SULL'INFERENZA STATISTICA

Per capire alcuni principi base dell'inferenza statistica proviamo ad immaginare un esperimento volto a conoscere quale sia la proporzione p di bambini italiani di età compresa tra 1 e 10 anni che soffre d'asma.

Il primo passaggio per uno statistico consiste nel DISEGNARE L'ESPERIMENTO nel modo ottimale per l'obiettivo prefissato. Capiamo immediatamente che l'idea di osservare l'intera popolazione con le caratteristiche che stiamo cercando non è praticabile; per questo ci affidiamo alla via INFERENZIALE che si basa sull'idea (un po' strana) di osservare un sottoinsieme della popolazione ed arrivare poi a conclusioni che possano essere generalizzate all'intera popolazione. Ovviamente perché questo abbia senso devono essere rispettate condizioni precise circa il modo in cui tale sottoinsieme viene scelto. Altrettanto ovviamente il risultato finale sarà affetto da un errore: la proporzione di bambini affetti d'asma nel nostro campione sarà probbailmente diversa da quella presente nella popolazione. Molti dei nostri sforzi saranno tesi proprio a quantificare questo errore.

Iniziamo con il definire come popolazione di riferimento quella costituita dai bambini residenti in Italia di età compresa tra 1 e 10 anni. La definizione della popolazione è un elemento essenziale dell'inferenza e richiede molta accuratezza. Pensiamo a quanta attenzione è dedicata alla definizione dei criteri di inclusione ed esclusione nei protocolli sperimentali.

Il principio base per la selezione del campione è che tutte le unità della popolazione abbiano la stessa probabilità di entrare a far parte del nostro campione (campionamneto casuale semplice). Se la popolazione e' finita come nel nostro caso questa probbailità sarà 1/N dove N è la numerosità totale della popolazione (perché?).

Se ritorniamo al nostro esperimento, la via maestra prevederebbe il possesso di una lista contenente tutti i soggetti con le caratteristiche citate, reperibile ad esempio presso le anagrafi. Una volta ottenuta la lista potrei utilizzare un generatore di numeri casuali (o pseudo-causali) avendo associato un numero ad ogni soggetto delle mia popolazione. Avvenuta l'estrazione, andrò a verificare se questi bambini sono o meno affetti da asma. Purtroppo in generale, nel raccogliere i nostri dati, dobbiamo scontrarci con la fattibilità di questa procedura e con i relativi costi. Ne segue che spesso le modalità di campionamento e reclutamento dei pazienti sono diverse dal puro campionamento casuale. Tuttavia il principio base deve restare quello a cui tendere, cercando di controllare le fonti di distorsione.

Nell'esempio dei bambini, dopo aver estratto il campione dalla lista, tramite campionamento casuale, dovremmo definire la variabile risposta. È ormai tempo di rinunciare a pensare alla nostra osservazione come un numero; dobbiamo invece farci carico della sua incertezza descrivendola attraverso una variabile aleatoria. Nel nostro caso quello che andremo ad ossservare sull *i*-simo bambino estratto sarà una variabile aleatoria Bernoulliana X_i che assumerà valore 0 se il soggetto è sano e 1 se il bambino è affetto da asma. In questo modo il campione diventa formalmente un vettore di n variabili aleatorie $(X_1,\ldots,X_n)$ ciascuna delle quali può assumere valore 0 oppure 1 rispettivamente con probabilità $(1-\pi)$ e π.

Concettualmente π e p sono diversi perché il primo è una probabilità che caratterizza X_i la seconda è la proporzione dei bambini affetti d'asma nella popolazione, una costante incognita. Tuttavia il campionamento casuale crea uno stretto legame tra queste due quantità.

Per chiarire la natura di questo legame ricordiamo la definizione di probabilità proposta da Pascal come rapporto tra casi favorevoli e casi possibili, a condizione che tutti i casi siano ugualmente "possibili". Nel nostro caso la condizione di equiprobabilità è assicurata dal campionamento casuale. Ne segue che la probabilità che l'*i*-esimo bambino sia affetto d'asma coinciderà con il rapporto tra il numero di bambini affetti d'asma nella popolazione e la numerosità totale della popolazione stessa. Formalmente avremo π=p. Questo concetto può essere ovviamente esteso a tutte le successive osservazioni.

Osserviamo che quanto detto non dipende dalla numerosità campionaria n, restando valido anche nel caso di una singola osservazione. Tuttavia il ruolo di n diventa cruciale quando le nostre osservazioni dovranno essere utilizzate per stimare π (e quindi p). A questo scopo utilizzeremo come stimatore di π la frequenza relativa dei bambini affetti d'asma nel nostro campione $\hat{\pi}=\frac{\sum_{i=1}^{n} X_i}{n}$. Si noti che questa frequenza relativa può essere letta come media delle osservazioni campionarie.

