

Statistica descrittiva bivariata

si occupa di indagare le relazioni che intercorrono fra due caratteri rilevati sullo stesso collettivo di unità statistiche

Per esempio:

- relazione fra livello di soddisfazione di un cliente (qualitativo) e la sua zona residenziale (qualitativo)
- relazione fra provincia di residenza (qualitativo) e livello di reddito (quantitativo)
- relazione fra voto di maturità (quantitativo) e voto all'esame di statistica (quantitativo)

Vedremo come:

- rappresentare graficamente due caratteri congiuntamente
- costruire tabelle a doppia entrata
- studiare possibili legami esistenti fra due caratteri

Rappresentare graficamente due caratteri

Diagramma di dispersione: si usa quando almeno uno dei due caratteri oggetto di studio è continuo o comunque si è presentato con un **elevato** numero di modalità distinte

Diagramma a bolle: si usa quando entrambi i caratteri oggetto di studio sono qualitativi o si sono presentati con un numero **basso** di modalità distinte così che ci sono coppie di modalità che compaiono più volte nella rilevazione.

Il raggio della bolla corrispondente ad una certa coppia di modalità è proporzionale alla frequenza congiunta (relativa o assoluta) con cui quella coppia di modalità si è presentata

Distribuzione doppia unitaria: consiste nell'elenco delle modalità dei due caratteri osservate per ciascuna delle unità statistiche

Esempio:

<i>Nome</i>	<i>Sesso</i>	<i>Regione nascita</i>
<i>M.Rossi</i>	<i>M</i>	<i>Lombardia</i>
<i>A.Bianchi</i>	<i>F</i>	<i>Calabria</i>
<i>G.Gini</i>	<i>M</i>	<i>Piemonte</i>
<i>A.Verdi</i>	<i>M</i>	<i>Lombardia</i>

Se avessi osservato ciascun carattere distintamente avrei avute due distribuzioni *marginali* unitarie e poi passo alle distribuzioni di frequenza:

$$\text{Sesso} = \begin{cases} M & F \\ 3 & 1 \end{cases}$$

$$\text{Regione} = \begin{cases} \text{Lombardia} & \text{Calabria} & \text{Piemonte} \\ 2 & 1 & 1 \end{cases}$$

Tabelle a doppia entrata

Supponiamo di voler sintetizzare in una tabella il risultato di una rilevazione statistica in cui abbiamo analizzato il carattere X ed il carattere Y riportando le **frequenze relative congiunte**

Il carattere X ha presentato le seguenti modalità distinte:

$$X_1, X_2, \dots, X_i, \dots, X_r$$

Il carattere Y ha presentato le seguenti modalità distinte:

$$Y_1, Y_2, \dots, Y_j, \dots, Y_c$$

Nel'esempio di prima ponendo $X = \text{Sesso}$ abbiamo che $r = 2$ e $c = 3$

Costruiamo allora una **tabella a doppia entrata** con h righe e k colonne. In ciascuna casella della tabella riportiamo la frequenza con cui la corrispondente coppia di modalità è stata osservata:

X/Y	y_1	y_2	\dots	y_j	\dots	y_c
x_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1c}
x_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2c}
\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{ic}
\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_r	n_{r1}	n_{r2}	\dots	n_{rj}	\dots	n_{rc}

dove abbiamo indicato con n_{ij} la frequenza assoluta con cui la modalità x_i si è presentata congiuntamente alla modalità y_j

$$n_{ij} = \text{Fr. Ass.}(X = x_i, Y = y_j)$$

Distribuzioni marginali

dalla tabella a doppia entrata possiamo ricavare le distribuzioni marginali dei due caratteri cioè le distribuzioni che avremmo osservato se avessimo rilevato ciascun carattere singolarmente

