioni Condizionate o Parziali

Da una distribuzione doppia, oltre alle distribuzioni marginali di X e di Y si possono ricavare altre distribuzioni univariate: le distribuzioni di un solo carattere dopo aver fissato una modalità dell'altro.

Distribuzione condizionata di X rispetto a $Y = y_j$

X	x_1	x_2	$\cdots$	x_i	$\cdots$	x_s	Totale
y_j	n_{1j}	n_{2j}	$\cdots$	n_{ij}	$\cdots$	n_{sj}	$n_{\cdot j}$

Ci sono t distribuzioni condizionate di X, tante quante sono le modalità di Y

Distribuzione condizionata di Y rispetto a $X = x_i$

Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	Totale
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{it}	$n_{i\cdot}$

Ci sono s distribuzioni condizionate di Y, tante quante sono le modalità di X

Riassumendo, data una distribuzione secondo 2 caratteri con $s \times t$ combinazioni di modalità, abbiamo:

- 1 distribuzione doppia congiunta

- 2 ditribuzioni marginali

- t distribuzioni parziali di X

- s distribuzioni parziali di Y

Distribuzioni delle Frequenze Relative

Calcolando

$$f_{ij} = \frac{n_{ij}}{n} \qquad i = 1, \ldots, s \quad j = 1, \ldots t$$

si ottiene la distribuzione doppia delle frequenze relative

$X \backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	Totale
x_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1t}	$f_{1\cdot}$
x_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2t}	$f_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{it}	$f_{i\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_s	f_{s1}	f_{s2}	$\cdots$	f_{sj}	$\cdots$	f_{st}	$f_{s\cdot}$
Totale	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot j}$	$\cdots$	$f_{\cdot t}$	1

dove:

- $f_{ij} = \frac{n_{ij}}{n}$: $(i = 1, \ldots, s \quad j = 1, \ldots t)$ frequenze assolute relative
- $f_{i\cdot} = \sum_{j=1}^{t} f_{ij}$ $(i = 1, 2, \ldots, s)$ frequenze marginali relative di X;
- $f_{\cdot j} = \sum_{i=1}^{s} f_{ij}$ $(j = 1, 2, \ldots, t)$ frequenze marginali relative di Y;
- $\sum_{i=1}^{s} \sum_{j=1}^{t} f_{ij} = \sum_{i=1}^{s} f_{i\cdot} = \sum_{j=1}^{t} f_{j\cdot} = 1.$

Quindi: due distribuzioni marginali delle frequenze relative

Distribuzione marginale delle f.r. di i X

X	x_1	x_2	$\cdots$	x_i	$\cdots$	x_s	Totale
	$f_{1.}$	$f_{2.}$	$\cdots$	$f_{i.}$	$\cdots$	$f_{s.}$	1

Distribuzione marginale delle f.r. di i Y

Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	Totale
	$f_{.1}$	$f_{.2}$	$\cdots$	$f_{.j}$	$\cdots$	$f_{.t}$	1

Distribuzioni delle Frequenze Relative Condizionate

La distribuzione di frequenze relative di X condizionata a $Y = y_j$ si ottiene dividendo ogni elemento della colonna j-esima della distribuzione doppia per il totale

$$f_{x_i|y_j} = \frac{n_{ij}}{n_{.j}} \qquad\qquad i = 1, 2, \ldots, s$$

La distribuzione di frequenze relative di Y condizionata a $X = x_i$ si ottiene dividendo ogni elemento della riga i-esima della distribuzione doppia per il totale

$$f_{y_j|x_i} = \frac{n_{ij}}{n_{i.}} \qquad\qquad j = 1, 2, \ldots, t.$$

Volendo rappresentare tutte le distribuzioni di $X \mid Y = y_j$ si divide ogni colonna della distribuzione doppia di frequenze assolute per il corrispondente totale $n_{.j}$.

