Il trattamento per gruppi

Christian Genest *Università McGill, Montréal, Canadà*
Christiane Rousseau *Universitá di Montréal, Montréal, Canadà*

Il tracciamento massiccio è un elemento essenziale della lotta contro la propagazione del coronavirus. Ma come far fronte a una possibile penuria di reagenti e di materiale? Operando i *test* su miscele di prelievi e facendo appello alla matematica.

Con il timore anticipato di una seconda ondata di COVID-19, numerosi esperti sono concordi nel dire che l'instaurazione di un piano di tracciamento su grande scala è necessario per arrestare la propagazione del coronavirus. Realizzati su campioni rappresentativi della popolazione, i *test* immunologici, sierologici o antigenici permetterebbero anche di stimare la diffusione della malattia, di giudicare il grado di immunità collettiva e di adattare i mezzi di gestione della pandemia.

Perché sia coronata da successo, il varo di una strategia di tracciamento presuppone l'accesso a risorse adeguate di personale e di materiale. Con la crescita della richiesta a livello planetario si profila all'orizzonte una penuria dei reagenti necessari alle analisi di laboratorio; questa resta una preoccupazione delle autorità della sanità pubblica in Canadà e nel resto del mondo.

Sapendo che la maggior parte dei *test* si rivelano (molto fortunatamente) negativi, si può mettere a profitto la matematica per fare meglio? Sì, si può, segnatamente realizzando *test* di gruppo su miscele di prelievi costruiti in maniera giudiziosa.

Il tracciamento per gruppi

Immaginiamo che un laboratorio abbia ricevuto 100 prelievi per il tracciamento. Li divide a caso in 5 gruppi di 20 ciascuno. Poi, gruppo per gruppo, utilizza la metà di ciascuno dei 20 prelievi per costituire una miscela alla quale si applica il *test*.

Se il *test* effettuato su una miscela è negativo si può subito concludere che nessun membro del gruppo interessato è infetto. Se il *test* è invece positivo, allora si procede a *test* individuali sulla seconda metà di ognuno dei 20 prelievi.

Se i 100 prelievi originali provengono tutti da persone sane, questa procedura permette di accertarsene facendo 5 *test* invece di 100. Se un solo individuo è infetto, bastano $5 + 20 = 25$ *test* per individuarlo. Se sono infettate due persone, le si possono individuare ancora con 25 *test* se sono nello stesso gruppo, ma ne occorrono $5 + 20 + 20 = 45$ se appartengono a gruppi differenti. E così via se vi sono tre o più persone infette.

Come si può constatare, il tracciamento per gruppi permette dunque di realizzare economie importanti, purché la sensibilità e la specificità del *test* non siano modificate dalla miscela, come si è supposto qui e come è spesso il caso in pratica.

Il laboratorio avrebbe anche potuto applicare la stessa strategia di tracciamento a 10 gruppi di 10 prelievi. Se un solo individuo è infettato, non vi sarebbe stato allora bisogno che di 20 *test* per identificarlo. In cambio, sarebbero stati necessari 10 *test* per concludere che nessuno è infetto.

Quale di queste due strategie è migliore? E ne esistono altre che siano preferibili a queste? La risposta dipende dalla prevalenza della malattia, cioè dalla proporzione della popolazione che è infetta.

In una nota apparsa nel 1943 in *The Annals of Mathematical Statistics*, l'americano Robert Dorfman [1] riporta che, nella sua forma più elementare, il rintracciamento per gruppi era già stato usato nella Seconda Guerra mondiale per individuare i casi di sifilide tra i coscritti. Questo approccio si è affermato e ne esistono oggi molte varianti che sono utilizzate dovunque in Nord America per controllare la presenza dell'HIV, dell'influenza, o del Virus del Nilo occidentale.

Ottimizzare l'algoritmo

Dorfman ha mostrato come determinare la numerosità ottimale di un gruppo in funzione della prevalenza $p \in [0,1]$ della malattia. Indichiamo con $n \geq 2$ la numerosità del gruppo e supponiamo che i suoi membri costituiscano un campione aleatorio rappresentativo della popolazione.

