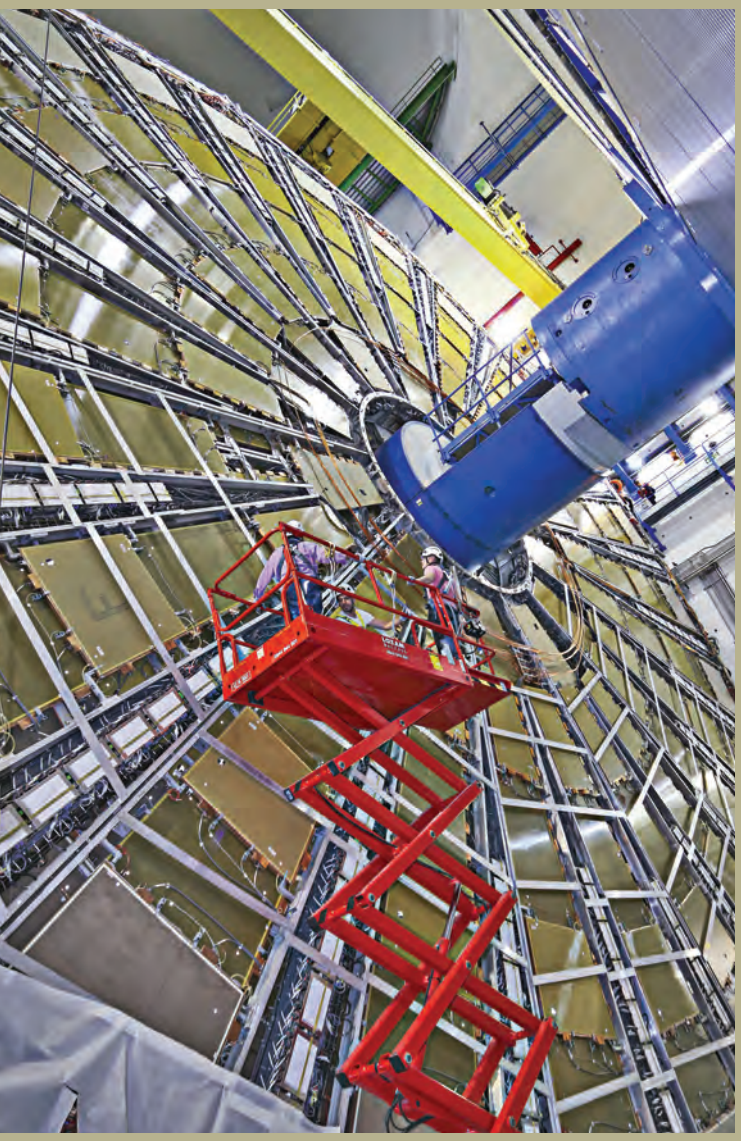
# Per qualche sigma in più

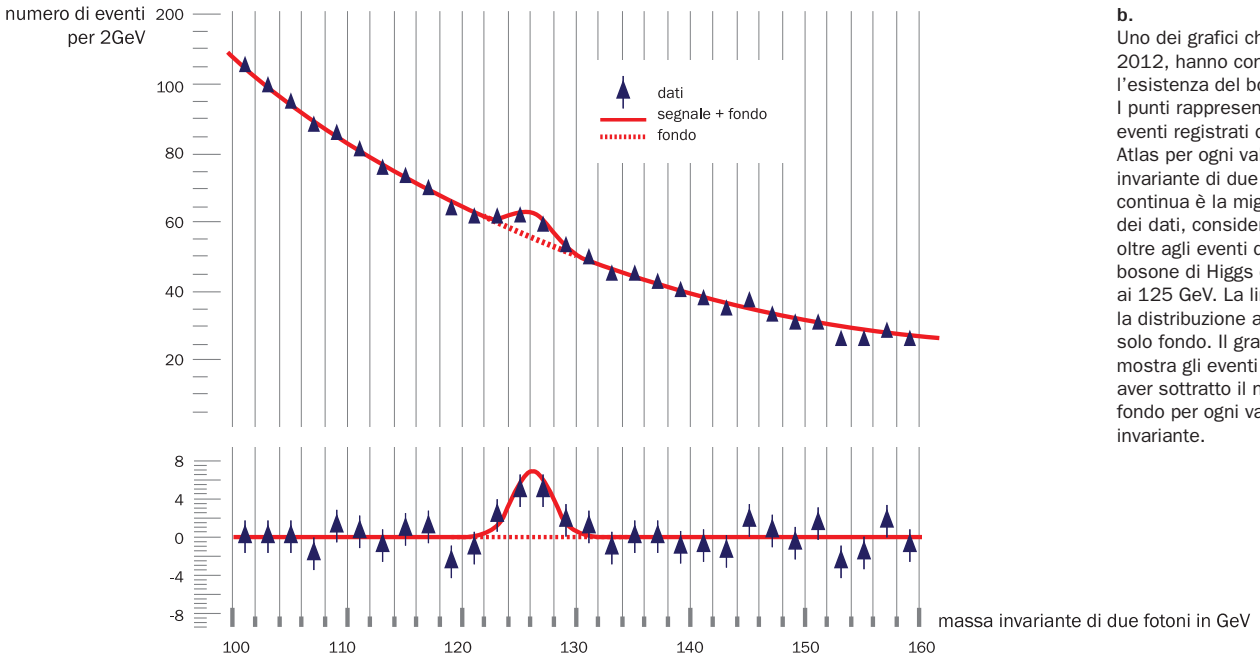
## Metodi statistici in fisica delle particelle

