Modelli statistici per l'analisi della prestazione eccezionale nell'attività sportiva

Andrea Spizzichino

Introduzione

È prassi ormai comune, quando si parla di un tema relativo ai fenomeni sportivi, iniziare con informazioni di tipo statistico; questo dimostra che la metodologia statistica è inserita a pieno titolo nell'ambito della scienza dello sport. I temi trattati spaziano su vari aspetti, tra cui la definizione di indicatori di partecipazione individuale, l'analisi della diffusione e dell'evoluzione della pratica sportiva, le implicazioni sociali, demografiche ed economiche connesse con tale attività.

Le potenzialità di un approccio statistico allo studio dello sport non si limitano soltanto alle problematiche prima citate; nel corso degli ultimi anni si è assistito a una crescente domanda di tecniche statistiche avanzate per la soluzione di problemi legati alla valutazione delle prestazioni nell'attività agonistica.

Sia per gli sport di squadra sia per quelli individuali l'interesse degli statistici si è concentrato sulla 'prestazione eccezionale'. In particolare per gli sport individuali, si è studiata la prestazione eccezionale più come evoluzione dei record (da chiunque stabiliti) nelle diverse discipline, che come analisi delle performance del singolo atleta eccezionale.

Come era facile aspettarsi, tra le discipline individuali è l'atletica quella che ha richiamato maggiormente l'attenzione degli studiosi; l'applicazione di modelli statistici necessita infatti di variabili quantitative quali tempi e distanze. In particolare, i contributi più interessanti sono rivolti alla previsione dei record futuri per le specialità della corsa. Un limite in questo tipo di studi è la carenza di dati statistici completi; oltre alle serie storiche delle prestazioni realizzate dagli atleti è difficile trovare basi di dati che facciano riferimento a fattori ambientali (o personali) che sicuramente influiscono sui risultati ottenuti.

Questo lavoro si sviluppa ispirandosi a metodi statistici usati da Robinson e Tawn (1995), basati su tecniche del valore estremo, per stimare la migliore prestazione possibile di una popolazione di atleti in competizione. Un clichè in atletica è che i record esistano per essere battuti. Con lo sviluppo delle tecniche d'allenamento e delle attrezzature, con il miglioramento della scienza dietetica e l'aumento della partecipazione data dal coinvolgimento di nuove 'popolazioni' di atleti, i record sono battuti con maggiore frequenza e con miglioramenti più consistenti.

Anno	Africa	Europa	Nord America	Centro e Sud America	Asia	Oceania
1940	0	19	0	1	0	0
1960	1	13	1	0	0	5
1980	4	9	2	0	3	2
1990	6	10	4	0	0	0
1995	16	0	1	2	0	1
1999	17	2	0	1	0	0
2007	20	0	0	0	0	0
2013	18	1	1	0	0	0

Tabella 1 - I 20 migliori atleti maschi nella specialità dei 10 mila metri per continente d'appartenenza e anno.

Nel contesto di una popolazione che migliora è molto probabile che un record venga battuto, ma se il margine è sorprendentemente ampio si può sospettare che la prestazione sia stata raggiunta con metodi non corretti, ad esempio l'uso di sostanze farmacologiche. Un'alternativa all'ipotesi del doping si verifica nel caso in cui l'atleta che realizza un record 'sospetto' provenga da una popolazione differente da quelle che fino a quel momento hanno partecipato a quel tipo di competizione.

Questo lavoro si pone l'obiettivo di valutare se i miglioramenti in una determinata specialità sono più o meno plausibili, e stabilire di conseguenza se le prestazioni eccezionali sono dovute al doping o all'appartenenza dell'atleta a una popolazione nuova per quello sport.

Il quadro di riferimento

Tra le diverse specialità dell'atletica leggera, molte hanno fatto registrare prestazioni eccezionali che hanno rotto la continuità con il passato, la specialità scelta in questo caso sono i 10 mila metri, in quanto, sia in campo maschile sia femminile, ci sono stati miglioramenti improvvisi e notevoli a opera di atleti provenienti da paesi senza una tradizione in questa disciplina.

Per gli uomini, grossi miglioramenti si sono avuti a partire dal 1990, quando diversi atleti africani hanno cominciato a monopolizzare la scena mondiale, realizzando tempi ai quali nessun corridore extra-africano è ancora riuscito ad avvicinarsi.