Se X denota il numero sconosciuto di persone infette nel gruppo, questa variabile obbedisce a una legge binomiale[1] di parametri n e p, donde

$$Pr(X = 0) = (1-p)^n \ .$$

Poiché ogni individuo ha probabilità $1-p$ di essere sano

$$Pr(X > 0) = 1 - Pr(X = 0) = 1 - (1-p)^n \ .$$

Se $X = 0$, non si faranno allora che $N = 1$ *test*. Tuttavia, se $X > 0$, si faranno $N = n+1$ *test*. In media, il numero di *test* che saranno effettuati, chiamato speranza di N e denotato con $E(N)$, è eguale a

$$\begin{aligned} & E(N) \\ &= \quad 1 \times Pr(X=0) + (n+1) \times Pr(X>0) \\ &= \quad n + 1 - n(1-p)^n \ . \end{aligned}$$

Questa è una funzione crescente di p. Se $p = 0$, si ha $E(N) = 1$, questo che è evidente perché nessuno ha la malattia e dunque basta un solo *test* per confermarlo. Se $p = 1$, si ha $E(N) = n+1$ perché il primo *test* sarà necessariamente positivo.

Per ogni valore di $p \in [0,1]$, è possibile determinare il `costo relativo` legato all'uso del tracciamento per gruppi studiando il comportamento del rapporto

$$\frac{E(N)}{n} = 1 + \frac{1}{n} - (1-p)^n \ ,$$

in funzione di n. Più $E(N)/n$ è piccolo, più conviene ricorrere ai *test* per gruppo, naturalmente a condizione che il rapporto sia inferiore a 1. Quando $p = 0$, si trova $E(N)/n = 1/n$, di modo che si ha interesse a prendere n più grande possibile. Quando $p = 1$, si ha sempre

$$\frac{E(N)}{n} = 1 + \frac{1}{n} > 1 \ ,$$

perché il *test* sul gruppo è sempre positivo dunque e non fa quindi che aggiungere altri *test*.

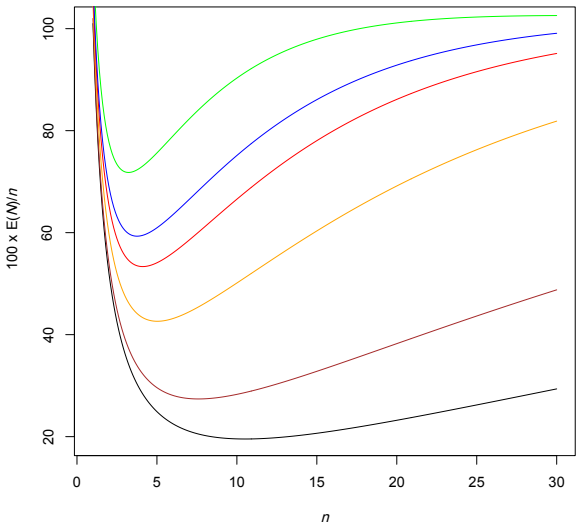

Figura 1: *Tracciato della curva* $100 \times E(N)/n$ *in funzione di* n *per diversi valori di* p*: 1% (nero), 2% (marrone), 5% (arancio), 8% (rosso), 10% (blu) e 15% (verde).*

Per un valore fissato di p, la funzione $100 \times E(N)/n$ rappresenta la percentuale media di *test* effettuati in funzione della numerosità, n, del gruppo. La Figura 1 mostra il grafico di questa funzione per diversi valori di p, corrispondenti a una prevalenza dell'1% (nero), 2% (marrone), 5%

[1] Si tratta qui di un'approssimazione che è giustificata nella misura in cui la popolazione è molto grande rispetto alla numerosità dei gruppi.

(arancio), 8% (rosso),10% (blu) e 15% (verde). Come si può vedere, la numerosità ottimale dipende dalla miscela, che corrisponde al minimo della curva, e varia in funzione della proporzione, p, degli individui infetti nella popolazione. La Tabella 1, riportata da Dorfman [1], dà la scelta ottimale di n per qualche valore di p.

p (%)	n	Costo relativo (%)
1	11	20
2	8	27
5	5	43
8	4	53
10	4	59
15	3	72

Tabella 1: *Scelta ottimale di n per alcuni valori della percentuale p degli infetti.*

Generalizzazioni

Il protocollo di *test* descritto sopra è l'esempio di `algoritmo adattativo` a due passi. Si definisce `adattativo` perché la scelta (e dunque il numero) dei *test* da effettuare nel secondo passo dipende dal risultato del *test* realizzato al primo. Esistono diversi modi di migliorare il successo di questo tipo di algoritmo. In particolare la procedura può essere estesa aumentando il numero di passi.