di Concezio Bozzi



**a.**
Uno scorcio dell'esperimento Atlas al Cern.

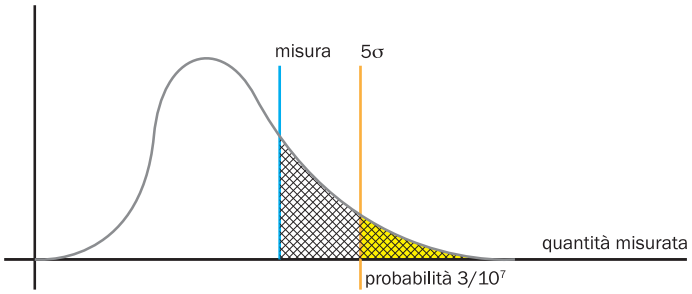
Il sogno di ogni fisico sperimentale è di poter lavorare con una "statistica infinita", vale a dire con campioni di dati grandissimi che permettono di misurare le variabili osservabili sotto studio (chiamate dai fisici semplicemente "osservabili") con incertezze trascurabili. Esiste infatti un'incertezza "statistica" in ogni misura di fisica, che dipende dalla dimensione dei dati analizzati e che si riduce in proporzione all'inverso della radice quadrata delle dimensioni del campione analizzato. Più grande è il campione, minore sarà l'incertezza statistica, e più facile sarà stabilire la verità o la falsità di un'ipotesi, come per esempio l'esistenza di nuove particelle o di nuovi processi al di là di quello che è comunemente noto come modello standard.
In pratica, esistono parecchi fattori che limitano la grandezza dei campioni a nostra disposizione. Tanto per cominciare, il tempo a nostra disposizione è necessariamente finito. Potremmo "aumentare la luminosità" costruendo acceleratori che siano in grado di fornire più collisioni al secondo e realizzando rivelatori che siano in grado di raccogliere i risultati di tutte queste collisioni. Tuttavia, i processi più interessanti sono anche molto rari, per cui anche con acceleratori ad altissima luminosità rimaniamo con un campione di dati esiguo, in cui le fluttuazioni statistiche potrebbero indurci a credere di vedere qualcosa di nuovo, quando in realtà abbiamo a che fare solamente con processi noti. Cerchiamo di approfondire con un esempio concreto.
Il 4 luglio 2012 viene annunciata al Cern la scoperta del bosone di Higgs. La scoperta viene fatta tramite studi il cui risultato si può esemplificare con il grafico in fig. b. Nel grafico si riporta la distribuzione di una variabile, che i fisici chiamano "massa invariante" (vd. in Asimmetrie n. 19 p. 16, ndr), ottenuta mediante un'analisi effettuata su particelle rivelate dall'apparato sperimentale (due fotoni nel caso della fig. b), e successivamente ricostruite e selezionate. Se le particelle provengono dal decadimento di un'altra particella, allora questa analisi dà invariabilmente la massa di questa particella, entro gli errori sperimentali, e quindi un picco nel grafico. Il grafico di fig. b mostra effettivamente un picco di "segnale" e un "fondo" monotonamente decrescente al di fuori.