È evidente che i risultati ottenuti sono dovuti all'appartenenza degli atleti a etnie che fino agli anni '90 non partecipavano alle gare in questione; la tavola 1 mostra come sia cambiata nel tempo la geografia dei migliori atleti; negli ultimi 20 anni gli africani sono arrivati ad avere sempre la maggior parte dei migliori 20 corridori nell'arco di un anno.

Ad aiutare queste performance contribuiscono le condizioni climatiche ed economiche nonché il modo di vivere che porta alla nascita naturale di grandi atleti; è altrettanto vero che grandi corridori hanno stimolato lo spirito di emulazione nelle nuove generazioni e che vari incentivi arrivano anche dall'estero (borse di studio da parte delle università americane e offerte dei club europei), il che stimola i giovani a impegnarsi anche in vista di un futuro migliore mediante lo sport.

Per le donne, l'attenzione si concentra sulla prestazione dell'atleta cinese Junxia Wang, e di altre sue connazionali. Nel 1993 il fondo e mezzofondo mondiale assistono a prestazioni strabilianti da parte di varie atlete cinesi che hanno sollevato curiosità e dubbi in tutto in mondo. Junxia Wang, l'8 settembre a Pechino riusciva a percorrere i 10 mila metri in 29 minuti 31 secondi e 78 centesimi migliorando di oltre 40 secondi il precedente record e realizzando una prestazione che ancora oggi nessuno ha mai avvicinato.

Ci sono diversi atteggiamenti degli studiosi rispetto ai record cinesi.

I più scettici, ricordando che la Cina è stata varie volte al centro di casi di doping, sospettano che le atlete facessero uso di sostanze non facili da rilevare con esami clinici. Veniva inoltre posta l'attenzione sul fatto che così tante atlete di una stessa nazione, nello stesso momento, avevano iniziato a raggiungere risultati eccezionali con pochissimi anni di preparazione, mentre le più grandi atlete al mondo, avevano dovuto sopportare lunghi anni di allenamento prima di arrivare ad alti livelli.

In contrapposizione agli scettici, ci sono studiosi che hanno motivato in maniera diversa questi grandi risultati, puntando l'attenzione sul fatto che mai una nazione si era concentrata in modo specifico su queste discipline, ma al contrario ci si era sempre affidati ad atleti cresciuti quasi per caso. A tutto ciò si aggiungeva l'enorme bacino di reclutamento, organizzato su varie province, con un'organizzazione centralizzata che portava in superficie solo i veri talenti, e l motivazioni delle cinesi che erano superiori a quelle delle altre atlete, basandosi sia su fattori ideologici per lo più superati in Occidente (immagine del proprio paese e veicolo dell'orgoglio nazionale) sia su vantaggi concreti superiori a quelli ottenibili altrove; basti pensare che vincendo la maratona di Tianjin, la Wang ha guadagnato l'equivalente di circa 40 anni di stipendio.

I dubbi sui risultati delle atlete cinesi non derivano però solo dal modo in cui sono stati battuti i record, ma anche dal fatto che dopo una strepitosa annata ne la Wang ne altre atlete cinesi sono riuscite a replicare quelle performance facendo cadere l'ipotesi che elementi strutturali latenti abbiano determinato questo exploit.

È in questo quadro che si cerca di stabilire il tempo limite a cui arriveranno in futuro gli atleti e confrontarlo con i tempi già realizzati, ponendo in particolare l'attenzione sulle prestazioni che hanno destato maggiore stupore tra gli esperti di atletica.

Il tempo limite verrà determinato con due differenti metodologie che utilizzano come informazio-

