Statistica bayesiana

G. D'Agostini

30 marzo 2005

Domanda

Ho studiato un poco di statistica: curva di gauss, binomiale, valore medio... Mi hanno detto recentemente che accanto alla statistica di gauss oggi si usa una statistica di bayes. In cosa differiscono?

Risposta

Chiariamo subito che il termine *statistica* ha diversi significati e questo provoca una certa confusione non solo fra la gente comune, ma anche fra i ricercatori che usano regolarmente 'la statistica' nel loro lavoro.

Nel significato originario, da cui trae il nome, essa rappresenta "la scienza che si occupa della raccolta e la classificazione di certi fatti concernenti la popolazione di uno *Stato*" (Webster's). Detto con le parole di Trilussa, *"È 'na cosa / che serve pe' fa' un conto in generale / de la gente che nasce, che sta male, / che more, che va in carcere e che sposa."*. In questa accezione essa è più propriamente nota come *statistica descrittiva*.

Ma per 'statistica' si intende anche il ramo della matematica applicata che si occupa di inferire i 'parametri' di una popolazione a partire da dati statistici parziali (un classico esempio sono i sondaggi preelettorali e gli exit poll), ovvero i 'valori veri' di grandezze fisiche a partire dalle osservazioni sperimenatali, e così via. In questa accezione il termine più appropriato è *statistica inferenziale*. La statistica inferenziale fa uso della *teoria della probabilità* e per questo quest'ultima è spesso vista, e anche insegnata, come una parte della 'statistica'.

Quando si parla di 'statistica gaussiana' (e lo stesso vale per statistica poissoniana, binomiale, di Fermi-Dirac, etc) ci si riferisce al 'comportamento statistico' di una variabile aleatoria, che, per dirla alla buona, si comporta 'in media' secondo la distribuzione di probabilità di Gauss. Questa espressione, e le analoghe per le altre 'statistiche', che avrà pure le sua brava ragione storica, sembra fatta apposta per confondere le persone, soprattutto quando si incontrano altre espressioni simili ma con diverso significato, come 'statistica frequentista', 'statistica classica' e 'statistica bayesiana'.[1] In particolare, quest'ultima non si riferisce ad una inesistente distribuzione di Bayes, come suggerirebbe l'analogia con 'statistica gaussiana'.

Le espressioni *statistica frequentista* e *statistica bayesiana* si riferiscono a due modi diversi di intendere la teoria della probabilità e, di consequenza, di affrontare l'inferenza statistica ('statistica classica' è più o meno sinonimo di 'statistica frequentista' e non ha niente a che vedere con la cosiddetta 'definizione classica' di probabilità — tanto per aumentare la confusione!).

Nella statistica *frequentista* si assume che il concetto di probabilità sia strettamente legato a quello di frequenza (relativa). Più precisamente, secondo questo approccio si può parlare di probabilità soltanto con riferimento agli esiti aleatori di esperimenti ripetuti nelle stesse condizioni. La probabilità *sarebbe* il limite 'per n che tende ad infinito' della frequenza relativa con cui un particolare esito si è verificato in n prove.

Nell'approccio *bayesiano* invece, il concetto di probabilità è semplicemente legato al suo significato intuitivo e all'etimologia (latina) dell'aggettivo probabile, ovvero alla plausibilità che eventi dall'esito incerto possano accadere, o che delle proposizioni possano risultare vere. Questo concetto di probabilità è anche detto *soggettivo* in quanto diverse persone possono avere un diverso stato di informazione e quindi è più che naturale che esprimano diverse valutazioni di probabilità. Come si può facilmente intuire, questo approccio è di più ampia applicazione di quello legato al limite delle frequenze relative. In particolare esso racchiude, come caso particolare e sotto ben

[1] Per completezza, ci sembra doveroso accennare ad altri significati di 'statistica'. A volte si incontrano espressioni del tipo "usare tutta la statistica", "una statistica di un milione di eventi" o "raddoppiare la statistica", ove 'statistica' sta semplicemente per 'quantità di dati sperimentali'. Un'altra accezione di 'statistica' indica una variabile calcolata sui dati sperimentali, ad esempio la media aritmetica. In questo caso il termine più appropriato sarebbe 'riassunto' o 'riassunto statistico'. Quando si incontra l'espressione 'statistica sufficiente', ad esempio riferito alla media aritmetica, ci si riferisce a questo significato del termine.

precise ipotesi, la valutazione della probabilità effettuata usando le frequenze relative con cui un certo tipo di eventi è accaduto nel passato.