La **distribuzione marginale di X** si ottiene **sommando per riga**

La **distribuzione marginale di Y** si ottiene **sommando per colonna**

Sommando invece tutte le frequenze assolute congiunte otteniamo il numero totale delle osservazioni

$$\sum_{i=1}^r \sum_{j=1}^c n_{ij} = n$$

Dalla tabella delle frequenze **assolute** si ricava la tabella delle frequenze **relative** semplicemente dividendo per il numero totale delle osservazioni. Abbiamo che

$$\sum_j^c f_{ij} = f_{i.} = Fr(X = x_i)$$

$$\sum_i^r f_{ij} = f_{.j} = Fr(Y = y_j)$$

$$\sum_i^r \sum_j^c f_{ij} = 1$$

formule simili alle prime due sono valide anche per le frequenze assolute:

$$\sum_j^c n_{ij} = n_{i.}$$

$$\sum_i^r n_{ij} = n_{.j}$$

ESEMPIO: la seguente tabella descrive la distribuzione del numero di addetti (classificati in 3 classi), ripartita per area geografica, del fatturato mensile di 1000 aziende:

area/addetti	(0 – 10]	(10 – 50]	(50 – 100]
nord	0.1	0.1	0.2
centro	0.15	0.05	0.05
sud	0.15	0.05	0.15

La distribuzione marginale del numero di addetti è la seguente:

$$\text{n. addetti} = \begin{cases} (0 - 10] & (10 - 50] & (50 - 100] \\ 0.4 & 0.2 & 0.4 \end{cases}$$

La distribuzione marginale per area geografica è la seguente:

$$\text{area} = \begin{cases} \text{nord} & \text{centro} & \text{sud} \\ 0.4 & 0.25 & 0.35 \end{cases}$$

Distribuzioni subordinante o condizionate

per rispondere a domande del tipo: *fra le aziende del nord sono più numerose quelle piccole (con al più 10 addetti) o quelle grandi (con più di 50 addetti)* bisogna considerare la sottopopolazione della aziende del nord (400 aziende in tutto). Questo equivale a lavorare con distribuzioni subordinate:

$$\begin{aligned} Fr(Y > 50|X = \text{nord}) &= \frac{Fr(Y > 50, X = \text{nord})}{Fr(X = \text{nord})} = \\ &= \frac{0.2}{0.4} = 0.5 \end{aligned}$$

$$\begin{aligned} Fr(Y \leq 10|X = \text{nord}) &= \frac{Fr(Y \leq 10, X = \text{nord})}{Fr(X = \text{nord})} = \\ &= \frac{0.1}{0.4} = 0.25 \end{aligned}$$

Analisi dell'associazione fra due caratteri

Ha lo scopo di stabilire se un carattere (X) ha influenza sull'altro Y .

Se X non ha alcuna influenza su Y si dice che Y è **statisticamente indipendente** (o indipendente) da X .

Se due caratteri non sono statisticamente indipendenti si dicono **connessi**.

Si ha **perfetta connessione di Y da X** se per ogni modalità di X (area) si ha una frequenza congiunta positiva per una sola modalità di Y (numero addetti) cioè nota la modalità con cui si è presentato X posso determinare univocamente la modalità con cui si è presentato per quel soggetto il carattere Y (esempio: se avessi avuto che tutte le imprese del sud avevano meno di 10 dipendenti e tutte quelle al centro e al nord più di 50 addetti)

L'indipendenza statistica

due caratteri X ed Y si dicono statisticamente indipendenti se

$$f(x_i, y_j) = f(x_i)f(y_j) \quad \forall i, j$$

queste sono altre due definizioni equivalenti alla precedente:

$$f(x_i|y_j) = f(x_i) \quad \forall i, j$$

$$f(y_j|x_i) = f(y_j) \quad \forall i, j$$

L'indipendenza statistica è un **concetto simmetrico**: se X è statisticamente indipendente da Y anche Y è statisticamente indipendente da X e viceversa

La connessione:

se due caratteri non sono statisticamente indipendenti si dicono connessi.