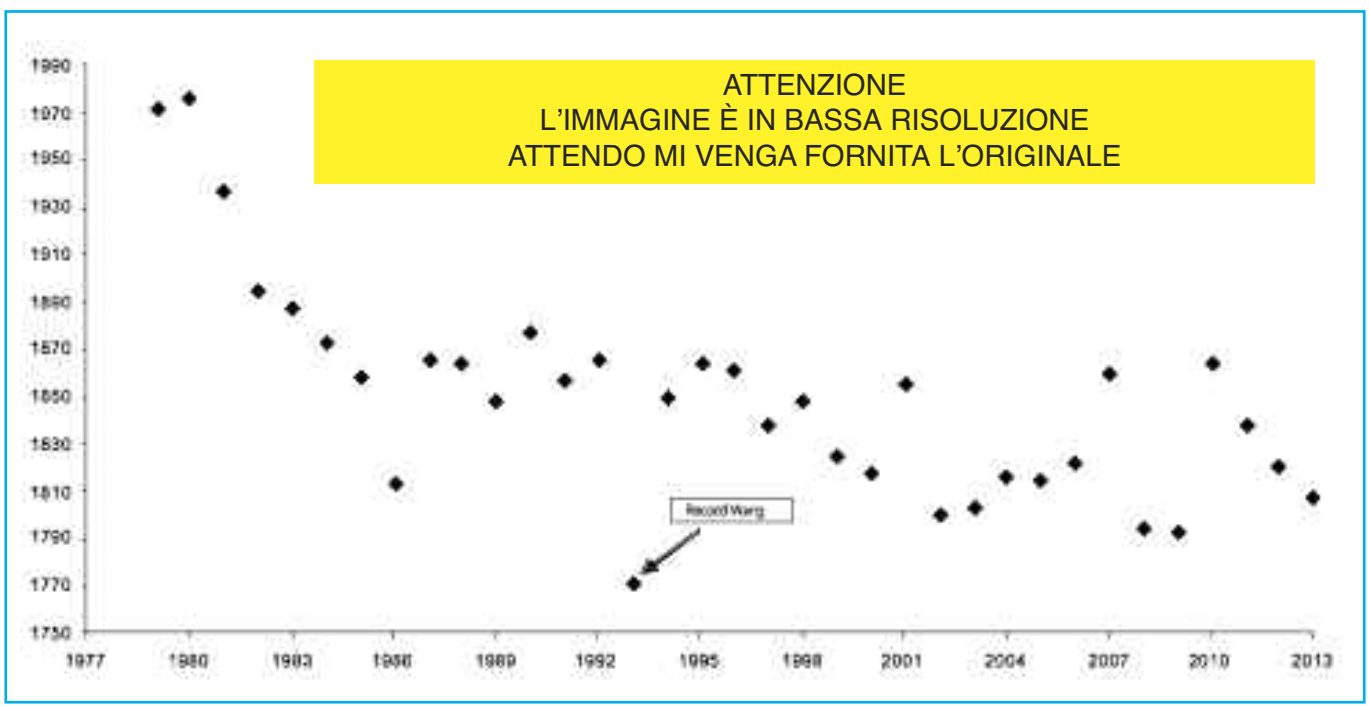

Figura 1 - Miglior tempo nei 10 mila metri femminili dal 1978 al 2013. (In secondi)

ne la migliore prestazione annuale l'una e i primi cinque tempi realizzati all'interno dello stesso anno l'altra. Come già accennato la metodologia utilizzata si basa su tecniche del valore estremo volte alla ricerca del valore limite per il futuro attraverso l'analisi degli estremi già osservati.

Inizialmente furono essenzialmente i fenomeni metereologici e idrologici che spinsero ad approfondire questi studi con riguardo alla previsione delle precipitazioni, delle piene, delle magre eccezionali ecc.

Gli estremi oggetto di studio possono essere sia i massimi sia i minimi; in questo caso l'interesse è per la distribuzione dei minimi in quanto le misurazioni riguardano i tempi migliori, e quindi più bassi, realizzati nei vari anni.

In questo caso non si usa esattamente la teoria degli estremi ma una sua generalizzazione che prende appunto il nome di teoria degli estremi generalizzata.

Secondo questa teoria, si ha come funzione di distribuzione del minimo la seguente espressione:

$$GEV(\mu,\sigma,\lambda) = G(x) = \\ = 1 - \exp\left[-\{1-\lambda(x-\mu)/\sigma\}_{+}^{-1/\lambda}\right]$$

Dove i parametri soddisfano i vincoli $\sigma>0$, $\mu \in R$, $\lambda \in R$ e $\{y\}_{+}=\max(y,0)$.

Questa distribuzione è stata scelta per l'asintoticità dei dati oggetto di studio; la distribuzione del minimo di variabili casuali identicamente distribuite è infatti ben modellata dalla distribuzione *GEV*.

I dati utilizzati per le elaborazioni sono le graduatorie internazionali di atletica leggera reperibili per gli ultimi anni in rete (http://www.alltime-athletics.com), per gli anni passati su riviste specializzate (Atletica leggera). Nelle graduatorie vengono riportate le migliori prestazioni annuali dei vari atleti, oltre al tempo (o misura) realizzato viene riportato il luogo e il giorno in cui la gara si è tenuta.

