Azzolini Riccardo 2020-05-26

Covarianza e correlazione

1 Covarianza

Siano X e Y due variabili aleatorie qualunque. Nel caso in cui X e Y sono indipendenti, si è già data una formula per la varianza di $X + Y$:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \qquad \text{se } X \text{ e } Y \text{ indipendenti}$$

Adesso, si vuole invece determinare una formula per la varianza di $X + Y$ in generale, che valga anche se X e Y non sono indipendenti:

$$\begin{aligned}
&\mathrm{Var}(X+Y)\\
&\quad \text{definizione di varianza:}\\
&\quad = E\Big[\big((X+Y) - E(X+Y)\big)^2\Big]\\
&\quad \text{linearità del valore medio:}\\
&\quad = E\Big[\big(X+Y-E(X)-E(Y)\big)^2\Big]\\
&\quad = E\Big[\big((X-E(X)) + (Y-E(Y))\big)^2\Big]\\
&\quad \text{sviluppo del quadrato:}\\
&\quad = E\big[(X-E(X))^2 + (Y-E(Y))^2 + 2(X-E(X))(Y-E(Y))\big]\\
&\quad \text{linearità del valore medio:}\\
&\quad = E\big[(X-E(X))^2\big] + E\big[(Y-E(Y))^2\big] + 2E\big[(X-E(X))(Y-E(Y))\big]\\
&\quad \text{definizione di varianza:}\\
&\quad = \mathrm{Var}(X) + \mathrm{Var}(Y) + \underbrace{2E\big[(X-E(X))(Y-E(Y))\big]}_{\mathrm{Cov}(X,Y)}
\end{aligned}$$

La quantità $E[(X - E(X))(Y - E(Y))]$ che compare in questa formula prende il nome di **covarianza** di X e Y, $\mathrm{Cov}(X, Y)$.

Come per la varianza, anche per il calcolo della covarianza esiste una formula alternativa, spesso più comoda:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E\big[(X-E(X))(Y-E(Y))\big]\\
&= E\big[XY - XE(Y) - E(X)Y + E(X)E(Y)\big]\\
&= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)\\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

1.1 Varianza della somma di n variabili aleatorie

Come caso particolare, la covarianza $\mathrm{Cov}(X, X)$ di una variabile X con se stessa è semplicemente la varianza di X:

$$\mathrm{Cov}(X, X) = E(XX) - E(X)E(X) = E(X^2) - (E(X))^2 = \mathrm{Var}(X)$$

Quest'osservazione è utile per generalizzare la formula della varianza di una somma di due variabili al caso di n variabili:

$$\begin{aligned}
&\mathrm{Var}(X_1 + \cdots + X_n) \\
&\quad = E\Big[\big((X_1 + \cdots + X_n) - E(X_1 + \cdots + X_n)\big)^2\Big] && \text{(definizione di varianza)} \\
&\quad = E\Big[\big((X_1 - E(X_1)) + \cdots + (X_n - E(X_n))\big)^2\Big] && \text{(linearità del valore medio)} \\
&\quad = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - E(X_i))(X_j - E(X_j))\right] && \text{(sviluppo del quadrato)} \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{n} E\big[(X_i - E(X_i))(X_j - E(X_j))\big] && \text{(linearità del valore medio)} \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) && \text{(definizione di covarianza)} \\
&\quad = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \mathrm{Cov}(X_i, X_j) && (\mathrm{Cov}(X_i, X_j) = \mathrm{Var}(X_i) \text{ se } i = j)
\end{aligned}$$

1.2 Covarianza e indipendenza

Se due variabili aleatorie X e Y sono indipendenti, allora, ricordando che in questo caso $E(XY) = E(X)E(Y)$, la loro covarianza è 0:

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

Viceversa, *non è in generale vero* che, se $\mathrm{Cov}(X, Y) = 0$, allora X e Y sono indipendenti. In ogni caso, se $\mathrm{Cov}(X, Y) = 0$, X e Y si dicono **scorrelate**.