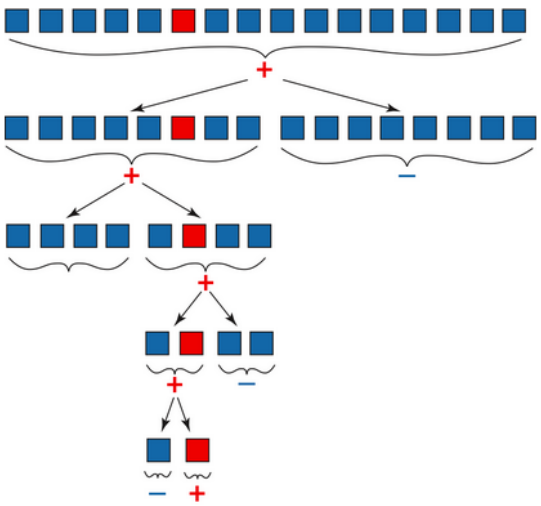

Figura 2: *Rappresentazione grafica di un algoritmo di divisione binaria adattivo a 5 passi applicato ad un insieme iniziale di $n = 2^4 = 16$ prelievi, di cui solo uno è infetto.*

Ecco un algoritmo classico, che chiameremo `algoritmo di divisione binaria`, e che possiede certe proprietà di ottimalità (si veda la Figura 2).

a) Si prende un intero n della forma $n = 2^s$ e si effettuano k passi del *test*, ove $k \leq s + 1$.

b) Al primo passo si effettua un *test* della miscela dei prelievi di tutto il gruppo.

c) Se il *test* risulta positivo, si divide allora il gruppo in due sottogruppi di 2^{s-1} prelievi e si effettua un *test* su ognuno di essi.

d) Si procede nello stesso modo sino al k-esimo passo, al quale si effettua il *test* sui membri di un sottogruppo dichiarato positivo al passo precedente. Nel caso particolare $k = s + 1$, questo sottogruppo ha solo due elementi.

Se il gruppo contiene un solo individuo infetto, questo algoritmo premetterà di identificarlo in esattamente $s + 1 = \log_2(n) + 1$ passi. Come regola generale, più alto è il numero di passi, migliori sono le economie realizzate tramite questo approccio. Tuttavia, se bisogna attendere da 24 a 48 ore perché il *test* dia il risultato, i ritardi nella consegna dei risultati rischiano di essere controproducenti. Si noti anche che questa miglioria richiede prelievi biologici più importanti. Questo non è considerato veramente un problema e lo si ritroverà in tutti gli algoritmi presentati di seguito.

Un algoritmo non adattativo

Per controllare meglio il tempo di risposta, si può anche pensare di utilizzare metodi non adattativi di tracciamento per gruppi. Questi protocolli comportano un solo passo, ciò che permette di effettuare tutti i *test* simultaneamente. Essi si rivelano inoltre molto efficaci per il tracciamento dei casi se si dispone di una stima affidabile della prevalenza della malattia.

Spieghiamo questo concetto mediante il seguente esempio, sviluppato da un gruppo di ricercatori ruandesi nel quadro della lotta attuale al COVID-19. Si forma dapprima un campione aleatorio di numerosità $n = 3^m$. Si stabilisce quindi una corrispondenza tra i 3^m individui e i

punti di un ipercubo discreto $\{0,1,2\}^m$. Si veda la figura 3 per un esempio nel caso $m = 3$.

L'approccio proposto consiste allora nell'effettuare contemporaneamente $3m$ *test* su miscele di campioni che comprendono ognuna 3^{m-1} individui. Le miscele sono tuttavia formate secondo modalità molto precise, ossia dei tagli nell'ipercubo. In effetti, se $x_1, \ldots, x_m$ denotano gli assi coordinati dellipercubo, allora ogni miscela corrisponde agli individui situati nell'iperpiano $x_i = t$, ove $i \in \{1, \ldots, m\}$ e $t \in \{0,1,2\}$ è una tranche di 3^{m-1} individui.