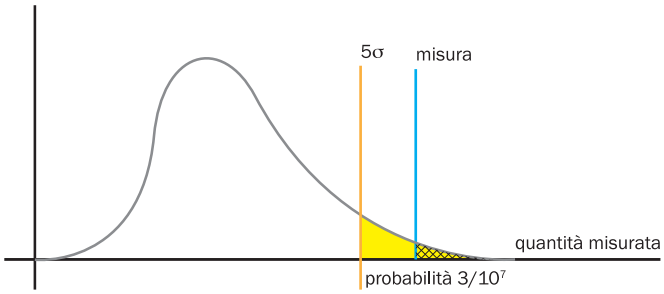

**b.**
Uno dei grafici che, nel luglio 2012, hanno confermato l'esistenza del bosone di Higgs. I punti rappresentano il numero di eventi registrati dall'esperimento Atlas per ogni valore della massa invariante di due fotoni. La linea continua è la miglior descrizione dei dati, considerando che esista, oltre agli eventi di fondo, anche un bosone di Higgs con massa intorno ai 125 GeV. La linea tratteggiata è la distribuzione attesa nel caso di solo fondo. Il grafico inferiore mostra gli eventi che restano dopo aver sottratto il numero di eventi di fondo per ogni valore di massa invariante.

Osserviamo però come il picco non sia così distinto dal fondo – siamo sicuri di aver scoperto una nuova particella? Oppure forse si tratta semplicemente di una fluttuazione? Se osserviamo la distribuzione dei punti sperimentali attorno alla linea tratteggiata, che rappresenta la nostra stima migliore del fondo, vediamo che essi fluttuano un po' sopra e un po' sotto – potrebbe essere perfettamente plausibile che due punti vicini fluttuino entrambi positivamente e simulino un picco.
Come ne usciamo? Innanzitutto notiamo che due esperimenti distinti, Atlas e Cms, osservano un eccesso rispetto al fondo esattamente nello stesso punto del grafico. Diventa già un poco meno plausibile che si tratti di una fluttuazione. Inoltre, possiamo ricostruire un decadimento in particelle diverse da quelle utilizzate per ottenere la fig. b. Otteniamo di nuovo un eccesso rispetto alla previsione per il fondo, sempre nello stesso punto. Il passaggio fondamentale è che riusciamo ad avere una stima quantitativa di quanto sia probabile che i processi di fondo producano quanto osservato sperimentalmente. Mettendo tutto assieme, giungiamo alla conclusione che questa probabilità è minore o uguale a 3 parti per dieci milioni, che è il limite convenzionale (le famose "5 sigma") che i fisici utilizzano per stabilire che c'è qualcos'altro oltre il fondo: una scoperta.

Ricapitoliamo: abbiamo aggiunto i dati raccolti da più esperimenti e abbiamo aggiunto più modi di decadimento della particella di cui siamo a caccia. In altri termini, abbiamo reso il nostro campione statisticamente più significativo. Poi abbiamo definito un cosiddetto "livello di confidenza" dei nostri risultati, dicendo che se l'ipotesi dell'esistenza dei soli processi di fondo fosse vera, e ripetessimo l'esperimento dieci milioni di volte, allora otterremmo quanto osservato sperimentalmente al massimo 3 volte. Come facciamo a stabilire questo?
Per fortuna abbiamo degli strumenti che ci aiutano a lavorare anche con un campione di dati finito, e uno dei più potenti a nostra disposizione è la possibilità di simulare i processi sotto esame, in maniera del tutto analoga a come li ricostruiremmo nei nostri rivelatori, con il vantaggio di poter ottenere campioni molto più grandi di quanto qualsiasi acceleratore possa fornire. Per fare ciò alla scala richiesta dagli esperimenti a Lhc utilizziamo la Grid, l'infrastruttura di calcolo distribuita (vd. in Asimmetrie n. 13 p. 21, ndr). Il modo in cui stimiamo il livello di confidenza di un'ipotesi utilizza proprio grandi campioni di dati che sono simulati secondo l'ipotesi sotto esame. Le simulazioni in questione sono dette "giocattolo". Infatti anche l'enorme potenza di



1.



2.