Osservazioni:

- Dalle formule generali per la varianza di una somma si possono ritrovare quelle per variabili indipendenti:

$$\begin{aligned}\operatorname{Var}(X+Y) &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\overbrace{\operatorname{Cov}(X,Y)}^{0} = \operatorname{Var}(X) + \operatorname{Var}(Y) \\ \operatorname{Var}(X_1 + \cdots + X_n) &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \underbrace{\operatorname{Cov}(X_i, X_j)}_{0} = \sum_{i=1}^{n} \operatorname{Var}(X_i)\end{aligned}$$

- Il fatto che $\operatorname{Cov}(X,Y) = 0$ non indichi l'indipendenza non è solitamente un problema. Tipicamente, infatti, l'indipendenza è l'ipotesi nulla: se $\operatorname{Cov}(X,Y) > 0$ la si può rifiutare (perché allora X e Y sono sicuramente *non* indipendenti), mentre se $\operatorname{Cov}(X,Y) = 0$ non la si rifiuta (e ciò, come detto in precedenza, non significa accettarla, cioè affermare che X e Y sono indipendenti).

2 Coefficiente di correlazione

Quando la covarianza viene utilizzata per misurare la correlazione tra due variabili X e Y, essa risulta "scomoda" perché i valori di $\operatorname{Cov}(X,Y)$ dipendono dalla grandezza dei valori di X e Y, quindi non è immediato capire quali valori della covarianza debbano essere considerati "grandi" (cioè corrispondenti a una correlazione forte) e quali "piccoli" (correlazione debole).

Perciò, è utile considerare invece il **coefficiente di correlazione** delle due variabili X e Y, indicato con $\rho_{X,Y}$ o r e definito come segue:

$$\rho_{X,Y} = r = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Si può dimostrare che esso assume sempre valori compresi tra -1 e 1, ovvero $|\rho_{X,Y}| \leq 1$. Allora, la correlazione tra X e Y è tanto più forte quanto più il valore assoluto di $\rho_{X,Y}$ è vicino a 1, e, in particolare:

- valori positivi indicano una *correlazione positiva*: intuitivamente, quando X ha valori superiori alla sua media $E(X)$, allora anche Y tende ad assumere valori superiori a $E(Y)$, e quando $X < E(X)$ tende a essere anche $Y < E(Y)$, dunque $X - E(X)$ e $Y - E(Y)$ tendono ad avere lo stesso segno,

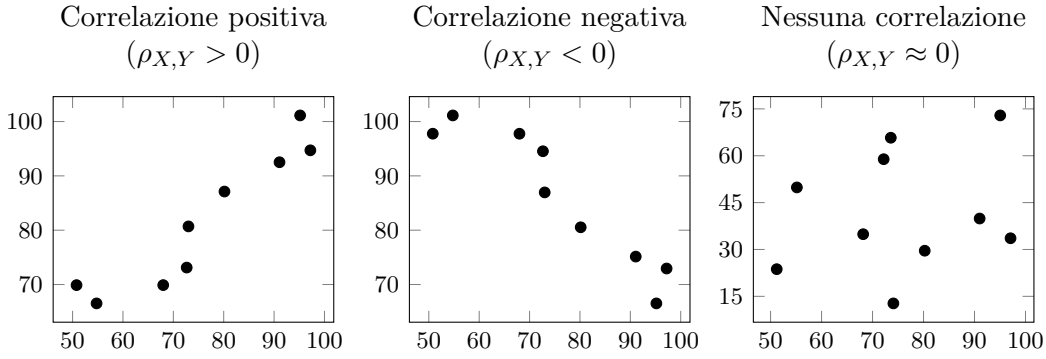
  $$(X - E(X))(Y - E(Y)) > 0$$

  da cui segue $\operatorname{Cov}(X,Y) > 0$, e infine $\rho_{X,Y} > 0$;

- valori negativi indicano una *correlazione negativa*: quando $X > E(X)$, si ha solitamente $Y < E(Y)$, e viceversa, dunque tende a essere $(X - E(X))(Y - E(Y)) < 0$, che implica $\operatorname{Cov}(X,Y) < 0$, e quindi $\rho_{X,Y} < 0$;

- valori prossimi a 0 indicano che non c'è nessuna correlazione.

