Scienza, Dati e Statistica

Anthony C. Davison

Institute of Mathematics,
Ecole Polytechnique Fédérale de Lausanne

Lectio Magistralis, Padova, 18 dicembre 2009

Sono assai dispiaciuto di non saper parlare nella vostra lingua così espressiva e di essere costretto a proporre questa *lectio* in inglese.

Mi permetto di iniziare esprimendo profondo apprezzamento per l'onore che mi fate conferendomi una laurea *honoris causa*. Nel corso degli anni, con le mie visite a Padova, ho maturato un grande rispetto per l'alta qualità dei vostri docenti e studenti ed è particolarmente gratificante per me che il mio lavoro venga riconosciuto così generosamente da una delle più antiche e prestigiose università d'Europa, in special modo nell'occasione di questo anniversario della vostra Facoltà di Scienze Statistiche.

Galileo e Gaia

Questo è un anno di anniversari importanti nella storia della scienza. Quattrocento anni fa, in questo mese, qui a Padova, Galileo Galilei costruì un cannocchiale che riusciva ad ingrandire 20 volte, per mezzo del quale egli rapidamente individuò le quattro lune più luminose di Giove, scoprì che la Via Lattea è composta da miriadi di stelle singole e osservò i crateri e le montagne della Luna. Gli accademici dei tempi più recenti sono abituati alla collaborazione internazionale e alla competizione serrata ma tutto ciò non è una novità: Galileo dovette lavorare molto velocemente perché un olandese era venuto a Venezia per cercare di vendere al Senato veneziano dei cannocchiali rudimentali — di valore immenso per una importante potenza marittima. Il Senato chiese ad un amico di Galileo di esaminare con cura il cannocchiale olandese, cosa che egli fece molto lentamente, e nel frattempo lo stesso Galileo fabbricò uno strumento migliore, un buon esempio di scaltrezza accademica e commerciale. Fatte le sue prime scoperte astronomiche, Galileo le pubblicò nel marzo 1610, appena poche settimane dopo averle compiute. Nessun contrattempo con risultati importanti tenuti in sospeso a causa di processi di revisione editoriale lenti o direttori di riviste scientifiche incompetenti, ma sto divagando ...

Le straordinarie osservazioni di Galileo contribuirono a capovolgere l'immutabile universo aristotelico, ma l'impatto che egli ebbe sulla scienza fu ancora più profondo. Egli non fu il primo ad utilizzare ciò che noi oggi chiamiamo metodo scientifico — osservazione, sperimentazione e attenta valutazione delle ipotesi alla luce dell'evidenza empirica — anziché l'investigazione attraverso il puro pensiero come propugnato dopo Aristotele dai filosofi peripatetici, ma fu il primo ad applicarlo sistematicamente ad un ampio spettro di fenomeni: navigazione, pendoli, azione della gravità, meccanica, idrostatica, magnetismo, resistenza dei materiali, ... Riconosciuto oggi come il padre della scienza moderna, egli intraprese le sue ricerche nelle vesti di