Cerchiamo ora un criterio per misurare la forza della connessione fra due caratteri. Per fare questo possiamo vedere come la distribuzione congiunta osservata si discosta dalla distribuzione che avremmo osservato nel caso di indipendenza statistica.

Questo equivale a confrontare due tabelle a doppia entrata, quella osservata appunto e quella che si ottiene mettendo in ciascuna delle caselle il prodotto delle corrispondenti frequenze marginali

Consideriamo per esempio la seguente tabella

X/Y	-1	1
-1	0.1	0.4
0	0.05	0.3
1	0.05	0.1

la tabella che si avrebbe nel caso i due caratteri fossero indipendenti è :

X/Y	-1	1
-1	0.1	0.4
0	0.07	0.28
1	0.03	0.12

La precedente tabella è ottenuta ponendo in ciascuna casella la frequenza relativa ottenuta facendo il prodotto delle corrispondenti frequenze relative marginali cioè ponendo nella casella (i, j) la frequenza:

$$f(x_i, y_j) = f(x_i) \times f(y_j)$$

.

Notiamo che per la tabella originale le frequenze congiunte sulla prima riga sono effettivamente uguali al prodotto delle corrispondenti marginali ma questo non basta per concludere che i due caratteri siano statisticamente indipendenti, occorre continuare e verificare che lo stesso valga per tutte le righe, cosa che non è vera per la nostra tabella. Concludiamo quindi che i due caratteri *non* sono statisticamente indipendenti ma connessi.

Per valutare il grado di connessione si costruisce la tabella delle contingenze facendo la differenza fra la tabella originale e quella di indipendenza. Ricorda che la contingenza fra x_i ed y_j vale

$$c(x_i, y_j) = f(x_i, y_j) - f(x_i) \times f(y_j)$$

otteniamo quindi la seguente tabella:

X/Y	-1	1
-1	0	0
0	-0.02	0.02
1	0.02	-0.02

Nota bene: la somma delle contingenze vale sempre zero, questo può essere usato per controllare di avere fatto i conti giusti!

Dalla tabella delle contingenze si fa il quadrato e si procede poi al calcolo dell'indice di contingenza:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{c(x_i, y_j)^2}{f(x_i) \times f(y_j)}$$

nel nostro caso abbiamo:

$$\chi^2 = \frac{0.0004}{0.07} + \frac{0.0004}{0.28} + \frac{0.0004}{0.03} + \frac{0.0004}{0.12} = 0.0024$$

L'indice di contingenza relativo si ottiene rapportando l'indice di contingenza assoluto al suo valore massimo.

Si dimostra che

$$0 \leq \chi^2 \leq \min[r - 1, c - 1]$$

quindi l'**indice di contingenza relativo** è :

$$\tilde{\chi}^2 = \frac{\chi^2}{\min[r - 1, c - 1]}$$

e risulta:

$$0 \leq \tilde{\chi}^2 \leq 1$$

con **$\tilde{\chi}^2 = 0$** se e solo se c'è indipendenza statistica fra X ed Y

mentre **$\tilde{\chi}^2 = 1$** se e solo se c'è perfetta connessione (unilaterale o bilaterale) fra X ed Y

Nel nostro esempio abbiamo che:

$$\begin{aligned}\tilde{\chi}^2 &= \frac{\chi^2}{\min[r-1, c-1]} = \frac{\chi^2}{\min[3-1, 2-1]} \\ &= \frac{\chi^2}{\min[2, 1]} = \frac{\chi^2}{1} = 0.0024.\end{aligned}$$

concludo che questo valore è molto basso indice del fatto che le due variabili sono molto vicine ad una situazione di indipendenza statistica.

Si parla di **perfetta connessione bilaterale** fra X ed Y se la tabella ha lo stesso numero di righe e colonne e presenta un unico valore positivo per ogni riga e per ogni colonna per esempio:

X/Y	1	-2	2
0	0	0	0,2
1	0.5	0	0
-1	0	0.3	0

in questo caso, nota la modalità con cui si è presentato il carattere X riesco a determinare univocamente la modalità con cui si è presentato il carattere Y .