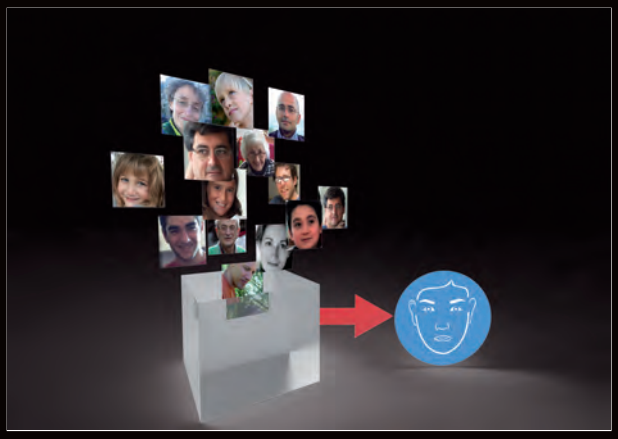
calcolo della Grid non basta a simulare in dettaglio tutte le interazioni che le particelle produrrebbero nei nostri rivelatori, e ci si deve accontentare di parametrizzare unicamente le distribuzioni finali delle osservabili misurate, per esempio la massa invariante o il numero di volte in cui si verifica un processo ritenuto raro o impossibile (vd Asimmetrie n. 15 p. 15, ndr). Anche con questi limiti, si tratta di una simulazione che costituisce un modo pratico e sufficientemente accurato per capire quanto sia affidabile la verifica di un'ipotesi e che genere di incertezze siano legate al campione di dati reali a nostra disposizione. Nella simulazione, generiamo un numero altissimo di campioni, ciascuno statisticamente equivalente a quello raccolto dall'esperimento. Ciascun campione è generato in base all'ipotesi sotto esame – processi di solo fondo, oppure processi in cui oltre al fondo c'è anche il segnale. Effettuiamo la nostra analisi su ciascun campione, e facciamo la distribuzione dei risultati ottenuti. È come se avessimo ripetuto il nostro esperimento un numero di volte pari al numero di campioni simulati. Confrontiamo adesso il risultato ottenuto sui dati reali con la distribuzione dei risultati simulati. Se lo troviamo parecchio al di fuori di quanto previsto in base all'ipotesi che stiamo testando, allora possiamo rigettare questa ipotesi, con un livello di confidenza pari alla frazione dell'area della distribuzione che si trova oltre il valore osservato sperimentalmente (vd. fig. c). La simulazione è inoltre utile per capire quali sono le incertezze sistematiche sulla nostra misura, dovute ad esempio all'incertezza con cui conosciamo la nostra ipotesi di "test" oppure all'incertezza con cui conosciamo la risposta dei nostri rivelatori alle particelle che li attraversano. Infine, la stessa tecnica consente una stima delle incertezze in processi intrinsecamente non ripetibili. Ad esempio, le osservazioni cosmologiche utilizzano l'intero universo come laboratorio, e non c'è modo di ripetere l'esperimento. Tutte le osservazioni cosmologiche hanno un'incertezza osservativa che viene generalmente indicata come "varianza cosmica", e che può essere stimata con tecniche simili.

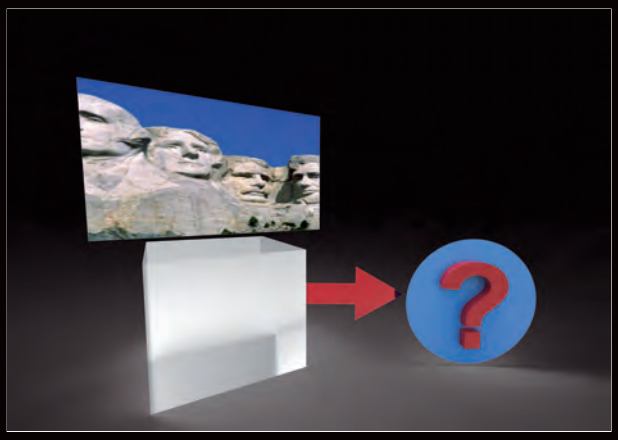
Un altro concetto importante nelle nostre misure è quello di "purezza", definita come la frazione di eventi di segnale presenti nel nostro campione complessivo. A parità di segnale, un campione più puro sarà anche statisticamente più significativo o, equivalentemente, un campione più piccolo ma più puro sarà statisticamente equivalente a un campione più grande e meno puro. Tornando di nuovo alla fig. b, il picco risulterebbe molto più convincente se il fondo fosse cento volte minore. Riusciamo a fare questo? Negli ultimi anni i fisici delle particelle hanno imparato a utilizzare con successo tecniche di *machine learning* in uso in campi del tutto diversi, come la classificazione delle immagini o la diagnostica medica. In fin dei conti, la separazione tra segnale e fondo nei nostri esperimenti è un problema concettualmente equivalente a quello del riconoscimento facciale, affrontato e brillantemente risolto nei programmi di fotografia esistenti su qualsiasi smartphone (vd. approfondimento).