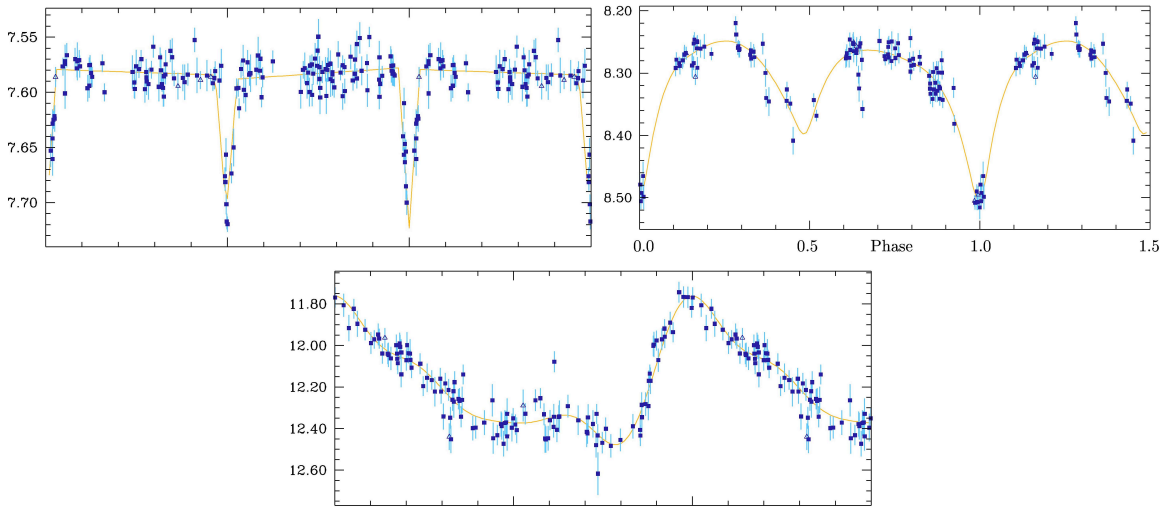
professore di matematica: un precursore dell'impatto che gli scienziati di ambito matematico possono avere sull'intera estensione del sapere. Egli pubblicò numerosi libri importanti, tra i quali *Il Saggiatore*, stampato nel 1623, riassume la sua visione del metodo scientifico. Vi enuncia:

> La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto.

Oggi potremmo aggiungere a questa lista altri caratteri quali derivate, matrici e variabili casuali.

Raccogliere dati di alta qualità è cruciale, ma senza un criterio organizzatore i dati sono solo raccolte arbitrarie di numeri e fatti, ossia ciò che Rutherford chiamava, spregiativamente, collezioni di francobolli. Il ragionamento matematico fornisce strutture coerenti all'interno delle quali classificare i dati, creare collegamenti tra loro, estrapolarne altri e, con l'aiuto di idee statistiche, valutare se divergenze tra i dati e una particolare ipotesi indichino una carenza della teoria corrente.

Dal 1609 l'astronomia ha compiuto un lungo cammino. La missione Gaia è un progetto dell'Agenzia Spaziale Europea, che intende lanciare nel 2012 un satellite che raccoglierà dati su un miliardo di oggetti astronomici. Ne conseguirà il compito di classificare in gruppi distinti circa 100 milioni di stelle variabili, sulla base di dati come quelli mostrati nei grafici sottostanti.



L'occhio umano può svolgere il compito forse per poche centinaia di immagini, ma per ottenere il risultato in modo oggettivo, veloce ed automatico, classificando almeno una stella al secondo, occorre disporre di strumenti di classificazione sofisticati e veloci. Questo è un immenso terreno di gioco per statistici che abbiano familiarità sia con l'astrofisica che con le tecniche di classificazione più recenti.