Degli esempi di questi tre casi sono mostrati nei seguenti **diagrammi di dispersione** (**scatter plot**):



2.1 Problema: estrazione di due palline

Problema: Un'urna contiene b palline bianche e r rosse. Da essa, si estraggono senza rimpiazzo due palline. Date le variabili aleatorie

$$X_1 = \begin{cases} 1 & \text{se la prima pallina estratta è rossa} \\ 0 & \text{altrimenti} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{se la seconda pallina estratta è rossa} \\ 0 & \text{altrimenti} \end{cases}$$

calcolare il coefficiente di correlazione di X_1 e X_2.

Soluzione: Prima di tutto, per poter calcolare la covarianza mediante la formula

$$\mathrm{Cov}(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2)$$

bisogna determinare i valori medi che compaiono in tale formula:

- Siccome X_1 e X_2 sono variabili di Bernoulli (possono assumere solo i valori 0 e 1), lo è anche X_1X_2: in particolare, $X_1X_2 = 1$ se e solo se $X_1 = 1$ e $X_2 = 1$. Allora, ricordando che il valore medio di una Bernoulli è dato dalla probabilità che essa assuma il valore 1, si ha:

$$\begin{aligned} E(X_1X_2) &= P\{X_1X_2 = 1\} \\ &= P\{X_1 = 1,\, X_2 = 1\} \\ &= P\{X_2 = 1 \mid X_1 = 1\}P\{X_1 = 1\} \\ &= \frac{r-1}{b+r-1} \cdot \frac{r}{b+r} = \frac{r(r-1)}{(b+r)(b+r-1)} \end{aligned}$$

- Il calcolo di $E(X_1)$ è immediato:

$$E(X_1) = P\{X_1 = 1\} = \frac{r}{b+r}$$

- $E(X_2)$ si ricava con il teorema della probabilità totale:

$$\begin{aligned}
E(X_2) &= P\{X_2 = 1\} \\
&= P\{X_2 = 1,\, X_1 = 1\} + P\{X_2 = 1,\, X_1 = 0\} \\
&= P\{X_2 = 1 \mid X_1 = 1\}P\{X_1 = 1\} + P\{X_2 = 1,\, X_1 = 0\}P\{X_1 = 0\} \\
&= \frac{r-1}{b+r-1} \cdot \frac{r}{b+r} + \frac{r}{b+r-1} \cdot \frac{b}{b+r} \\
&= \frac{r(r-1)+rb}{(b+r)(b+r-1)} \\
&= \frac{r(r-1+b)}{(b+r)(b+r-1)} \\
&= \frac{r}{b+r}
\end{aligned}$$

  È interessante osservare che $P\{X_2 = 1\} = P\{X_1 = 1\}$, ovvero che X_2 ha una densità di probabilità marginale uguale a X_1, al contrario di quanto ci si potrebbe aspettare sapendo che l'estrazione è senza rimpiazzo. Di conseguenza, anche i parametri della distribuzione di X_2 (valore medio, varianza, ecc.) sono uguali a quelli di X_1.