Si parla di **perfetta connessione unilaterale di X da Y** se nella tabella ho un numero di colonne *maggiore* del numero di righe inoltre la tabella deve presentare un unico valore positivo per ogni *colonna* come per esempio succede nella seguente tabella:

X/Y	1	-2	2
0	0	0,3	0,2
1	0.5	0	0

Nella precedente tabella, se conosco che per un certo soggetto il valore di Y è risultato essere, per esempio, pari a 2 posso immediatamente indovinare che il valore di X per quello stesso soggetto non può che essere stato pari a zero quindi, noto il valore di Y ricavo univocamente il valore di X . Non è però vero il contrario, per esempio se sapessi che $X = 0$ per quel soggetto il valore di Y potrebbe essere sia -2 che +2.

Si parla invece di **perfetta connessione unilaterale di Y da X** se nella tabella ho un numero di colonne *minore* del numero di righe, inoltre la tabella deve presentare un unico valore positivo per ogni *riga* come per esempio succede nella seguente tabella:

X/Y	1	2
-1	0	0,3
0	0.5	0
+1	0	0.2

Nella precedente tabella, se conosco che per un certo soggetto il valore di X è risultato essere, per esempio, pari a zero posso immediatamente indovinare che il valore di Y per quello stesso soggetto non può che essere stato pari ad uno quindi, noto il valore di X ricavo univocamente il valore di Y . Non è però vero il contrario, per esempio se sapessi che $Y = 2$ per quel soggetto il valore di X potrebbe essere sia -1 che +1.

ESERCIZIO

Un prestigiatore afferma che, lanciando due monetine, è in grado di pilotare il risultato facendo in modo che presentino entrambe la stessa faccia. Non sempre però la sua forza del pensiero è sufficientemente forte e quindi non sempre l'esperimento riesce. Osservo il prestigiatore ripetere l'esperimento del lancio delle due monetine 10 volte ed ottengo i seguenti risultati:

Prima moneta	<i>T</i>	<i>C</i>	<i>C</i>	<i>T</i>	<i>C</i>	<i>T</i>	<i>C</i>	<i>T</i>	<i>C</i>	<i>T</i>
Seconda moneta	<i>C</i>	<i>C</i>	<i>C</i>	<i>T</i>	<i>C</i>	<i>C</i>	<i>C</i>	<i>C</i>	<i>C</i>	<i>T</i>

Riclassificate le osservazioni ottenute indicando con *X* il risultato del lancio della prima moneta e con *Y* il risultato del lancio della seconda. Indicate inoltre con 0 il risultato "esce testa" e con 1 il risultato "esce croce".

RISPOSTA:

<i>X</i>	0	1	1	0	1	0	1	0	1	0
<i>Y</i>	1	1	1	0	1	1	1	1	1	0

Posso dire che i lanci delle due monete sono statisticamente indipendenti e concludere che quindi il prestigiatore è in realtà un truffatore oppure l'esperimento mi fa pensare che effettivamente il prestigiatore abbia dei poteri paranormali?

RISPOSTA: posso costruire una tabella a doppia entrata e studiare se esiste connessione fra i lanci delle due monetine (o assenza di connessione, ossia indipendenza statistica):

	Y	0	1
X			
0		2	3
1		0	5

da cui ricavo la tabella delle frequenze relative congiunte:

	Y	0	1
X			
0		0.2	0.3
1		0	0.5

e le due distribuzioni marginali: la distribuzione marginale di X è la distribuzione che avrei ottenuto se, durante i lanci, mi fossi dimenticato completamente della seconda moneta:

$$X = \begin{cases} 0 & 1 \\ 0.5 & 0.5 \end{cases}$$

La distribuzione marginale di Y è la distribuzione che avrei ottenuto se, durante i lanci, mi fossi dimenticato completamente della prima moneta:

$$Y = \begin{cases} 0 & 1 \\ 0.2 & 0.8 \end{cases}$$

Sembra che la prima moneta sia bilanciata mentre ho dei dubbi sul fatto che la seconda moneta sia invece truccata a favore dell'evento "esce croce". Per verificare i miei dubbi potrei chiedere al prestigiatore di effettuare più lanci ma devo comunque poi seguire il corso di statistica inferenziale per imparare come *verificare delle ipotesi statistiche sulla base di un campione di osservazioni!*