Stima del valore estremo - prima tecnica

La prima metodologia è basata sull'analisi del miglior tempo, calcolato in secondi, realizzato dalle donne tra il 1978 e il 2013; la figura 1 mostra l'andamento nel tempo delle migliori prestazioni.

Viene subito all'occhio lo stupefacente tempo realizzato nel 1993 da Junxia Wang (1771.78), che migliora di oltre 40 secondi il record precedente, nonchè l'asintoticità dei tempi con il passare degli anni, che avvalora la scelta della *GEV* come funzione di distribuzione.

Si suppone che ci sia indipendenza tra le prestazioni nei vari anni e su questa base si ipotizza che i tempi realizzati dai vari atleti seguano un trend che può essere incorporato all'interno della distribuzione; ammettendo quindi che m varia negli anni avrò:

$$G_t(x) = P\{X_t \le x\} = \\ = 1 - \exp\left[-\{1-\lambda(x-\mu_t)/\sigma\}_{+}^{-1/\lambda}\right]$$

dove X_t indica il minor tempo nell'anno t e l'anno di partenza è quello per il quale abbiamo la prima osservazione.

Il trend che considero per μ_t è dato dalla curva interpolatrice del grafico precedente, per la quale è appropriato un decadimento esponenziale;

$$\mu_t = \alpha - \beta\{1 - \exp(-\gamma t)\}$$

dove $\beta>0$ e $\gamma>0$.

Robinson e Tawn fanno notare che anche le variazioni di λ e s possono intervenire sui tempi, ma in questa analisi, come nella loro, i dati sono insufficienti per esaminare questo aspetto; la funzione di distribuzione con al suo interno la μ_t diventa:

$$G_t(x) = 1 - \exp\left[-\{1-\lambda(x-\alpha+\beta(1- \\ -\exp(-\gamma t)))/\sigma\}_{+}^{-1/\lambda}\right]$$

L'intento è quello di stimare contemporaneamente tutti i parametri della *GEV*, per farlo si determina una stima di massima verosimiglianza mediante il software di calcolo Mathematica da cui

$$\alpha = 2021{,}1,\ \beta = 185{,}858,\ \sigma = 21{,}1595, \\ \lambda = -0{,}210043,\ \gamma = 0{,}163562.$$

Avendo unito il trend alla funzione di distribuzione, e avendo determinato i parametri, rimane da trovare il tempo limite x_{ult}, cui si può, teoricamente, arrivare. Per farlo si definisce $x_{p,t}$ il tempo che è battuto con probabilità p nell'anno t; ponendo quindi $G(x_{p,t}) = p$, avrò:

$$1 - \exp\left[-\{1-\lambda(x_{p,t}-\mu_t)/\sigma\}_{+}^{-1/\lambda}\right] = p$$

da cui

$$x_{p,t} = \mu_t + \sigma\left[1-[-\log(1-p)]^{-\lambda}\right]/\lambda.$$

Sostituendo in quest'equazione l'espressione di μ_t, si ha:

$$x_{p,t} = \alpha - \beta(1 - \exp(-\gamma t)) + \\ + \sigma\,[1-[-\log(1-p)]^{-\lambda}]/\lambda$$

Ponendo p=0 e calcolando il limite per t che tende a infinito si ha l'espressione per l'ultimo tempo

$$x_{ult} = \begin{cases} \alpha - \beta + \sigma/\lambda & se\,\lambda < 0 \\ -\infty & se\,\lambda \geq 0 \end{cases}$$

La scelta di porre p=0 è dettata dal fatto che l'ultimo tempo, in quanto più basso in assoluto, ha probabilità 0 di essere battuto.

Si osserva che per $\lambda \geq 0$ il valore di $x_{p,t}$ non ha senso, si giunge così a definire come tempo limite nella specialità dei 10 mila metri femminili il valore ottenuto nel caso in cui λ risulta negativo.

Avendo a disposizione i valori dei parametri necessari per determinare il tempo finale, li sostituisco nell'equazione e ho x_{ult}=1734,51.

Confrontando questo tempo con quello realizzato dall'atleta cinese si nota che x_{ult} è più basso di circa 37 secondi e che quindi il tempo della Wang potrebbe considerarsi 'pulito'.