Adesso, si può procedere al calcolo della covarianza:

$$\begin{aligned}
\mathrm{Cov}(X_1, X_2) &= E(X_1X_2) - E(X_1)E(X_2) \\
&= \frac{r(r-1)}{(b+r)(b+r-1)} - \frac{r}{b+r} \cdot \frac{r}{b+r} \\
&= \frac{r}{b+r}\left(\frac{r-1}{b+r-1} - \frac{r}{b+r}\right)
\end{aligned}$$

Poi, sapendo che la varianza di una Bernoulli di parametro p è $p(1-p)$, si ricavano le varianze di X_1 e X_2, che sono uguali perché le due variabili hanno la stessa distribuzione (cioè lo stesso parametro $p = \frac{r}{b+r}$):

$$\begin{aligned}
\mathrm{Var}(X_1) = \mathrm{Var}(X_2) &= \frac{r}{b+r}\left(1 - \frac{r}{b+r}\right) \\
&= \frac{r}{b+r} \cdot \frac{b}{b+r} \\
&= \frac{rb}{(b+r)^2}
\end{aligned}$$

Infine, si inseriscono i dati nella formula del coefficiente di correlazione:

$$\begin{aligned}
\rho_{X_1,X_2} &= \frac{\operatorname{Cov}(X_1,X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}} \\
&= \frac{\dfrac{r}{b+r}\left(\dfrac{r-1}{b+r-1}-\dfrac{r}{b+r}\right)}{\sqrt{\dfrac{rb}{(b+r)^2}\cdot\dfrac{rb}{(b+r)^2}}} \\
&= \frac{r}{b+r}\left(\frac{r-1}{b+r-1}-\frac{r}{b+r}\right)\frac{(b+r)^2}{rb} \\
&= \frac{b+r}{b}\left(\frac{r-1}{b+r-1}-\frac{r}{b+r}\right) \\
&= \frac{b+r}{b}\cdot\frac{(r-1)(b+r)-r(b+r-1)}{(b+r-1)(b+r)} \\
&= \frac{(r-1)(b+r)-r(b+r-1)}{b(b+r-1)} \\
&= \frac{rb+r^2-b-r-rb-r^2+r}{b(b+r-1)} \\
&= \frac{-b}{b(b+r-1)} \\
&= -\frac{1}{b+r-1}
\end{aligned}$$

Osservazione: $\rho_{X_1,X_2} \to 0$ per $n = b+r \to +\infty$. Infatti, intuitivamente, X_1 e X_2 non saranno mai indipendenti, ma, all'aumentare del numero n di palline presenti inizialmente nell'urna, la prima estrazione cambierà sempre meno la composizione dell'urna, rendendo la dipendenza di X_2 da X_1 sempre più debole: come già detto in precedenza, per n grandi un'estrazione senza rimpiazzo è ben approssimata da una senza rimpiazzo (e ciò corrisponde all'approssimazione della distribuzione ipergeometrica con la binomiale).

3 Stimatori per covarianza e coefficiente di correlazione

Per stimare la covarianza $\operatorname{Cov}(X,Y)$ di una popolazione a media incognita a partire da un campione di ampiezza n, è necessario utilizzare il seguente stimatore, che, come la varianza campionaria, è reso non distorto dividendo per $n-1$ anziché per n:

$$S_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$$

Tale valore può poi essere usato, insieme alle varianze campionarie S_X^2 e S_Y^2, per stimare il coefficiente di correlazione:

$$r = \frac{S_{X,Y}}{\sqrt{S_X^2 S_Y^2}}$$

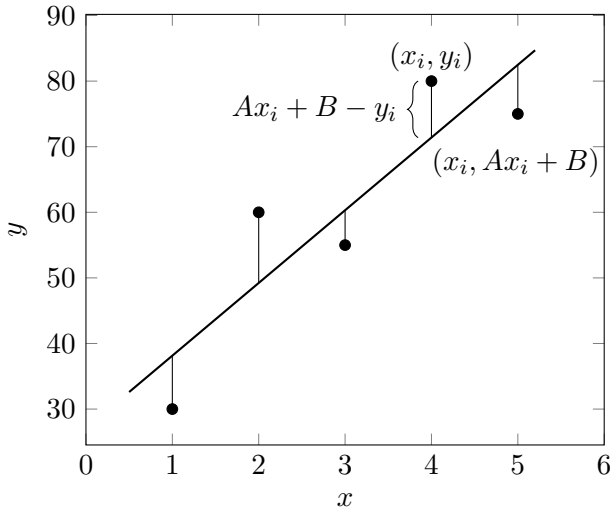
4 Retta di regressione

Dato un campione di ampiezza $n > 1$ estratto da una coppia di variabili aleatorie, è sempre possibile dare una retta, chiamata **retta di regressione**, che interpoli al meglio i punti del campione. Il metodo tipicamente usato è quello di minimizzare la somma degli scarti quadratici: partendo dall'equazione di una generica retta,