Distribuzioni Condizionate delle Frequenze Relative di X

$X \backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	Totale
x_1	$\frac{n_{11}}{n_{\cdot 1}}$	$\frac{n_{12}}{n_{\cdot 2}}$	$\cdots$	$\frac{n_{1j}}{n_{\cdot j}}$	$\cdots$	$\frac{n_{1s}}{n_{\cdot s}}$	$\frac{n_{1\cdot}}{n}$
x_2	$\frac{n_{21}}{n_{\cdot 1}}$	$\frac{n_{22}}{n_{\cdot 2}}$	$\cdots$	$\frac{n_{2j}}{n_{\cdot j}}$	$\cdots$	$\frac{n_{2t}}{n_{\cdot t}}$	$\frac{n_{2\cdot}}{n}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	$\frac{n_{i1}}{n_{\cdot 1}}$	$\frac{n_{i2}}{n_{\cdot 2}}$	$\cdots$	$\frac{n_{ij}}{n_{\cdot j}}$	$\cdots$	$\frac{n_{it}}{n_{\cdot t}}$	$\frac{n_{i\cdot}}{n}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_s	$\frac{n_{s1}}{n_{\cdot 1}}$	$\frac{n_{s2}}{n_{\cdot 2}}$	$\cdots$	$\frac{n_{sj}}{n_{\cdot j}}$	$\cdots$	$\frac{n_{st}}{n_{\cdot t}}$	$\frac{n_{s\cdot}}{n}$
Totale	1	1	$\cdots$	1	$\cdots$	1	1

Nota Bene

- Nella precedente tabella la lettura per riga ha perso significato.
- In modo analogo a quanto fatto sopra per rappresentare tutte le distribuzioni delle frequenze relative condizionate di Y si divide ogni riga della distribuzione doppia di frequenze assolute per il corrispondente totale di riga $n_{i\cdot}$.

Esempio

In un collettivo di $n = 100$ famiglie sono stati rilevati i caratteri X, numero di componenti, e Y, numero di stanze dell'abitazione. Si è ottenuta la seguente

Distribuzione di frequenze doppia delle frequenze assolute

$x_i \backslash y_j$	1	2	3	4	Totale
1	15	10	3	2	30
2	10	22	12	6	50
3	5	8	5	2	20
Totale	30	40	20	10	100

Distribuzione doppia di frequenze relative

$x_i \backslash y_j$	1	2	3	4	Totale
1	0.15	0.10	0.03	0.02	0.30
2	0.10	0.22	0.12	0.06	0.50
3	0.05	0.08	0.05	0.02	0.20
Totale	0.30	0.40	0.20	0.10	1

Distribuzione marginale delle frequenze relative di X

x_i	1	2	3	Totale
$f_{i\cdot}$	0.3	0.5	0.2	1

Distribuzione marginale di frequenze relative di Y

y_j	1	2	3	4	Totale
$f_{\cdot j}$	0.3	0.4	0.2	0.1	1

Distribuzioni di X condizionate alle diverse modalità di Y

$x_i \backslash y_j$	1	2	3	4
1	0.5	0.25	0.15	0.2
2	0.33	0.55	0.6	0.6
3	0.17	0.2	0.25	0.2
Totale	1	1	1	1

Ad esempio: distribuzione condizionata relativa del numero di componenti riferita alle famiglie che vivono in case con tre stanze:

x_i	1	2	3	Totale
$f_{x_i \mid Y=3\cdot}$	0.15	0.6	0.25	1

Distribuzioni di Y condizionate alle diverse modalità di X

$x_i \backslash y_j$	1	2	3	4	Totale
1	0.5	0.33	0.1	0.07	1
2	0.2	0.44	0.24	0.12	1
3	0.25	0.4	0.25	0.1	1

Relazioni tra due Caratteri Statistici.

Consideriamo tre diversi approcci, distinguendo tra:

- distribuzioni doppie rispetto a due caratteri qualitativi

- distribuzioni doppie rispetto a due caratteri di cui uno almeno sia quantitativo

- distribuzioni doppie rispetto a due caratteri quantitativi

Introduciamo quindi i concetti di:

- dipendenza o connessione tra caratteri statistici

- dipendenza in media

- correlazione

Infine parleremo di *interpolazione* e *regressione*.