Le differenti concezioni di probabilità dei due approcci si riflettono sui metodi di inferenza statistica che da essi derivano.

Nell'approccio bayesiano l'inferenza è basata sulle regole di base della probabilità (essenzialmente i famosi *assiomi* della probabilità) in quanto lo stesso concetto di probabilità può essere applicato sia alle osservazioni (condizionatamente da certe ipotesi) che alle ipotesi (condizionatamente da certe osservazioni). In particolare, la cosiddetta 'inversione di probabilità', che permette di valutare la probabilità condizionate delle varie ipotesi a partire dalle probabilità condizionate delle varie osservazioni, è basata sul *teorema di Bayes*, da cui il nome all'approccio.

Nell'approccio frequentista, invece, è proibito parlare di probabilità di ipotesi, di valori veri, di parametri di un modello o di una popolazione, etc. Di consequenza l'inferenza non può essere *basata* sulle regole di base della probabilità (che comunque servono, per così dire, come ausilio) e bisogna inventarsi dei metodi *ad hoc* per i diversi problemi che si incontrano. I famosi *test di ipotesi* di cui si sente spesso parlare appartengono a tali ingegnose *invenzioni*.

L'approccio frequentista, sviluppato nei primi decenni del XX secolo è ancora quello 'numericamente' dominante anche se difficilmente difendibile a livello teorico e filosofico. In pratica è quello che si insegna comunemente nelle università, a parte eccezioni oggi ancor rare. Per tale motivo esso è anche chiamato 'convenzionale' e quindi l'espressione *statistica convenzionale* fa riferimento a tale approccio.

L'approccio bayesiano, nonostante appaia a molti una novità, si rifà alle idee originarie dei padri fondadori della teoria della probabilità, inclusi Bernoulli, Poisson, Laplace e Gauss. Nella metà del XX secolo era stato praticamente spazzato via dal mondo delle applicazioni dalla scuola frequentista. Ma negli ultimi decenni c'è un deciso revival di questo antico modo di intendere la probabilità e quindi l'inferenza statistica. Tale recupero è stato possibile grazie sia al lavoro teorico chiarificatore di matematici e statistici (fra i quali citiamo l'italiano Bruno de Finetti) che ai grandi progressi nel calcolo sia simbolico che numerico, legati anche all'avvento dei potentissi computer a basso costo che sono presenti oggi in quasi tutte le case e gli uffici. Infatti, sebbene i metodi dell'approccio bayesiano siano in genere concettualmente semplici, essi richiedono, per problemi al di là di quelli elementari da manuale, calcoli complicati (tipicamente *integrali* di funzioni

non elementari e su spazi a molte dimensioni). Una delle ragioni di successo dei metodi frequentisti era quella offrire a ricercatori frettolosi e con scarse conoscenze di matematica formulette risolutive semplici o soluzioni tabulate. Ma, purtroppo, quando si tratta di risolvere problemi veri e si pretende di usare soluzioni 'semplici' la deriva al 'semplicismo' è abbastanza rapida, con risultati talvolta completamente errati e paradossali.

Oggigiorno le applicazioni di metodi bayesiani sono in costante aumento e una ricerca su Google della parola chiave *Bayesian* ne dà un'idea. Si va dai filtri anti-spamming alle applicazioni in medicina e biologia, dall'ingegneria alla finanza e addirittura alla scienza forense. L'approccio più moderno e promettente all'intelligenza artificiale, dopo il fallimento dei tentativi basati sulla logica del certo (booliana), usa le 'reti bayesiane' ('bayesian networks' or 'belief networks') e molto probabilmente chi sta leggendo ora questo testo ha nel suo computer delle reti bayesiane (presenti in Windows dalla versione 98) per aiutare l'utente in caso di difficoltà. Infine, lo stesso Google usa metodi bayesiani nel suo motore di ricerca e addirittura ritiene interessanti, come possibili futuri collaboratori, gli utenti che fanno una ricerca della parola chiave 'bayesian', come si evince dal banner pubblicitario *"We can't hire smart people fast enough!"*.

Per ulteriori informazioni sul teorema di Bayes e sulle sue applicazioni si raccomanda la consultazione delle dispense dell'autore *Probabilità e incertezze di misura* (in particolare il capitolo 5), disponibili in rete all'indirizzo http://www.roma1.infn.it/~dagos/teaching.html. Sulle stesso sito altre note, pubblicazioni e link sull'argomento.