$$y = Ax + B$$

si scelgono i valori di A e B per i quali è minima la quantità

$$\sum_{i=1}^{n} (Ax_i + B - y_i)^2$$



Il fatto che questa retta esista sempre implica però che essa non sia necessariamente significativa. Lo strumento che permette di valutare se abbia senso trovare la retta di regressione, cioè se quest'ultima costituirà una buona interpolazione dei dati, è il coefficiente di correlazione: più è forte la correlazione tra i dati, migliore sarà la "bontà" dell'interpolazione.

4.1 Esempio

La seguente tabella riporta i valori ottenuti osservando il tempo che un computer impiega per processare dei dati; x_i è il numero di dati processati, e y_i è il tempo impiegato in secondi:

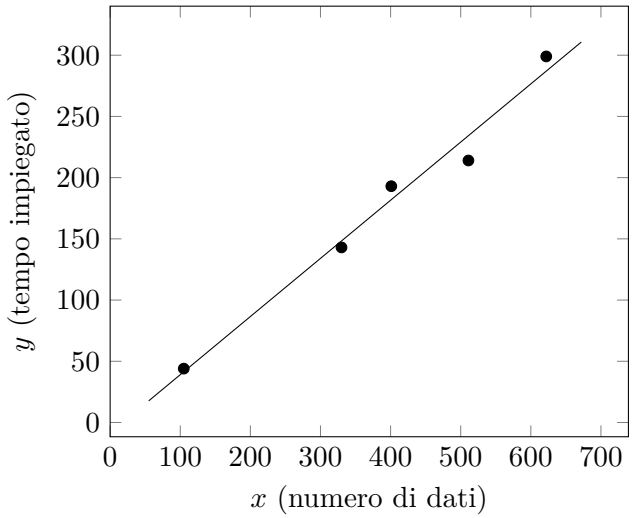
x_i	y_i
105	44
511	214
401	193
622	299
330	143

Come tipicamente accade, anche se il campione è piccolo ($n = 5$), il calcolo del coefficiente di correlazione è piuttosto laborioso. Eseguendolo, ad esempio, mediante un apposito strumento software, si trova che $r = 0.99$, cioè che esiste una forte correlazione (positiva).

Allora, è significativo determinare la retta di regressione. L'equazione della retta che minimizza la somma degli scarti quadratici può essere ricavata, ad esempio, usando ancora opportuni strumenti software, ed è approssimativamente:

$$y = 0.475x - 8.455$$

La rappresentazione grafica del campione e della retta di regressione conferma che quest'ultima costituisce effettivamente una buona interpolazione dei dati:



5 Altre misure di associazione

Il coefficiente di correlazione non è l'unica misura di associazione tra due variabili aleatorie.

Ad esempio, un caso tipico è quello di variabili dicotomiche (che possono assumere solo due valori, ovvero variabili di Bernoulli) relative all'esposizione a un determinato fattore di rischio e al successivo insorgere di una malattia. Per rappresentare tale situazione, si costruisce una **tabella di contingenza**:

	Malati	Non malati	Totale
Esposti	a	b	$a+b$
Non esposti	c	d	$c+d$

Da questa tabella, si può calcolare l'incidenza della malattia tra gli esposti, $a/(a+b)$, e tra i non esposti, $c/(c+d)$. Il rapporto tra queste due incidenze è una misura di associazione chiamata **rischio relativo**:

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

- $RR = 1$ indica l'assenza di un'associazione tra le variabili;
- $RR > 1$ indica un'associazione positiva;
- $RR < 1$ indica un'associazione negativa.