Connessione

Riprendiamo in considerazione una distribuzione doppia di frequenze per due generici caratteri X e Y:

$X \backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_t	Totale
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2t}	$n_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{it}	$n_{i\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_s	n_{s1}	n_{s2}	$\cdots$	n_{sj}	$\cdots$	n_{st}	$n_{s\cdot}$
Totale	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot t}$	n

Definizione

In una distribuzione doppia secondo i caratteri (X, Y), il carattere Y è *indipendente* da X (o *non connesso con* X) se sono *simili* tra loro tutte le distribuzioni parziali di Y, corrispondenti alle modalità di X.

Due distribuzioni parziali sono *simili* se presentano stesse modalità e uguali frequenze relative.

Se invece almeno una coppia di distribuzioni parziali non è uguale, i caratteri X e Y si dicono dipendenti o connessi.

Connessione nulla o indipendenza

Y non è connesso con X se, per ogni coppia di modalità x_i e x_h di X, le distribuzioni parziali di Y sono *simili* tra loro e simili alla distribuzione marginale di Y, ovvero se per ogni coppia (h, i) e per ogni j, si ha

$$\frac{n_{hj}}{n_{h0}} = \frac{n_{ij}}{n_{i0}} = \frac{n_{0j}}{n}.$$

Invertendo i termini medi nella proporzione, otteniamo che, per ogni i, j

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i0}}{n}$$

ovvero che ogni distribuzione parziale di X è simile alla distribuzione parziale di X, che è la definizione di indipendenza di X da Y.

Quindi: se Y è indipendente da X, anche X è indipendente da Y.

IMPORTANTE: il concetto di indipendenza tra variabili in statistica è simmetrico, diversamente da quanto avviene in matematica.

Pertanto, nel caso di indipendenza, la frequenza assoluta della coppia i, j, che indichiamo con $\bar{n}_{ij}$, è

$$\bar{n}_{ij} = \frac{n_{i0} n_{0j}}{n}$$

mentre la frequenza relativa, che indichiamo con $\bar{f}_{ij}$, è

$$\bar{f}_{ij} = \frac{\bar{n}_{ij}}{n} = \frac{n_{i0} n_{0j}}{n \cdot n} = f_{i0} f_{0j}$$

Dipendenza perfetta

Y dipende perfettamente da X se a ogni modalità x_i di X corrisponde una sola modalità di Y con $n_{ij} \neq 0$, e se Y assume almeno due modalità diverse al variare di X. Questo vuol dire che x_i individua una modalità di Y.

Ad esempio

$X \backslash Y$	y_1	y_2	Totale
x_1	0	3	3
x_2	0	2	2
x_3	5	0	5
Totale	5	5	10

N.B. La dipendenza perfetta non è bilaterale.

Si ha dipendenza perfetta bilaterale se a ogni modalità x_i di X corrisponde una sola modalità di Y con $n_{ij} \neq 0$, e viceversa

$$\Rightarrow$$

si può avere solo se $s = t$.

Misura della connessione

Due approcci alternativi:

- misurare la dissomiglianza tra le distribuzioni marginali

- misurare la "distanza" tra la distribuzione osservata e la tabella di connessione nulla, (quella con frequenze teoriche $\bar{n}_{ij}$).

Consideriamo solo la seconda strada, che porta al principale indice di connessione, il χ^2, definito come

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(n_{ij} - \bar{n}_{ij})^2}{\bar{n}_{ij}}$$

che, con semplici passaggi, si può scrivere come

$$\chi^2 = n \left(\sum_{i=1}^{s} \sum_{j=1}^{t} \frac{n_{ij}^2}{n_{i0} n_{0j}} - 1 \right).$$

Osservazioni:

- assume il valore zero nel caso di indipendenza
- cresce al crescere delle quantità $(n_{ij} - \bar{n}_{ij})^2$
- dipende da $n \Rightarrow \Phi^2 = \frac{\chi^2}{n}$ (contingenza quadratica media)
- Φ^2 è un indice assoluto, che assume il valore massimo pari a

$$\min(s-1, t-1)$$

Quindi, l'indice relativo di connessione che consideriamo è

$$\phi^2 = \frac{\Phi^2}{\min(s-1, t-1)}$$

Esempio

La seguente tabella rappresenta la distribuzione doppia rispetto al carattere *sopravvivenza* e alla *tipologia di passeggero* riferita all'affondamento del Titanic, avvenuta il 15 aprile 1925.