Il risultato ottenuto con la stima puntuale è interessante ma per averne uno più completo si ricorre alla stima per intervalli.

Per costruire l'intervallo di confidenza si determina la log-verosimiglianza profilo (profile log-likelihood) di x_{ult}, esplicitando uno dei parametri della log-verosimiglianza costruita in precedenza rispetto all'espressione

$$x_{ult} = \alpha - \beta + \sigma/\lambda$$

$$l(\alpha,\alpha - x_{ult} + \sigma/\lambda,\gamma,\sigma,\lambda) =$$

$$= \log \prod_{1=1}^{21} \partial x(1-\exp(-(1 - \lambda(x_i - \alpha +$$

$$+ (\alpha - x_{ult} + \sigma/\lambda(1-\exp(-\gamma t)))/\sigma)^{-1/\lambda}))$$

e sostituendo a x_{ult} una serie di valori in funzione dei quali calcolo $l(\alpha,\alpha - x_{ult} + \sigma/\lambda,\gamma,\sigma,\lambda)$.

La figura 2 mostra la curva della log-verosimiglianza profilo con in ascissa i tempi associati a x_{ult} e in ordinata i risultati della log-verosimiglianza.

A partire dai valori della log-verosimiglianza, sapendo che

$$-2\log \frac{l(\alpha,\beta,\gamma,\sigma)}{l(\alpha,\alpha - x_{ult} + \sigma/\lambda,\gamma,\lambda,\sigma)} \sim X_1^2$$

l'intervallo al 90% di x_{ult} è dato da tutti i valori tali che

$$-2\log \frac{l(\alpha,\beta,\gamma,\lambda,\sigma)}{l(\alpha,\alpha - x_{ult} + \sigma/\lambda,\gamma,\lambda,\sigma)} \leq 2{,}706.$$

La retta inserita nel grafico indica che i valori inferiori cadono nell'intervallo, da cui l'intervallo di confidenza al 90% di x_{ult} è (1639,06;1781,99) che contiene il tempo realizzato dalla Wang.

Stima del valore estremo - seconda tecnica

È interessante giudicare il tempo di un atleta confrontandolo con altre buone performance riferite allo stesso anno. Queste performance possono essere incorporate all'interno di un modello di valori estremi usando la distribuzione congiunta di r statistiche ordinate.

Questo approccio è basato sulle stesse argomentazioni asintotiche che giustificano la *GEV* per il minimo annuale, con in più il requisito che r sia fissato e che i valori estremi della variabile siano asintoticamente indipendenti.

Per il generico anno t la funzione di densità congiunta è:

$$g_t(\underline{x}) = (-1)^r \left(\prod_{i=1}^{r} \frac{\overline{G}_t\,'(x^{(i)})}{\overline{G}_t(x^{(i)})}\right) \overline{G}_t(x_t^{(r)}),$$

dove $x_t^{(1)} \leq x_t^{(2)} \leq \ldots \leq x_t^{(r)}$ sono gli r migliori tempi nell'anno t, G_t è la funzione di sopravvivenza data da 1-G, e infine

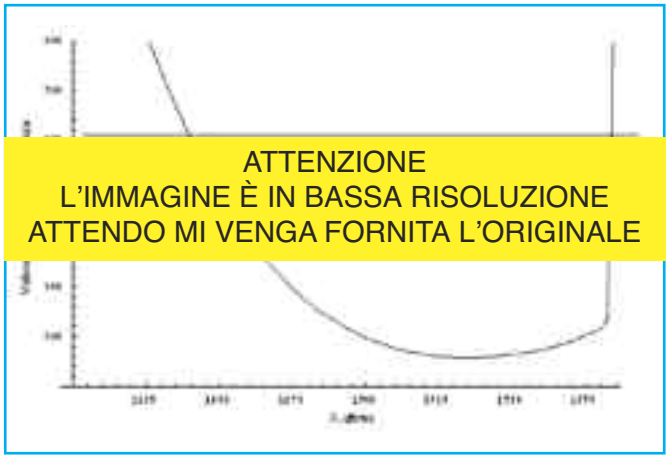

Figura 2 - Log-verosimiglianza profilo di x_{ult}

$\overline{G}_t\text{‘}(x) = \partial \overline{G}_t(x)/\partial x.$

Il problema da considerare è che i tempi corsi da uno stesso atleta, all'interno dello stesso anno non sono indipendenti, infatti un atleta che realizza un buon tempo è probabile che ne realizzi anche un altro, ma prendendo solo i tempi corsi da atleti diversi nella stessa disciplina nel medesimo anno non ci sono problemi di dipendenza.