Spesso, però, quando si effettua uno studio, il campione è scelto fissando arbitrariamente un rapporto tra il numero di pazienti malati ("casi") e non ("controlli"): ad esempio, si potrebbero scegliere, per ogni caso, due controlli con le stesse caratteristiche di quest'ultimo (stessa età, residenza, professione, ecc.), e allora si avrebbe $(a+c)/(b+d) = 0.5$. Così, il campione non dà alcuna informazione riguardo alla reale incidenza della malattia nella popolazione, e non è quindi possibile determinare il rischio relativo.

In questa situazione, può invece essere calcolata un'altra misura di associazione: l'**odds ratio**. In generale, se un evento ha probabilità p, si definisce **odds** di tale evento la quantità $p/(1-p)$. L'odds ratio è il rapporto tra l'odds di malattia tra gli esposti al fattore di rischio, a/b, e l'odds di malattia tra i non esposti, c/d:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Anche se non ha un'interpretazione particolarmente intuitiva, l'odds ratio è molto usato in pratica, sia per la sua utilità negli studi caso–controllo (come appena detto), sia perché lo si ritrova in un metodo di regressione chiamato *regressione logistica*.

6 Intervalli di confidenza per la varianza

Dato un campione di ampiezza n estratto da una popolazione normale, si può dimostrare che la statistica

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

(dove S^2 è la varianza campionaria e σ^2 è la varianza della popolazione) ha la distribuzione χ^2 (**chi quadro**) con grado di libertà $\nu = n - 1$.

Essa può essere usata per dare un intervallo di confidenza per la varianza della popolazione, in modo analogo a quanto fatto per la media (dove si usavano invece la distribuzione normale e la t di Student). Per prima cosa, si individuano i valori critici $\chi^2_{1-\alpha/2}$ e $\chi^2_{\alpha/2}$[1] tali che

$$P\left\{\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right\} = 1 - \alpha$$

ovvero tali che valga con probabilità $1 - \alpha$ la disuguaglianza:

$$\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}$$

Risolvendo quest'ultima rispetto a σ^2, e inserendo il valore s^2 della varianza campionaria S^2 per lo specifico campione considerato, si ottiene il seguente intervallo di confidenza per la varianza, con grado di fiducia $(1 - \alpha) \cdot 100$ %:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

7 Test di adattamento dei dati

Le tabelle di contingenza 2×2 possono essere generalizzate estendendole a più classi (corrispondenti a variabili che assumono più di 2 valori):

		Classi				
		1	2	3	...	c
	1	O_{11}	O_{12}	O_{13}	...	O_{1c}
	2	O_{21}	O_{22}	O_{23}	...	O_{2c}
Classi	3	O_{31}	O_{32}	O_{33}	...	O_{3c}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	r	O_{r1}	O_{r2}	O_{r3}	...	O_{rc}

[1] Siccome la distribuzione χ^2 è asimmetrica, questi due valori critici non sono semplicemente uno l'opposto dell'altro, ma devono invece essere letti separatamente dalle tavole.

In una tabella del genere, relativa ai possibili valori assunti di due variabili aleatorie X e Y, ciascun valore O_{ij} è la **frequenza osservata** dell'evento $\{X = i, Y = j\}$. Spesso, ci si aspetta che questi dati seguano una qualche distribuzione teorica, che prevede per ciascuna classe una **frequenza attesa** A_{ij}. Un procedimento che permette di sottoporre a test tale ipotesi è il **test chi-quadro di adattamento**.