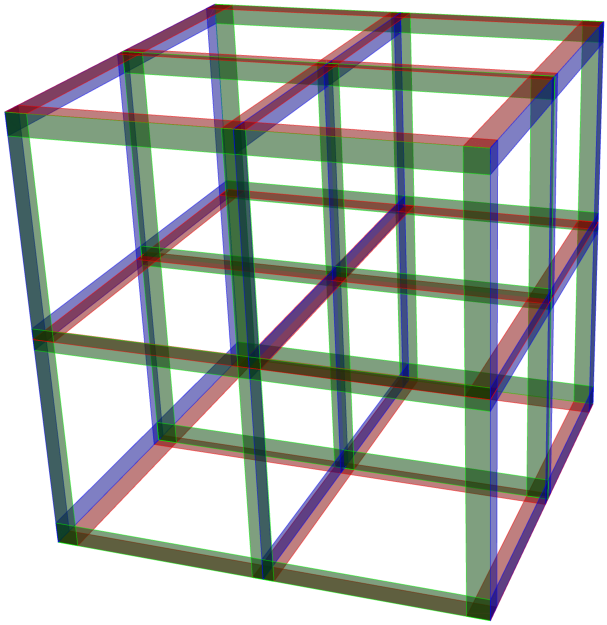

Figura 3: *Ipercubo discreto $\{0,1,2\}^3$. Ogni punto del reticolo corrisponde ad un individuo di un campione casuale di dimensioni $n = 3^3 = 27$. Gli individui $3^2 = 9$ di ciascuna delle 9 miscele si trovano lungo una fetta rossa, blu o verde.*

Quando $m = 3$, come nella Figura 3, si effettuano allora $3 \times 3 = 9$ *test* su gruppi di $3^2 = 9$ individui. Quando $m = 4$, che è il valore riportato nel caso del Ruanda, si effettuano piuttosto 12 *test* a partire da un campione di $n = 81$ individui. Ciò significa che ogni prelievo è diviso in quattro porzioni uguali e contribuisce a quattro *test* differenti. Inoltre ogni *test* è fatto su una miscela di 27 campioni.

Questo approccio si basa su una tecnica di costruzione dei codici correttori d'errori descritta nel riquadro[2]. Uno dei grandi vantaggi è che la composizione delle miscele è tale che se nel campione vi è un solo individuo infetto questo può essere individuato con certezza. Di contro, se è presente più di una persona infetta, occorre allora procedere a un secondo giro di *test*.

Esaminiamo ora l'esempio ruandese nel caso $n = 81 = 3^4$. Sapendo che il numero X di individui infetti nel campione obbedisce a una legge binomiale di parametri $n = 81$ e p, si ha

$$Pr(X \leq 1) = (1-p)^{81} + 81p(1-p)^{80} \ .$$

Questa strategia è interessante se è grande la probabilità che sia $X \leq 1$. Ma perché sia $Pr(X \leq 1) \geq 0.95$, per esempio, occorre avere $p \leq 0.44\%$; in altre parole, la prevalenza della malattia deve essere bassa. Già a 1% di prevalenza si ha $Pr(X \leq 1) = 0.806$ e in quasi il 20% dei casi bisognerà ricorrere a un secondo giro di *test*. Tuttavia, ancora per l'1% di prevalenza si ha

$$Pr(X = 2) = \binom{81}{2} p^2(1-p)^{79} = 0.146 \ ,$$

da cui $Pr(X \leq 2) = 0.952$.

È possibile controllare facilmente i costi se ci si mostra astuti nel modo di condurre il secondo passo, quando vi è più di un individuo infetto. Per esempio, tutti i casi che corrispondono a $X = 2$ soddisfano alla seguente proprietà (si faccia riferimento alla Figura 4 dove si mostrano, nel caso $m = 3$, le *tranche* per le quali il *test* è positivo):

> (P) Per ogni valore di $i \in \{1, \ldots, m\}$, vi sono al più due *tranche* della forma $x_i = s$ e $x_i = t$ che portano a un *test* positivo, ed esiste almeno un valore di i per il quale si hanno esattamente due *test* positivi di questa forma.

Se k è il numero di valori di $i \in \{1, \ldots, m\}$ per i quali si hanno due *test* positivi della forma specificata in (P), il numero dei *test* aggiuntivi richiesti per individuare tutti gli individui infetti è dato nella Tabella 2.