Un altro problema associato a questo modello riguarda il giusto numero di valori estremi da prendere per un singolo anno; un r grande da più informazioni ma può portare a violare le condizioni di asintoticità e indipendenza; si è scelto di considerare le prime cinque prestazioni annuali con la funzione di densità che diventa:

$$g_t(\underline{x}) = (-1)^5 \left(\prod_{i=1}^{5} \frac{\overline{G}_t\text{‘}(x^{(i)})}{\overline{G}_t(x^{(i)})}\right) \overline{G}_t(x_t^{(5)}).$$

Come nel paragrafo precedente si procede alla stima dei parametri, alla definizione del tempo limite e dell'intervallo di confidenza cambiando solo la funzione di densità.

La figura 3 mostra i migliori 5 tempi realizzati in ogni anno dal 1978 al 2013, anche in questo caso viene all'occhio la straordinarietà della prestazione dell'atleta cinese, ma si vede anche che fino al 2000, l'unico avvicinamento al record della Wang è sempre nel 1993, non a caso da parte di un'altra atleta cinese; la cosa che desta stupore è che tutte e cinque le migliori performance di quel anno furono realizzate a Pechino nello stesso giorno, nella stessa gara, da atlete dello stesso paese!

I valori dei parametri stimati con questo metodo sono: α =1999,0, β =176,9, σ =19,05, λ =-0,289, γ=0,1675 da cui x_{ult} =1756,18.

Si nota che questo tempo è notevolmente maggiore di quello ottenuto con il primo metodo di stima (1734,51), e di 15 secondi inferiore a quello della Wang (1771,78) che quindi rimane consistente.

Anche l'intervallo di confidenza da risultati differenti (1620,4;1771,85); si abbassa di circa 20 secondi il limite inferiore e di 10 il superiore.

Conclusioni

L'esercizio svolto si proponeva di applicare delle tecniche statistiche basate sulla teoria del valore estremo per stimare la migliore prestazione possibile di una popolazione di atleti in competizione in una disciplina.

In particolare l'attenzione è stata rivolta ai 10 mila metri femminili, specialità in cui nel 1993 l'atleta cinese Junxia Wang fece registrare una 'pre-