$X \backslash Y$	uomini	donne	ragazzi	ragazze	Totale
sopravv	332	318	29	27	706
non sopravv	1360	104	35	18	1517
Totale	1692	422	64	45	2223

Dai dati risulta che sono sopravvissuti il 19.6% degli uomini, il 75.4% delle donne, il 45% dei ragazzi e il 60 % delle ragazze.

Vogliamo stabilire se vi è o meno dipendenza tra i due caratteri considerati e, se si, misurare il grado di connessione con l'indice ϕ^2.

La tabella delle frequenze teoriche, $\bar{n}_{ij}$ risulta

$X \backslash Y$	uomini	donne	ragazzi	ragazze	Totale
sopravv	537.360	134.122	20.326	14.291	706
non sopravv	1154.640	287.978	43.674	30.709	1517
Totale	1692	422	64	45	2223

L'indice χ^2 risulta

$$\chi^2 = \frac{(332 - 537.360)^2}{537.360} + \ldots + \frac{(18 - 30.709)^2}{30.709} = 507.084.$$

L'indice relativo di connessione ϕ^2 è:

$$\phi^2 = \frac{507.084}{2223} = \frac{0.228}{\min(2-1, 4-1)} = 0.228$$

che indica un livello piuttosto basso di dipendenza tra i due caratteri.

A livello descrittivo non si può aggiungere molto.

Usando tecniche inferenziali ed effettuando un *test di indipendenza*, l'ipotesi di dipendenza tra i due caratteri viene rifiutata al livello 0.05.

Possiamo concludere che la sopravvivenza nel disastro del Titanic non è statisticamente legata al fatto di essere maschio o femmina, giovane o meno giovane (!!!)

Dipendenza in media

Supponiamo ora che Y sia un carattere quantitativo e X un carattere qualsiasi.

In tal caso, per ciascuna delle s distribuzioni parziali di Y, possiamo calcolare la media aritmetica:

$$\bar{y}_1 \qquad \bar{y}_2 \qquad \dots \qquad \bar{y}_i \qquad \dots \qquad \bar{y}_s,$$

dove

$$\bar{y}_i = \frac{1}{n_{i0}} \sum_{j=1}^{t} y_j n_{ij}.$$

Possiamo anche calcolare la media aritmetica $\bar{y}$ della distribuzione marginale di Y:

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{t} y_j n_{0j}.$$

Definizione

Si dice che Y è *indipendente in media* da X se le medie delle distribuzioni parziali $\bar{y}_i$ risultano tutte uguali tra loro, e uguali anche alla media generale $\bar{y}$, ovvero se:

$$\bar{y}_1 = \bar{y}_2 = \ldots = \bar{y}_i = \ldots = \bar{y}_s = \bar{y}.$$

L'indice di dipendenza in media è il *rapporto di correlazione* di Pearson:

$$\eta_{y|x} = \frac{\sigma_{\bar{y}}}{\sigma_y},$$

dove

$$\sigma^2_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{s}(\bar{y}_i - \bar{y})^2 n_{i0} \qquad \text{e} \qquad \sigma^2_y = \frac{1}{n}\sum_{j=1}^{t}(y_j - \bar{y})^2 n_{0j}.$$

Osservazioni:

- il numeratore è la radice quadrata della varianza delle medie aritmetiche $\bar{y}_i$

- il denominatore è la radice quadrata della varianza totale della distribuzione del carattere Y

- $\eta_{y|x}$ è un indice relativo, che risulta uguale a zero quando tutte le medie delle distribuzioni parziali di Y sono uguali tra loro e vale 1 nel caso di dipendenza perfetta.