Se non è verificata la proprietà (P), si debbono allora controllare tutti gli individui delle *tranche* che hanno portato ad un *test* positivo, ciò che si traduce in al più 81 *test* supplementari. Alla fine, il costo totale è dunque inferiore o eguale a $100 \times 13.06/81 \approx 16.1\%$.

[2] Si veda anche http://accromath.uqam.ca/accro/wp-content/uploads/2020/02/Codes.pdf per una introduzione ai codici correttori d'errori.

Costruire un algoritmo non adattativo

Si vuole costruire un algoritmo non adattativo per *test*are un gruppo di persone. Si rappresenterà questo algoritmo mediante una tabella di T righe e n colonne, o, in maniera equivalente, mediante una matrice di dimensione $T \times n$ con tutti gli elementi eguali a 0 o a 1. Gli elementi della j-esima colonna rappresentano i *test* ai quali partecipa il j-esimo individuo. Così, $m_{ij} = 1$ se il j-esimo individuo contribuisce all'i-esimo *test* e 0 altrimenti.
Costruiamo un vettore X di lunghezza n che rappresenti il gruppo che ci si accinge a *test*are: la sua j-esima coordinata, x_i, vale 1 se il j-esimo individuo è infetto, 0 altrimenti. Si decide di trattare X come una parola di lunghezza n e le sue coordinate eguali a 1 come errori in una parola iniziale X_0 in cui tutte le coordinate erano nulle. Degli algoritmi di codici correttori d'errori permettono di correggere gli errori che si sarebbero prodotti nella trasmissione di X_0, ciò che equivale a identificare quali siano le coordinate x_j di X che valgono 1, ossia esattamente lo scopo cercato. Come illustrato nell'esempio, un algoritmo di correzione degli errori non può correggere che un numero massimo prefissato di errori k scelto al momento della sua costruzione.
La matrice M è la `matrice del codice`. Una proprietà sufficiente perché il codice possa correggere k errori è che la matrice sia k-`disgiunta`. Definiamo questa nozione. Non si vuole che se degli individui $j_1, \ldots, j_k$ (non necessariamente distinti) sono infetti, allora un j_{k+1}-esimo individuo infetto rimane inosservato. Dare la colonna j (ossia l'individuo j) è lo stesso che dare il sottoinsieme A_j di $\{1, \ldots, T\}$ degli elementi uguali a 1 (vale a dire l'insieme dei *test* ai quali contribuisce quell'individuo). Se gli individui $j_1, \cdots, j_k$ sono infetti, allora tutti i *test* corrispondenti all'unione $A_{j1} \cup \cdots \cup A_{jk}$ saranno positivi. Perché non passi inosservato il j_{k+1}-esimo individuo infetto bisogna che A_{jk+1} non sia incluso in $A_{j1} \cup \cdots \cup A_{jk}$. La matrice M è k-`disgiunta` se ciò si verifica per ogni $j_1, \ldots, j_k$ e per ogni j_{k+1} distinto da $j_1, \ldots, j_k$. La matrice 12×81 che corrisponde all'esempio precedente utilizzato in Ruanda sarebbe una matrice 1-`disgiunta`.
Esistono due principali tipi di metodi per costruire matrici k-`disgiunte`. Il primo è probabilistico: si generano a caso matrici i cui elementi sono 1 con probabilità q e 0 altrimenti. Per n, T e k ben scelti la probabilità che la matrice sia k-`disgiunta` è non nulla; per tentativi si finisce dunque per generare una matrice k-`disgiunta`.
Il secondo metodo, algebrico, è preso a prestito dalla teoria dei codici correttori d'errori di Reed-Solomon. Questo permette di costruire matrici M nelle quali tutte le righe hanno lo stesso numero m di elementi non nulli e tutte le colonne hanno lo stesso numero c di elementi non nulli. Così si effettuano tutti i *test* su sottogruppi di numerosità m e ogni prelievo individuale è diviso in c parti e può essere incluso in c *test* distinti.

k	Numero di *test* supplementari
1	0
2	4
3	8
4	16

Tabella 2: *Scelta ottimale di n per alcuni valori della percentuale p degli infetti.*

Prospettive

La ricerca di algoritmi adatti alla lotta contro il coronavirus raggiunge il culmine. Per esempio, recentemente è stato elaborato e implementato in laboratorio un algoritmo da un gruppo di ricerca israeliano: consiste nella realizzazione di 48 *test* sullo stesso campione di numerosità $8 \times 48 = 384$. Ogni prelievo individuale è diviso in sei parti uguali. Ogni *test* si basa sulla mescola di 48 di queste parti, una per ogni individuo. Ogni individuo è dunque presente in sei *test* differenti.