Per semplicità, si considera il caso di una singola variabile aleatoria, i cui valori definiscono k classi, ciascuna con frequenza osservata O_i e frequenza attesa A_i (per $i = 1, \dots, k$, appunto). A partire dalle frequenze osservate e attese, si ricava la seguente statistica test, che si dice essere il **chi-quadro** calcolato dal campione:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - A_i)^2}{A_i}$$

come suggerisce il nome, per n sufficientemente grandi[2] si può dimostrare che questa statistica ha approssimativamente la distribuzione χ^2, con grado di libertà $\nu = k-1-m$, dove k è il numero di classi, e m è il numero dei parametri della distribuzione teorica che sono stati stimati servendosi dei dati del campione.

L'ipotesi nulla H_0 è che i dati si adattino alla distribuzione teorica ipotizzata: essa viene rifiutata al livello di significatività α se $\chi^2 > \chi^2_\alpha$, cioè se il chi-quadro calcolato dal campione è maggiore del valore critico corrispondente ad α.

7.1 Test chi-quadro di indipendenza

Il test chi-quadro può essere usato per verificare se due variabili siano o meno indipendenti, ponendo come ipotesi nulla l'indipendenza.

Considerando una tabella di contingenza $r \times c$ come quella mostrata prima, si osserva che essa è di fatto la tabella della densità congiunta (osservata nel campione) delle due variabili. Allora, calcolando i totali di ciascuna riga e colonna si ricavano le densità marginali, e, se e solo se le variabili sono indipendenti, la densità congiunta sarà il prodotto delle marginali. Dunque, le frequenze attese (supponendo che valga H_0, cioè l'indipendenza) sono proprio quelle date da tale prodotto,

$$A_{ij} = \frac{(\text{totale riga } i)(\text{totale colonna } j)}{\text{totale generale}} = \frac{\left(\sum_{k=1}^{c} O_{ik}\right)\left(\sum_{k=1}^{r} O_{kj}\right)}{\sum_{k=1}^{r}\sum_{m=1}^{c} O_{km}}$$

e la statistica test (il chi-quadro calcolato dal campione) è

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}$$

[2] Come regola pratica, si richiede che tutte le frequenze attese siano ≥ 5.

che ha approssimativamente la distribuzione χ^2 con grado di libertà $\nu = (r-1)(c-1)$.

Una volta fissato il livello di significatività α, si rifiuta l'ipotesi nulla, dichiarando che invece esiste una dipendenza tra le variabili, se $\chi^2 > \chi^2_\alpha$.

7.1.1 Esempio

Per stabilire l'efficacia di un vaccino anti-influenzale è stata condotta una ricerca, somministrando il vaccino a 500 persone e controllando il loro stato di salute in un anno; lo stesso controllo è stato fatto per un gruppo di altre 500 persone non vaccinate. In base ai risultati dell'esperimento, si è ottenuta la seguente tabella di contingenza:

	nessuna influenza	una influenza	più di una influenza	Totale
vaccinati	252	145	103	500
non vaccinati	224	136	140	500
Totale	476	281	243	1000

Si vuole determinare se le variabili "vaccinazione" e "numero di influenze" siano o meno indipendenti: nel caso lo fossero, significherebbe che il vaccino non è efficace.

Applicando la formula mostrata in precedenza (o, più concretamente, usando uno strumento software), si ottengono le seguenti frequenze attese:

	nessuna influenza	una influenza	più di una influenza
vaccinati	238	140.5	121.5
non vaccinati	238	140.5	121.5

Da queste, si ricava che il valore del chi-quadro è $\chi^2 = 7.57$. Il grado di libertà è

$$\nu = (r-1)(c-1) = (2-1)(3-1) = 2$$

perciò:

- per $\alpha = 5$ % si ha $\chi^2 = 7.57 > 5.991 = \chi^2_{0.05}$, quindi si rifiuta l'ipotesi nulla, affermando che le variabili non sono indipendenti (ovvero che c'è evidenza statistica di efficacia del vaccino);
- per $\alpha = 1$ % si ha $\chi^2 = 7.57 < 9.210 = \chi^2_{0.01}$, quindi *non* si rifiuta l'ipotesi nulla.