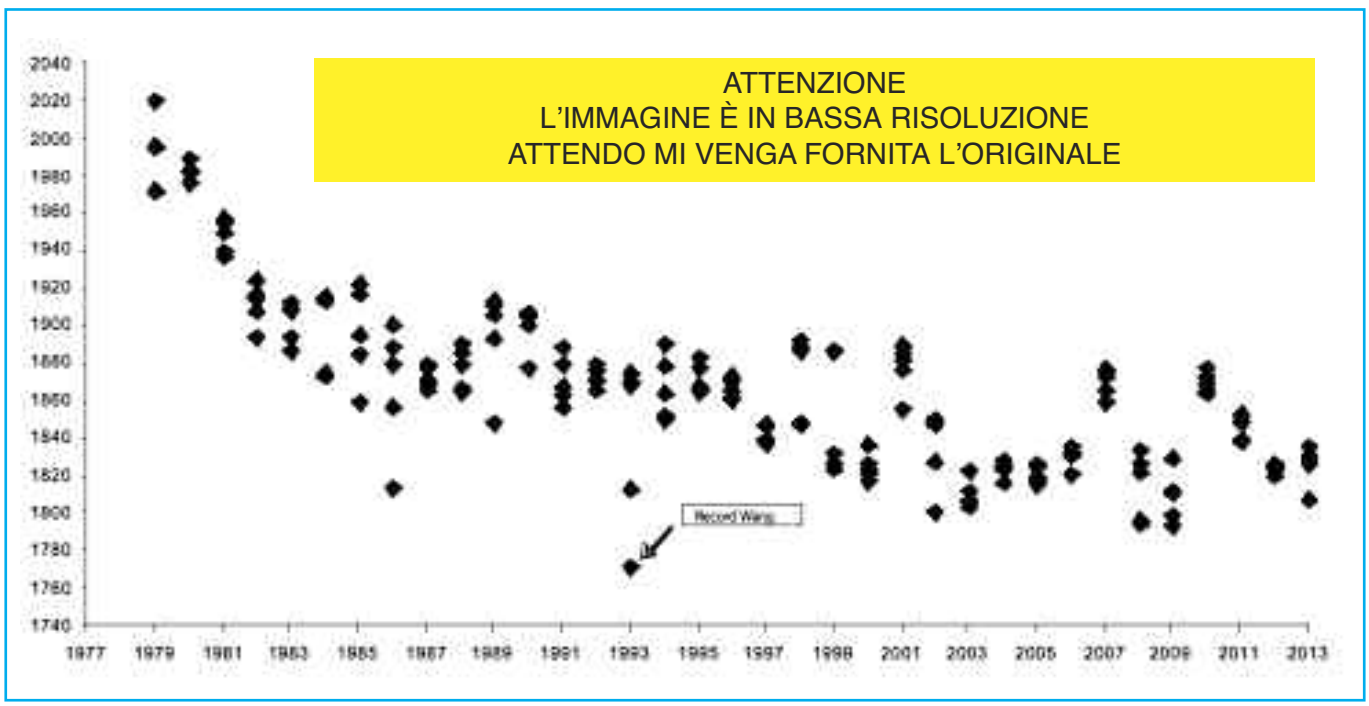

Figura 3 - Migliori cinque tempi nei 10 mila metri femminili dal 1978 al 2013. (In secondi)

stazione eccezionale' mai eguagliata in seguito. L'obiettivo è stato stabilire il tempo limite per questa specialità e confrontarlo con il record per vedere se la prestazione è consistente con i tempi realizzati prima e dopo o è frutto di aiuti illeciti (doping) che hanno consentito di andare oltre le possibilità naturali.

In particolare sono stati utilizzati due metodi basati su una generalizzazione della teoria del valore estremo che considerano o solo il miglior tempo realizzato dal 1978 al 2013 o i primi cinque tempi realizzati nello stesso periodo.

Per entrambe le tecniche il record cinese risulta consistente con il valore limite previsto per i 10 mila metri, non si può escludere quindi l'ipotesi di origine naturale del fenomeno secondo cui la 'prestazione eccezionale' sia veramente frutto di componenti naturali e non sia stata aiutata da agenti illeciti.

Una critica alle analisi di tipo statistico-sportivo, è di limitarsi a una sola metodologia per lo studio dei dati, con il rischio di tralasciare aspetti che con un'analisi simultanea di varie tecniche potrebbero essere evidenziati. Questo lavoro utilizza due metodi che comunque si basano sulla stessa teoria, è quindi molto probabile che si trascurino altri aspetti di rilievo.

Per concludere si pone l'attenzione sulla carenza di dati relativi allo sport e all'atletica in particolare; se alle semplici graduatorie che riportano i migliori tempi annuali si potessero aggiungere maggiori informazioni quali le condizioni del tempo durante le gare, il luogo in cui vengono svolte le corse con particolare riferimento all'altitudine, i materiali usati dagli atleti ecc., le analisi statistiche potrebbero essere più accurate.

Riferimenti bibliografici

Cerioli A., D'Arcangelo E. e Sanna F.M. (1997), studio dell'attività sportiva di alta prestazione: i contributi della metodologia statistica nella letteratura internazionale, Statistica e sport:non solo numeri, a cura di A. Mussino, Società Stampa Sportiva, Roma.Mussino A. (1999), "I contributi della metodologia statistica all'atletica leggera", Atletica studi, 3-4/99, pp.11-17.

Prescott P. And Walden A.T. (1980), "Maximum likelihood estimation of the parameters of the generalized extreme value distribution", Biometrika, 67, pp.723-724.

Quercetani, R. L., (1995). "Sfida alla distanza:i magnifici dei 5000 e 10000 metri", MagisBooks Editori (Reggio Emilia).

Robinson M.E. e Tawn J.A, (1997),"Reply to Comment on "Statistics for Exceptional Records", Applied statistics, 46, pp.127-128.

Robinson M.E. e Tawn J.A. (1995), "Statistics for Exceptional Records", Applied statistics, 44,4, pp. 499-511.

Smith R.L. (1986), "Estreme value theory based on the r largest annual events", Journal of Hydrology, 86, pp.27-43.

Smith R.L. (1988), "Forecasting records by maximum likelihood", Journal of the American Statistical Association, 83, pp.331-338.

spizzich@istat.it