Anche $\hat{\pi}$ è una variabile aleatoria poiché somma di variabili aleatorie. Lo stimatore è in generale una funzione delle nostre osservazioni e può essere ricondotto ad un momento pre-sperimentale in cui abbiamo disegnato, ma non svolto, l'esperimento e dobbiamo tener conto di tutti i possibili risultati sperimentali (spazio dei campioni). Una volta condotto l'esperimento e osservato i bambini, lo stimatore si trasformerà in un numero cioè in una stima di π. Si noti che che ad ogni campione osservabile nello spazio dei possibili campioni (come è fatto?) corrisponde una frequenza relativa di bambini affetti d'asma, cioè un possibile valore del nostro stimatore. Tuttavia la corrispondenza non è biunivoca poiché in generale ad ogni frequenza relativa corrispondono più campioni osservabili (quali?). Nel deriva che nel passare dal vettore $(X_1,\ldots,X_n)$ allo stimatore $\hat{\pi}$, se da un lato abbiamo operato una sintesi dei dati che ci consente di lavorare nella singola dimensione (siamo passati da R^n ad R), dall'altro abbiamo perso parte dell'informazione originaria (quale? E' una perdita accettabile? Perché?).

Ricapitolando, siamo partiti da p (proporzione dei bambini malati di asma nella popolazione), quantità che vogliamo stimare. Utilizzando il campionamento casuale abbiamo stabilito uno stretto legame tra p e le nostre esservazioni (π=p) che avranno la stessa distribuzione della popolazione di provenienza. Abbiamo, poi, definito uno stimatore di π, $\underline{\hat{\pi}}$. Una volta osservato il campione avremo π_0 che rappresenterà il valore numerico osservato cioè la nostra stima di p. Quale ruolo gioca n? Sappiamo che la frequenza campionaria $\hat{\pi}$ coinciderà con la probabilità π soltanto per n che tende da infinito (ricordate la definizione frequentista di probabilità). In altri termini se potessimo osservare il totale della popolazione conosceremmo esattamente il valore di p. Tuttavia ciò che andiamo ad osservare corrisponde solo un sottoinsieme della popolazione. Ne deriva che il nostro risultato sarà necessariamente affetto da errore. Molta della teoria statistica ruota proprio intorno al modo di quantificare tale errore. Non conoscendo infatti il valore vero di p non avremo mai la possibilità di misurare tale errore come $\pi_0 - p$.

La statistica classica si fonda sul principio del CAMPIONAMENTO RIPETUTO che associa l'errore al fatto che, ripetendo il campionamento (o l'sperimento) a parità di condizioni e numerosità, la stima finale sarebbe diversa. E' proprio a partire da questa diversità che possiamo valutare l'errore, considerando lo spazio dei campioni osservabili a ciascuno dei quali corrisponde un possibile risultato, nel nostro caso una stima di p. Se queste stima saranno molto simili tra loro, l'errore sarà basso; se, al contrario, i valori varieranno in un range ampio, l'errore sarà alto. E' naturale allora valutare l'errore sulla base della varianza del nostro stimatore, misura diretta della variabilità dei valori che potremo ottenere al variare del campione osservato. Vale la pena notare che i campioni osservabili non hanno tutti la stessa probabilità di essere osservati. Immaginiamo che p=0.1. La probabilità di osservare un bambino affetto da asma sara' anch'essa 0.1, cioè molto bassa. Come diretta conseguenza saranno più probabili i campioni che contengono pochi bambini

malati. Se immaginiamo di fare 10 osservazioni avremo $Pr(X_1=0,\ldots,X_{10}=0)=0.9^{10}=0.349$ mentre $Pr(X_1=1,\ldots,X_{10}=1)=0.1^{10}=0.0000000001$.

Si dimostra che $Var(\hat{\pi})=\frac{\pi\,(1-\pi)}{n}$ dove, più elevato è il valore di n minore sarà la variabilità delle nostre stime e quindi minore l'errore assciato al risultato finale. La radice delle varianza, cioè la deviazione standard del nostro stimatore, prende il nome di standard error.

Concludiamo osservando che una elevata numerosità ci garantisce che la stima finale sarà probabilmente vicina a π poichè per n che tende ad infinito $\hat{\pi}$ tenderà a π. Tuttavia solo se il campionamento assicura che π=p questo risultato ha un senso. Se il nostro campionamneto contiene delle distorsioni importanti $\hat{\pi}$ continuera' a tende a π ma π sara' diverso da p, la quantità che vogliamo conoscere. Se estraggo un campione di elevata numerosità dai bambini della mia città avrò una stima affidabile della percentuali di bambini affetti d'asma nella mia città ma non saprò della percentuale riferita alla popolazione Italiana. A meno che io non sia molto molto fortunato…..

Questo esempio è stato sviluppato considerando una varabilie risposta dicotomica. Nel caso di una variabile risposta continua il percorso logico è identico, mutatis mutandis. Vediamone i dettagli.