**c.**
Nello studio dei processi rari una delle osservabili interessanti è il numero di eventi prodotti dal processo raro in questione (il segnale). Questo numero deve essere confrontato con quello dovuto a processi diversi e in generale meno rari (il fondo). Attraverso le simulazioni "giocattolo" si può produrre molte volte un possibile risultato dell'esperimento nel caso che ci siano solo eventi di fondo (la curva in figura). L'area corrispondente alla parte di curva superiore a un determinato valore della quantità misurata indica la probabilità che un risultato maggiore o uguale a tale valore sia dovuto all'esistenza del solo fondo. Per affermare che il valore misurato non può essere prodotto dal solo fondo, i fisici hanno scelto il limite convenzionale di 3 parti per dieci milioni (le famose 5 sigma) indicate dall'area gialla in figura. L'area tratteggiata in figura esprime invece la probabilità di ottenere un valore maggiore di quello misurato.
Nel caso indicato in 1) la misura è compatibile con l'esistenza di solo fondo, mentre in 2) possiamo affermare che il solo fondo non è sufficiente a spiegare la misura, che può essere quindi giustificata solo con l'esistenza di un segnale dovuto al processo raro.

[as] approfondimento

# Indovina chi





1.
Schema di funzionamento di un classificatore (rappresentato in figura dalle scatole).

Vi è mai capitato di postare foto su Facebook e di vedere apparire l'invito a "taggare" i volti dei vostri amici, magari con il suggerimento del nome "è Stella? è Giovanni?". Come fa Facebook a sapere che in una foto sono ritratte delle persone e addirittura a riconoscerle? Le tecniche utilizzate fanno uso di "reti neurali", o più genericamente di cosiddetti "classificatori", che sono capaci di imparare a distinguere, a riconoscere le forme (in inglese *pattern recognition*). Per insegnare a un classificatore a riconoscere i volti (che sono il nostro "segnale") bisognerà fargli prima analizzare varie immagini di volti, e solo quelle. Il classificatore per esempio imparerà che in un volto ci devono essere due occhi, un naso e una bocca. Che ci possono essere i capelli, ma non è sempre detto, che ci possono essere dei baffi, e così via... In una seconda fase sarà necessario passare al classificatore solo immagini che non sono volti (cioè sono il nostro "fondo"): scorci di mare, montagne, automobili,... Alla fine del processo di apprendimento, il programma di classificazione produrrà un algoritmo, capace di analizzare un'immagine e calcolare la probabilità che questa raffiguri un volto (segnale) o qualcos'altro. Il risultato è una probabilità, perché c'è sempre una certa frazione di casi in cui l'algoritmo sbaglia, invitandovi a "taggare" il volto di un amico dove c'è in realtà una parete rocciosa, o non riconoscendo che un'immagine ritrae una persona. Questo può capitare perché la roccia può avere dei tratti che ricordano la struttura minima di un volto (occhi, naso, bocca) oppure perché la persona è ritratta di profilo. Gli algoritmi migliori sono quelli che sono stati "allenati" su un enorme numero di "segnali" (anche molto particolari, come una persona di spalle) e di "fondi" (montagne a forma di testa o macchine decorate con una faccia). Negli ultimi anni i fisici hanno iniziato a usare le stesse tecniche per distinguere il segnale, il raro evento cercato, dal fondo, i tanti eventi quasi uguali al segnale. Nel loro caso i classificatori vengono allenati su grandi campioni di dati simulati, sia di segnale che di fondo. Come nel caso delle immagini, il risultato non è mai sicuro al cento per cento, ma l'uso intelligente di queste tecniche ha permesso ai fisici di ottenere molta più informazione sui campioni di dati raccolti e di testare la validità di un'ipotesi o l'accordo dei dati con la teoria. [Barbara Sciascia]


**Biografia**

**Concezio Bozzi** è primo ricercatore presso la sezione Infn di Ferrara, attualmente in congedo presso il Cern. Ha partecipato a esperimenti al Cern e a Slac. Si interessa di fisica del *flavour* e di calcolo scientifico. Svolge attualmente la sua attività di ricerca nell'esperimento Lhcb. È stato membro di comitati di valutazione della ricerca presso l'Infn, il Cern e il Department of Energy degli Usa.


**Link sul web**

http://blogs.scientificamerican.com/observations/five-sigmawhats-that/

http://scikit-learn.org