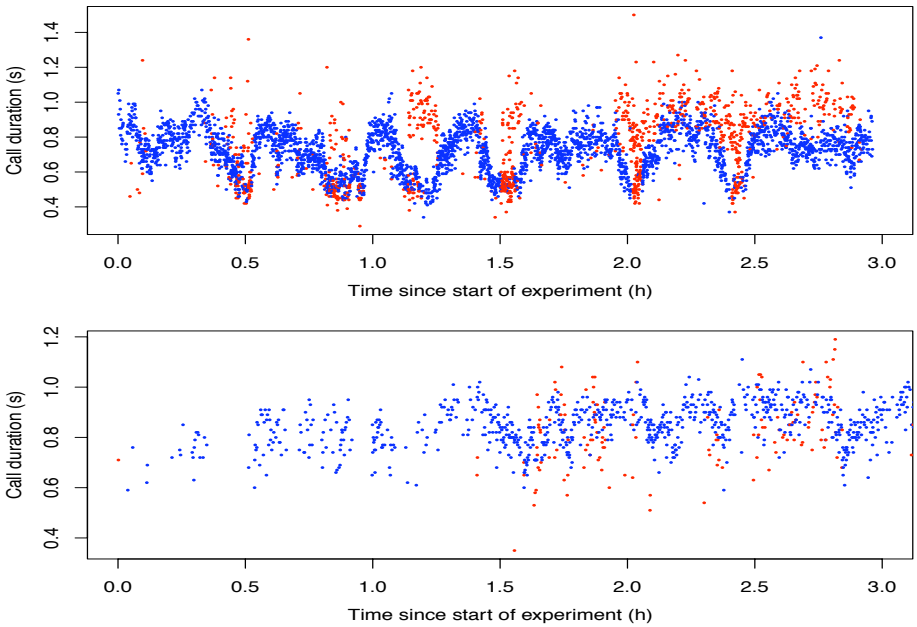
Darwin e l'evoluzione

L'anno 2009 segna anche i 150 anni dalla pubblicazione del libro di Charles Darwin *The Origin of Species*. Così come l'apporto di Galileo alla rivoluzione copernicana contribuì a mostrare che il nostro pianeta non è al centro dell'universo, la grande intuizione di Darwin mostrò che il genere umano non è unico nel creato, che ogni essere vivente è legato, tramite un albero evolutivo, a tutti gli altri, e forse a tutti gli esseri vissuti nel passato. Darwin, come Galileo, era un acuto osservatore e un assiduo sperimentatore e, come Galileo, ottenne molti dati, alcuni dei quali furono poi usati dal genetista e statistico R. A. Fisher nel suo grande libro *Design of Experiments* al fine di spiegare l'importanza della casualizzazione. Questo principio statistico ha forse dato un contributo alla salute del genere umano più importante di ogni specifica scoperta medica, attraverso la creazione della sperimentazione clinica controllata con tutte le sue varianti, e grazie al graduale delinearsi della medicina basata sulle prove d'evidenza. È memorabile l'affermazione di Fisher:

> Chiamare uno statistico dopo che l'esperimento è stato condotto non può essere altro che chiedergli di effettuare un'autopsia: egli potrà dire di cosa è morto l'esperimento.

La programmazione degli esperimenti è così sottovalutata che molti statistici qui presenti avranno avuto la spiacevole esperienza di aver dovuto spiegare a colleghi ricercatori che i loro dati, così faticosamente acquisiti, non possono in alcun modo fornire le conclusioni sperate per colpa della mancata inclusione di un gruppo di riferimento, che andava fatta utilizzando dei casi di controllo, o per colpa della mancata casualizzazione, indispensabile per evitare distorsioni, o ancora per colpa del mancato controllo della variabilità attraverso l'uso della stratificazione.

Anche la biologia, come l'astronomia, ha percorso una lunga strada. Ecco dei dati tratti da una serie di esperimenti condotti da Amélie Dreiss e Charlène Ruppli all'Università di Losanna sulle conversazioni tra piccoli di civetta.



L'obiettivo è capire se gli schemi di fonazione di ogni piccolo possano essere caratterizzati e se il loro essere affamati influenzi le conversazioni, al fine di capire le trattative tra i piccoli prima che un genitore torni al nido con il cibo. La teoria dei giochi evolutiva suggerisce che gli uccelli dovrebbero reagire agli stimoli reciproci in modo da rafforzare la lealtà di nidiata e assicurare un'allocazione ottimale delle risorse. Una volta che i dati sono stati registrati e trattati acusticamente, inizia la sfida statistica: è possibile caratterizzare i richiami di ognuno dei piccoli di civetta? Qual è la dinamica del loro dialogo? Esistono regole che determinano chi parla, e quando, e chi riceve il prossimo pasto?