Ipotizziamo di voler studiare il livello di espressione di un gene nella popolazione di pazienti affetti da una data patolgia. Possiamo ad esempio essere interessati al livello di espressione del gene BRCA1 valuata attraverso la quantità di mRNA o la quantità della proteina che codifica nelle donne affette da carcinoma mammario.

Ipotizziamo inoltre che la sua distribuzione nella popolazione sia appossimabile ad una densità Normale, in altri termini "*assumiamo un modello Normale*". Potete visualizzare questa distribuzione immagianando un istogramma di frequenze che descrive il livello di espressione del gene in tutti i pazienti che appartengono alla nostra popolazione: assumere un modello Normale vuol dire ipotizzare che la forma di questo istogramma sia approssimabile con una densità Normale dove μ sarà il livello medio di espressione del gene nella popolazione e σ^2 la sua varianza. Sottolineiamo il fatto che l'assunzione del modello (non necessariamente Normale) riguarda la distribuzione della variabile oggetto di studio nella popolazione e si basa su conoscenze a priori diverse dai dati che andremo ad osservare. Ogni modello si configura come una famiglia di densità che condividono alcune caratteristiche, quelle che caratterizzano il modello, mantenendo al contempo degli elementi di flessibilità, i parametri, che verranno adattati ai dati sperimentali. La famiglia Normale è caratterizzata dall'esistenza di un valore centrale μ nel cui intorno si concentrano i valori più probabili (più frequenti se pensiamo al il livello di espressione del gene nella nostra popolazione), dalla simmetria rispetto all'asse che passa per questo valore centrale e una probabilità che decresce (secondo σ^2) man mano che ci si allontana da μ . Ribadiamo che assumere un modello Normale vuol dire ipotizzare che la distribuzione del livello di espressione del gene nella nostra popolazione abbia le caratteristiche descritte. Restano incogniti i due parametri del modello che andremo a stimare sulla base dei dati osservati, adattando il modello scelto all'informazione sperimentale. Modello e dati costituiscono due diverse fonti di informazione tra loro complementari che tuttavia dovrebbero essere coerenti. Verificare tale coerenza significa valutare la bontà di adattemento del modello ai dati e, nel caso non sia soddisfacente, deve portare a riconsiderare entrambe le componenti. Questo passaggio è sempre fondamentale poiché, se il modello è di sostegno ai dati, un modello inadatto a descriverli può portare a conclusioni errate.

Immaginiamo poi di aver estratto un camione casuale dalla nostra popolazione che possiamo descrivere come un un vettore di n variabili aleatorie $(X_1,\ldots,X_n)$ dove X_i rappresenta il il livello di espressione del gene

BRCA1 nell'iesama donna reclutata nel nostro studio. A differenza del caso precedente abbiamo adesso una variabile continua che assumerà valori reali e che sarà acaratterizzata da una densità di probabilità. Quale?

Proviamo a domandarci con quale probabilità osserveremo un livello di espressione del gene in un dato intevallo [a,b]. In simboli $Pr(a \leq X_i \leq b) = \int_a^b f(x)dx$, dove f(x) è proprio la densità che cerchiamo. Come nel caso Bernoulliano, se il campionamento è casuale tutte le donne che appartengono alla popolazione condivideranno la stessa probabilità di entrare a far parte dello studio e potremo utilizzare la definizione classica di probabilità in base alla quale tale probabilità è uguale al rapporto tra la frequenza delle donne che hanno un livello di espressione del gene in questo intevallo nella popolazione e la numerosità totale della popolazione stessa. La probabilità che cerchiamo coinciderà anche adesso con la frequenza nella popolazione e l'area sotto le densità di X_i, f(x) sarà uguale all'area del rettangolo che ha base l'intervallo [a,b] nell'istogramma che abbiamo immaginato descriva la distribuzione del livello di espressione del gene nella nostra popolazione.

Poiché possiamo ripetere lo stesso ragionamento per qualsiasi intervallo, la densità f(x) approssimerà l'istogramma e coinciderà con la densità Normale che abbiamo ipotizzato nella nostra popolazione. Come nel caso Bernoulliamo le osservazioni campionarie, intese come variabili aleatorie, erediteranno il modello che abbiamo ipotizzato nella popolazione, $X_i \sim N(\mu, \sigma^2)$. Si noti che ha adesso una doppia interpretazione, μ è il livello medio di espressione del gene BRCA1 nella popolazione delle donne affette da carcinoma della mammella ma anche il valore atteso del livello di espressione del gene nell'iesima donna reclutata. Analogamente σ^2 misura la variabilità del gene nella popolazione ma anche la varianza di X_i.

Come nel caso precedente quella che abbiamo descritto è soltanto una garanzia in probabilità. Perche' il nostro campione sia rappresentativo della distribuzione del livello di espressione del gene nella popolazione dovremo avere una elevata numerosità. Solo questa può garantirci che l'errore campionario sia contenuto come vedremo meglio affrontando il problema di stima di μ e σ^2.