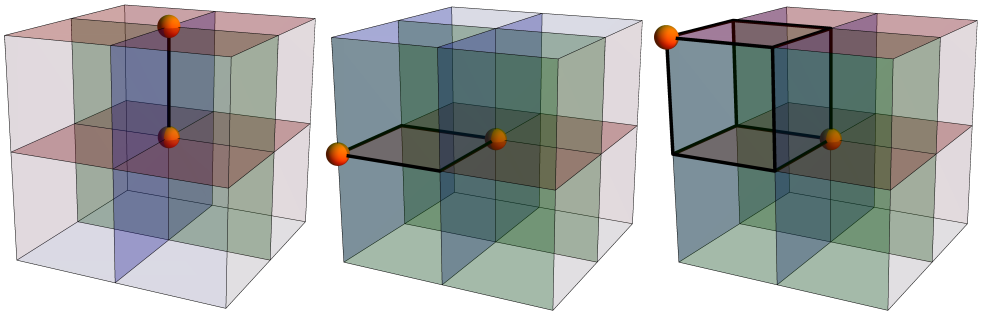

Figura 4: *L'ipercubo discreto $\{0,1,2\}^3$ e le tre possibilità per due individui infetti: sulla stessa linea retta parallela ad un asse (sinistra), ai vertici opposti di un quadrato in un piano (centro) o ai vertici opposti di un cubo (destra). In ogni caso, le fette che portano ad un test positivo sono identificate in rosso, verde o blu con le stesse convenzioni di colore della Figura 3.*

Nel laboratorio è stato programmato un *robot* per preparare le 48 miscele. L'algoritmo consente di identificare sino a quattro persone infette. Si ha così bisogno di un numero di *test* otto volte inferiore a quello degli individui. Di nuovo, minore è la percentuale di infetti, migliore è la *performance* dell'algoritmo.

Naturalmente, nell'elaborazione di una procedura di un *test* statistico entrano in gioco altre considerazioni. Per semplificare, noi abbiamo qui supposto implicitamente che il *test* usato sia infallibile. Nella pratica anche le migliori procedure possono dare falsi positivi o falsi negativi. La sensibilità e la specificità dei *test* sono elementi importanti da prendere in considerazione al momento di raccomandarne l'implementazione, come pure la loro fattibilità in termini di tempo, di costi e di complessità delle manipolazioni.

Questo articolo è apparso in lingua originale francese in:
`http://accromath.uqam.ca/volume/volume-15-2-ete-automne-2020/`
e nella sua traduzione italiana in:
`http://accromath.uqam.ca/2020/09/il-tracciamento-per-gruppi/`
Ringraziamo gli autori e Accromαth per il permesso di riprodurlo.

[1] R. Dorfman: *The Detection of Defective Members of Large Populations*, Ann. Math. Statist., 4 (1943) 436.

Christian Genest: è professore di Statistica e titolare di una cattedra di ricerca sui Modelli di dipendenza stocastica all'Università McGill, Montréal, Canadà. È stato presidente della Società di Statistica del Canadà (SSC) nel 2007-08 e redattore capo del *Journal of Multivariate Analysis* dal 2015 al 2019. Ha ricevuto la medaglia d'oro per la ricerca dalla SSC nel 2011 e il premio John L. Synge dalla Società reale del Canadà nel 2020.

Christiane Rousseau: è professoressa emerita di Matematica all'Università di Montréal. È stata presidentessa della Società matematica canadese dal 2002 al 2004. Ha promosso e coordinato l'anno internazionale della Matematica del pianeta Terra 2013 con il patrocinio dell'UNESCO. Dal 2011 al 2018 ha fatto parte del comitato esecutivo dell'Unione Matematica Internazionale per la quale ha guidato il dossier per la proclamazione del 14 marzo come Giornata internazionale della Matematica da parte dell'UNESCO nel novembre 2019.