Esempio

Si consideri la distribuzione doppia di un gruppo di laureati in Economia, secondo i caratteri *sesso* (X) e *voto di laurea* (Y)

$X \backslash Y$	≤ 87	88-98	99-109	≥ 110	Totale
Maschi	29	112	151	95	387
Femmine	3	44	83	61	191
Totale	32	156	234	156	578

Assumendo come rappresentativi delle classi di modalità di Y i valori:

$$84 \quad 93 \quad 104 \quad 111$$

si ha:

$$\bar{y}_1 = 101.04 \qquad \bar{y}_2 = 103.49 \qquad \bar{y} = 101.81$$

Inoltre

$$\sum_{i=1}^{s}(\bar{y}_i - \bar{y})^2 n_{i0} = \sum_{i=1}^{s} \bar{y}_i^2 n_{i0} - n\bar{y}^2 = \ldots = 5429.58$$

e

$$\sum_{j=1}^{t}(y_j - \bar{y})^2 n_{0j} = \sum_{j=1}^{t} y_j^2 n_{0j} - n\bar{y}^2 = \ldots = 36926.42.$$

Quindi

$$\eta_{y|x} = \frac{5429.58}{36926.42} = 0.15.$$

L'indice suggerisce una non forte dipendenza in media del voto di laurea dal sesso: il valore assunto è solo il 15% del massimo cui può giungere

Correlazione

Per parlare di correlazione e di come misurarla, partiamo da un esempio.

Esempio

Un laboratorio vuole studiare la relazione tra dose di uno stimolante per la crescita e aumento di peso (in dg) riscontrato su alcune cavie. I dati relativi a 7 animali, opportunamente selezionati (stesso sesso, età e peso iniziale) sono:

Dose	.0	1.0	2.0	3.0	4.0	5.0	6.0
Aumento Peso	1.0	1.2	2.0	2.4	3.4	4.9	5.1

Possiamo rappresentare graficamente i dati con uno “scatterplot”

Dalla lettura diretta dei dati si nota che, all'aumentare delle dosi, anche il guadagno in peso tende a crescere.

Dal grafico si nota che i punti che rappresentano la distribuzione doppia tendono a disporsi lungo una retta...

Diciamo allora che, nella distribuzione osservata c'è *concordanza*.

Come possiamo misurare questa caratteristica della distribuzione doppia?

In generale, se consideriamo una distribuzione doppia rispetto a due caratteri quantitativi X e Y,

Unità	X	Y
1	x_1	y_1
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

si è in presenza di **concordanza** se, nel complesso, le 2 variabili crescono o decrescono concordemente. Vi è **discordanza** se al crescere di una variabile l'altra decresce.

In entrambi i casi si parla anche di correlazione lineare.

Come possiamo misurare la correlazione lineare?

Idea: se

$$(x_i - \bar{x}) > 0 \qquad (y_i - \bar{y}) > 0$$

oppure se

$$(x_i - \bar{x}) < 0 \qquad (y_i - \bar{y}) < 0$$

allora

$$(x_i - \bar{x})(y_i - \bar{y}) > 0.$$

Nel caso di concordanza (correlazione lienare positiva) ci aspettiamo che la maggior parte dei prodotti tra scarti $(x_i - \bar{x})(y_i - \bar{y})$ sia positiva.

Quindi, una misura della concordanza è la **covarianza,**

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

ovvero la media del prodotto degli scarti delle x_i e y_i dalle rispettive medie aritmetiche.

Ovviamente:

$\sigma_{xy} > 0 \quad \Rightarrow \quad$ correlazione positiva

$\sigma_{xy} < 0 \quad \Rightarrow \quad$ correlazione negativa

$\sigma_{xy} = 0 \quad \Rightarrow \quad$ correlazione nulla o indifferenzapositiva

Nell'esempio, si ha che

$$\sigma_{xy} = 3.014.$$

OSSERVAZIONI

- La covarianza è un indice assoluto. Si mostra (disuguaglianza di Cauchy) che

$$-\sigma_x \sigma_y \leq \sigma_{xy} \leq \sigma_x \sigma_y$$

e quindi l'indice relativo di correlazione è

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

che si chiama *coefficiente di correlazione di Bravais*.