I modelli standard per serie temporali difficilmente possono essere d'aiuto in questo caso e possono essere necessarie idee innovative per dare risposte incisive a queste domande. Il commento di Fisher vale in parte anche per questa applicazione: poiché un piccolo di civetta non è mai scelto due volte, ma è considerato sempre con un fratello, è impossibile rispondere a quesiti sulle caratteristiche individuali e tutto ciò che può essere descritto è il comportamento in presenza di un'altra civetta.

Insiemi di dati grandi e piccoli

I due problemi che ho descritto esemplificano come le attuali tecnologie di misurazione e di elaborazione dei dati forniscano agli studiosi una quantità senza precedenti di dati complessi, per i quali le idee e gli strumenti statistici sono più essenziali che mai. Questo è vero non solo in ambito scientifico ma anche commerciale: quando selezioniamo i link di una pagina web, vengono effettuati esperimenti casualizzati per confrontare le nostre reazioni di fronte a pubblicità differenti. Non per nulla il *Chief Economist* di Google, Hal Varian, dichiara in un'intervista di inizio 2009

> Continuo a dire che il lavoro *sexy* nei prossimi dieci anni sarà quello dello statistico. La gente pensa che stia scherzando, ma . . . l'abilità di prendere possesso dei dati — essere capaci di capirli, elaborarli, estrarne un significato, visualizzarli, comunicarli — questa sarà un'abilità immensamente importante nei prossimi decenni.

Uno dei ruoli chiave di una Facoltà di Scienze Statistiche come la vostra è formare giovani studiosi che possano essere all'altezza di questa sfida e riescano a raccogliere i frutti derivanti dal possesso di abilità oggi molto richieste.

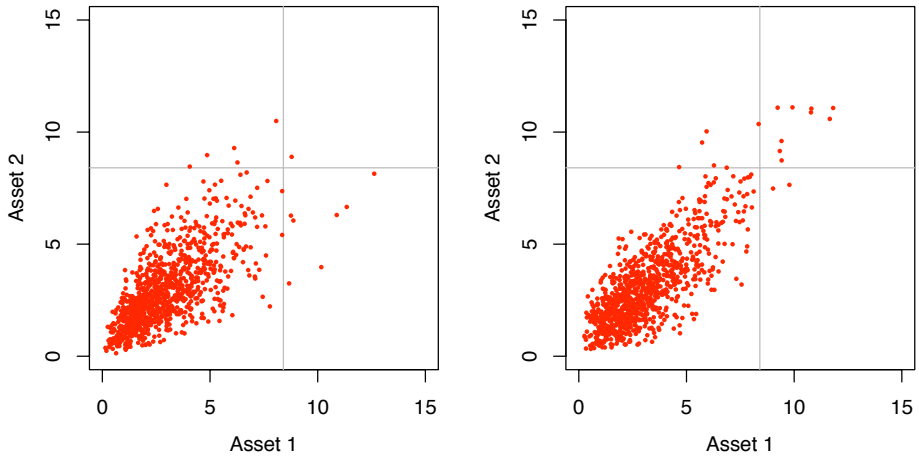
Enormi basi di dati pongono grandi sfide in termini sia statistici che computazionali, ma questi dati possono ridursi moltissimo quando si deve rispondere ad una domanda molto specifica. Per esempio, ci si attende che il Large Hadron Collider (LHC) al CERN vicino a Ginevra, arrivando finalmente a piena operatività dopo una serie di ritardi, getti luce sull'origine dell'universo e sulla correttezza del *modello standard* della fisica delle particelle. Uno degli obiettivi principali è l'attesa scoperta del bosone di Higgs, la cosiddetta 'particella di Dio', che si presume abbia un ruolo chiave nel dare massa alle altre particelle ma che è l'unica particella la cui esistenza non sia stata ancora confermata. Tuttavia, anche se il LHC produrrà più dati dell'intera rete di telecomunicazioni europea, una volta che sarà stato rimosso il rumore, la scoperta o meno del bosone di Higgs dipenderà da poche osservazioni e saranno necessari metodi statistici adatti per piccoli campioni. Molti dei presenti hanno contribuito a sviluppare tali