Sembra inoltre che i due eventi non siano statisticamente indipendenti ... in quanto non è vero che le frequenze congiunte risultano esattamente uguali al prodotto delle rispettive

marginali. Ma significa allora che il prestigiatore ha effettivamente dei poteri para-normali?!? In realtà non possiamo rispondere a questa domanda in quanto è comunque molto difficile che, nella realtà, due caratteri, pur essendo, di fatto, statisticamente indipendenti, diano luogo ad una tabella a doppia entrata in cui tutte le frequenze congiunte siano esattamente uguali al prodotto delle rispettive marginali: provate voi stessi a prendere due monete, lanciarle per 10 volte e costruite la tabella a doppia entrata. Ottenete una tabella in cui i due lanci risultano statisticamente indipendenti? Verosimilmente no! Ma questo non basta per affermare che siete dei maghi ... basta forse per motivarvi a studiare un pò più di statistica (la statistica inferenziale)

Un altro modo per verificare l'esistenza o meno di indipendenza statistica è confrontare le distribuzioni marginali con quelle subordinate. Se i lanci delle due monetine fossero indipendenti, l'esito del primo lancio non deve modificare la probabilità che esca testa o croce nel secondo lancio. In termini statistici questo significa che la probabilità di avere, per esempio, testa sulla seconda monetina deve essere la stessa “*independentemente*” dall'esito del primo lancio cioè dovremmo avere che

$$Fr(\text{testa seconda moneta}|\text{testa prima moneta}) =$$

$$Fr(\text{testa seconda moneta}|\text{croce prima moneta}) =$$

$$Fr(\text{testa seconda moneta})$$

ovvero

$$Fr(Y = 0|X = 0) = Fr(Y = 0|X = 1) = Fr(Y = 0)$$

e, in modo simile, dovrà essere

$$Fr(Y = 1|X = 1) = Fr(Y = 1|X = 0) = Fr(Y = 1).$$

Nell'esperimento effettuato risulta invece che

$$Fr(Y = 0|X = 0) = 0.2/0.5 = 0.4$$

$$Fr(Y = 0|X = 0) > Fr(Y = 0) = 0.2$$

cioè , in base ai dati dell'esperimento, la probabilità di avere testa sulla seconda moneta raddoppia se sulla prima moneta ho ottenuto testa cioè *"condizionatamente"* al fatto che sulla prima moneta abbia ottenuto testa. In modo simile abbia che

$$Fr(Y = 1|X = 1) = 0.5/0.5 = 1$$

$$1 > Fr(Y = 1) = 0.3 + 0.5 = 0.8$$

Sempre per studiare l'indipendenza o meno fra i due caratteri, posso infine calcolare l'indice di connessione, $\tilde{\chi}^2$. La tabella di indipendenza è

X/Y	0	1
0	0.1	0.4
1	0.1	0.4

La tabella delle contingenze è

X/Y	0	1
0	0.1	-0.1
1	-0.1	0.1

Quindi abbiamo che

$$\begin{aligned} \chi^2 &= \tilde{\chi}^2 = \\ &= 0.1^2/0.1 + 0.1^2/0.4 + 0.1^2/0.1 + 0.1^2/0.4 = \\ &= 0.25 \end{aligned}$$

Anche in questo caso il valore è diverso da zero indicando che i due lanci sono connessi. Ma questo valore è *sufficientemente diverso da zero* per concludere che il prestigiatore ha poteri paranormali? La risposta al prossimo corso!