- L'uguagliaza nella disuguaglianza di Cauchy si ha solo se in presenza di una relazione lineare tra X e Y, cioè se

$$Y = aX + b.$$

In questo caso si può avere $r = -1$ oppure $r = +1$. Quindi:

$r = +1 \qquad \Rightarrow \qquad$ perfetta relazione lineare positiva tra X e Y

$r = -1 \quad \Rightarrow \quad$ perfetta relazione lineare negativa tra X e Y

$r = 0 \quad \Rightarrow \quad$ assenza di relazione lineare tra X e Y.

- Nel caso di distribuzioni doppie di frequenze, si ha:

$$r = \frac{\sum_{i=1}^{s}\sum_{j=1}^{t}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{\sqrt{\sum_{i=1}^{s}(x_i - \bar{x})^2 n_{i0}\sum_{j=1}^{t}(y_j - \bar{y})^2 n_{0j}}}$$

Nell'esempio, $r \approx 0.97$.

IMPORTANTE:

connessione nulla ($\phi^2 = 0$) implica assenza di correlazione lineare. Non è vero il viceversa.

Vediamo.

$$\phi^2 = 0 \qquad \Rightarrow \qquad r = 0.$$

Infatti:

$$\phi^2 = 0 \qquad \Rightarrow \qquad n_{ij} = \bar{n}_{ij} = \frac{n_{i0} n_{0j}}{n} \qquad \Rightarrow \qquad \sigma_{xy} = 0$$

Viceversa, l'incorrelazione non implica che i caratteri siano indipendenti. Ad esempio, se consideriamo la distribuzione doppia

X	-2	-1	1	2
Y	8	2	2	8

è semplice verificare che $r = 0$ ma che $\phi^2 \neq 0$. Infatti $Y = 2X^2$!!

Regressione Lineare

Il coefficiente di correlazione r fornisce una misura complessiva del legame di dipendenza lineare tra X e Y.

In generale, le variabili statistiche statistiche non sono legate da leggi matematiche esatte.

Tuttavia è utile cercare di individuare una funzione matematica che descriva bene la relazione esistente tra le due variabili.

Vogliamo cioè cercare una funzione matematica che metta in relazione una variabile dipendente Y a una variabile indipendente, X.

Il più semplice esempio di modello matematico che mette in relazione 2 variabili è la retta

$$Y = c + mX.$$

Poichè però i punti della distribuzione osservata (x_i, y_i) non sono

allineati, l'idea è di trovare una retta passante il più vicino possibile ai punti osservati.

Retta di Regresione e Metodo dei Minimi Quadrati

$$\hat{y} = b_0 + b_1 X$$

I coefficienti b_0 e b_1 vengono determinati in modo che, nell'insieme, siano piccole le differenze tra i valori y_i osservati e quelli teorici

$$\hat{y}_i = b_0 + b_1 x_i.$$

Determiniamo cioè b_0 e b_1 in modo tale che la quantità

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

sia minima.

Per trovare i valori di b_0 e b_1 che rendono minima

$$\sum_{i=1}^{n}[y_i - b_0 - b_1 x_i]^2$$

calcoliamo le derivate parziali rispetto a b_0 e b_1 ed uguagliamole a 0:

$$\begin{cases} -2\sum_{i=1}^{n}[y_i - b_0 - b_1 x_i] = 0 \\ -2\sum_{i=1}^{n} x_i[y_i - b_0 - b_1 x_i] = 0 \end{cases}$$

Dobbiamo ora cercare la soluzione rispetto a b_0 e b_1 del sistema di equazioni precedenti.

Dalla prima abbiamo

$$\sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

da cui

$$b_0 = \bar{y} - b_1 \bar{x}$$

con

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad\qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Sostituendo b_0 nella seconda equazione si ha

$$\sum_{i=1}^{n} x_i [y_i - \bar{y} + b_1 \bar{x} - b_1 x_i] = 0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - b_1 [\sum_{i=1}^{n} x_i^2 - n\bar{x}^2] = 0$$

da cui

$$b_1 = \frac{n\bar{y}\bar{x} - \sum x_i y_i}{n\bar{x}^2 - \sum x_i^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

Quindi:

$$b_1 = r \times \frac{\sigma_y}{\sigma_x}.$$

N.B.: r, $\sigma_{x,y}$ e b_1 hanno lo stesso segno.