procedure, che prendono corpo dalle nozioni fondamentali dell'inferenza statistica — verosimiglianza, sufficienza, ancillarità, marginalizzazione, e così via — introdotte da Fisher circa 90 anni fa, ma che ancora oggi sono assai rilevanti. È l'intreccio di queste idee generali, che derivano da considerazioni filosofiche e matematiche, con l'enorme varietà di possibili applicazioni che conferisce alla statistica un così grande interesse e fascino — in aggiunta, ovviamente, al brivido intellettuale dell'investigazione scientifica in un pagliaio di dati.

Eventi rari

Per concludere, vorrei volgermi al mio attuale interesse primario, lo studio degli eventi rari condotto tramite la statistica degli estremi. Due fenomeni recenti lo motivano fortemente.

Il primo è l'attuale crisi della finanza globale, la quale sta vacillando dopo una serie di shock importanti, con altri forse a venire. Una visione popolare di questi avvenimenti, sostenuta da giornali e programmi televisivi, è che *una formula matematica ha fatto scoppiare Wall Street*, fornendo un modo semplice per misurare il rischio di due insolvenze simultanee — un modo semplice, ma fondamentalmente errato. Secondo questa visione la copula gaussiana è stata utilizzata su vasta scala dalle banche e da altre istituzioni finanziarie le quali hanno, di conseguenza, pesantemente sottostimato certi rischi perché, come mostrato dal primo dei due grafici seguenti,



in questo modello eventi estremi, come il crollo simultaneo di due grandi istituti finanziari, non possono accadere. In realtà, come si è visto, questi eventi possono accadere con una probabilità spaventosamente alta perché, come molti rispettabili accademici avevano detto ben prima che la crisi colpisse, la copula gaussiana è decisamente troppo semplice per permettere un'analisi matematica realistica di questi rischi. Questi rischi permangono anche dopo gli shock che abbiamo sperimentato di recente e un'importante sfida statistica consiste nel trovare metodi migliori per stimarli e poi, cosa ancora più difficile, cercare di controllare che questi metodi vengano applicati correttamente.

Una seconda sfida ancora più importante è determinata dalle verosimili conseguenze del cambiamento climatico. Anche considerando uno scenario ottimistico, un aumento di 2–4°C

della temperatura media globale sembra ormai inevitabile, e ciò avrà ripercussioni profonde sul nostro ambiente — sui raccolti e sull'acqua dai quali dipendiamo, ad esempio — e, dunque, forse, sulla nostra sopravvivenza come specie. Eventi come l'ondata di calore dell'estate del 2003, che ha determinato, si pensa, circa 37.000 decessi in Europa, saranno verosimilmente normali nel 2050. Tuttavia, al momento, non abbiamo un metodo accettato e matematicamente solido per modellare le probabilità e le conseguenze realistiche di questi disastri. Tempeste, nubifragi, inondazioni, siccità e altre calamità saranno anch'essi, secondo le previsioni, più frequenti e colpiranno ambienti urbani e naturali la cui sicurezza oggi diamo per scontata. Costruzioni attualmente in fase di realizzazione in certi luoghi dovranno resistere ad eventi mai visti prima e gli ingegneri dovranno basarsi su previsioni determinate da modelli matematici e statistici che ad oggi sono ancora poco compresi e che giocheranno un ruolo fondamentale nella adeguata valutazione del livello di incertezza. Fornire questo tipo di valutazioni è compito specifico della statistica, ma se i risultati devono avere una reale utilità, sarà essenziale la stretta collaborazione tra studiosi di molte discipline differenti. È compito di ognuno di noi affrontare queste sfide sfruttando al meglio le nostre capacità di modo che possano essere assunte decisioni razionali, ispirate ai valori più alti della nostra comune umanità.

Conclusione

Permettetemi di riesprimere la mia profonda gratitudine per questa laurea *honoris causa*.