Il coefficiente di regressione b_1 esprime di quanto varia Y al variare di una unità di misura di X.

Inoltre

$$b_1 = 0 \quad \Rightarrow \quad \text{indifferenza tra } X \text{ e } Y$$

Previsione

Possiamo usare la retta di regressione per stimare un valore della variabile Y in corrispondenza di un valore x_0 della variabile X.

Semplicemente

$$\hat{y}_0 = b_0 + b_1 x_0.$$

Misura della Bontà di Accostamento

Idea: vorremmo che le differenze tra i valori y_i osservati e quelli stimati dalla retta, $\hat{y}_i$, siano basse.

Definiamo il *residuo* relativo all'osservazione (x_i, y_i) e al valore $\hat{y}_i = b_0 + b_1 x_i$

$$e_i = y_i - \hat{y}_i \qquad i = 1, \ldots, n.$$

Per sintetizzare gli n residui in un singolo valore:

$$D_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Possiamo costruire un indice relativo che misuri la bontà dell'accostamento della retta di regressione ai punti osservati partendo da D_e.

Scomposizione Devianza Totale

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\bar{y}_i - \bar{y})^2$$

ovvero

$$D_t = D_e + D_r.$$

Una intuitiva misura della bontà di accostamento della retta ai dati è quindi

$$R^2 = \frac{D_r}{D_t} = 1 - \frac{D_e}{D_t}.$$

Possiamo mostrare facilmente che

$$R^2 = r^2.$$

Quindi: il quadrato del coefficiente di correlazione di Bravais fornisce una misura della bontà di accostamento della retta di regressione ai

dati osservati. (Ricordarsi che $r^2 = 1$ se e solo se i punti (x_i, y_i) sono allineati).

Per verificare quanto detto, basta osservare che

$$D_e = \sum_{i=1}^{n} e_i^2 = n\sigma_y^2(1 - r^2)$$

e quindi che

$$R^2 = \frac{D_r}{D_t} = 1 - \frac{D_e}{D_t} = 1 - \frac{n\sigma_y^2(1 - r^2)}{n\sigma_y^2} = r^2.$$

L'indice R^2 prende nome di *indice di determinazione*

Esempio. I punteggi GPA di 10 studenti, relativi all'high school e al college sono, rispettivamente

$$X \qquad 2.7, 3.1, 2.1, 3.2, 2.4, 3.4, 2.6, 2, 3.1, 2.5$$

$$Y \qquad 2.2, 2.8, 2.4, 3.8, 1.9, 3.5, 3.1, 1.4, 3.4, 2.5$$

È lecito usare una retta di regressione per studiare la relazione tra X e Y?

Dal grafico sembra OK

Inoltre abbiamo che $r = 0.8439$.

Si ottiene

$$b_1 = 1.347 \qquad b_1 = -.950$$

e quindi

$$\hat{y} = -.950 + 1.347x.$$

Dall'equazione della retta di regressione possiamo trovare i valori teorici $\hat{y}_i$ e i residui e_i:

x_i	y_i	$\hat{y}_i$	e_i
2.7	2.2	2.687	-.487
3.1	2.8	3.225	-.425
$\vdots$	$\vdots$	$\vdots$	$\vdots$
2.5	2.5	2.417	.083

La misura della bontà di accostamento è $R^2 = 0.712$ che indica un discreto accostamento della retta ai dati: il 71.2% della variabilità complessiva dei dati è "spiegato" dalla relazione lineare tra X e Y, cioè dal modello di regressione lineare considerato.

Se vogliamo "prevedere" il GPA che otterrebbe al College uno studente con GPA pari a 3.5 all'High School:

$$-.950 + 1.347 * 3.5 = 3.7645$$

OSSERVAZIONI

- L'accuratezza della previsione dipende da quanto è buono il modello.
- Attenzione a previsioni con x_0 lontano dai valori osservati