GLOSSARIO SULLA PARTE FINALE DEL CORSO DI MATEMATICA, PARTE 3

INFERENZA STATISTICA SU POPOLAZIONI GAUSSIANE E NON

Inferenza statistica: E' l'insieme delle metodologie che sulla base di osservazioni estratte da una popolazione statistica (distribuzione di una variabile aleatoria X) hanno lo scopo di inferire i valori di alcune caratteristiche di queste popolazioni. Queste inferenze, riguardando delle distribuzioni di probabilita', avranno sempre un margine di incertezza che deve essere adeguatamente quantificato.

Campione casuale: Nella forma piu' semplice di un problema di inferenza le osservazioni sono estratte **indipendentemente** dalla popolazione di interesse. Le osservazioni $X_1, \ldots, X_n$ sono distribuite come una singola osservazione del carattere X oggetto di studio nella popolazione di interesse. I **parametri** di interesse della popolazione sono caratteristiche numeriche della distribuzione di probabilita' di X quali media, varianza, mediana, quantili, ecc.

Verifica delle ipotesi: In problemi di questo genere, si ipotizzano dei valori per un **parametro di interesse** (in presenza o meno di altri parametri, che vengono allora chiamati **parametri di disturbo**). Ad esempio possiamo essere interessati ad una ipotesi sulla media di una distribuzione gaussiana e non alla varianza. L'approccio piu' diffuso a problemi di questo tipo consiste nel determinare una **regione critica** per le osservazioni cui associamo il **rifiuto** dell'ipotesi. In caso contrario si parla di **non rifiuto** dell'ipotesi piuttosto che di accettazione di una **alternativa**, per motivi che saranno presto evidenti.

Statistica test: Le regioni critiche piu' usuali consistono nel superamento di una **soglia critica** da parte di un'opportuna statistica test. Questa e' una variabile aleatoria, funzione dei dati campionari $X_1, \ldots, X_n$ la cui distribuzione sotto tutte le alternative considerate dovrebbe essere "spostata" verso valori piu' grandi che sotto le ipotesi.

Errori di prima e seconda specie: Si ha un errore di prima specie quando la statistica test supera la soglia critica e il modello ipotizzato e' corretto. Si ha un errore di seconda specie qunado la statistica test non supera la soglia ma e' una delle alternatve ad essere corretta. Le probabilita' di questi errori vengono chiamate probabilita' di errore di prima e seconda specie. Nel primo caso la probabilita' di errore puo' variare se l'ipotesi permette una pluralita' di modelli (si pensi di nuovo al caso di popolazione gaussiana con media specificata e varianza qualsiasi). I test che verranno considerati, anche in presenza di una condizione di questo tipo, avranno una probabilita' di errore di prima specie costante, detta **livello di significativita'** del test. Tipicamente il livello di significativita' che si utilizza nell verifica delle ipotesi e' variabile a seconda del tipo di applicazione ma quasi sempre non maggiore di 0.05 (piccolo, per proteggerci nei confronti di rifiuti ingiustificati dell'ipotesi). L'errore di seconda specie, invece, varia con l'alternativa considerata: sara' vicino al complemento ad 1 dell livello di significativita', e quindi molto alto, per alternative molto "vicine" alle ipotesi e scendera' all'allontanarsi dell'ipotesi dall'alternativa. La funzione che associa ad ogni alternativa il complemento ad 1 della probabilita' di errore di seconda specie si dice **funzione di potenza** (probabilita' di rifiutare l'ipotesi sotto le varie alternative). A parita' di livello di significativita' preferiamo ovviamente utilizzare un test piu' potente, ma dato che la funzione di potenza varia al variare dell'alternativa, non sempre e' possibile determinare un test **uniformemente piu' potente**.

Media e varianza della distribuzione della media campionaria: Sia m la media ipotizzata della popolazione da cui si estraggono le osservazioni $X_!, \ldots, X_n$. La media campionaria $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ ha sempre media m, mentre la varianza σ^2 risulta divisa per n (brevemente la ragione: la varianza della somma di variabili indipendenti e' somma nelle varianze, quindi $n\sigma^2$ nel caso di estrazioni dalla stessa popolazione. Dividendo la somma per n la varianza

viene divisa per n^2 e quindi la distribuzione campionaria della media finisce per avere una varianza che e' quella originaria divisa per n. Al crescere di n essa tende a divenire sempre piu' piccola, segnalando che la distribuzione si stringe sempre di piu' intorno alla media (gli errori tendono ad elidersi l'uno con l'altro). Tutto questo vale per una qualsiasi distribuzione di (media e) varianza finita.

La legge dei grandi numeri: Il fenomeno di "restringimento" della distribuzione di $\bar{X}_n$ intorno alla media al crescere di n e' noto come **legge dei grandi numeri**, e vale anche se la varianza non e' finita, anche se occorrono degli strumenti piu' sofisticati per individuarlo. Se non esiste la media, invece, non si ha nessun fenomeno di questo tipo. L'esempio piu' eclatante lo offre la distribuzione di Cauchy. In questo caso la media campionaria di n osservazioni $X_1, \ldots, X_n$ di una stessa densita' di Cauchy continua ad avere la stessa densita' di una sola di esse. Non c'e' quindi alcun vantaggio nell'aumentare il numero delle osservazioni (fintanto che usiamo la media aritmetica per riassumerle, diverso e' il caso se si utilizza, come e' ragionevole, la mediana campionaria).

La particolarita' della gaussiana: Quanto detto precedentemente per popolazioni di varianza (e quindi di media) finita vale in particolare per la gaussiana. Ma in questo caso la distribuzone della media campionaria resta ancora gaussiana e quindi, se m e' la media e σ^2 la varianza, la funzione di distribuzione della standardizzata $Z_n = \frac{\sqrt{n}(\bar{X}_n - m)}{\sigma}$ rimane sempre Φ.

Inferenza sulla media di una popolazione gaussiana con varianza nota: Ora, se $m = m_0$ e' specificato dall'ipotesi, e la varianza σ^2 e' nota, si puo' utilizzare il valore di Z_n per verificare l'ipotesi che la media sia m. Supponiamo che le alternative prescrivano che la media della popolazione sia maggiore di m_0. E' chiaro allora che piu' $Z_n = \frac{\sqrt{n}(\bar{X}_n - m_0)}{\sigma}$ assume un valore grande, piu' tendiamo a rifiutare l'ipotesi. Il valore oltre il quale rifutiamo dipende precisamente dal livello di significativita' scelto. Se questo e' α dovremo rifiutare se $Z_n > \Phi^{-1}(1-\alpha)$ e quindi quando $\bar{X}_n > m_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)$.

Potenza del test: La potenza di questo test, quando la media e' invece $m_1 > m_0$, e' uguale, standardizzando la soglia sotto l'alternativa ipotizzata

$$P(\bar{X}_n > m_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)) = P(\frac{\sqrt{n}(\bar{X}_n - m_1)}{\sigma} > \Phi^{-1}(1-\alpha) - (\frac{\sqrt{n}(m_1 - m_0))}{\sigma}) = 1 - Phi(\Phi^{-1}(1-\alpha) - (\frac{\sqrt{n}(m_1 - m_0))}{\sigma})$$

che e' uguale a α per $m_1 \to m_0+$ e tende a 1 per $m_1 \to +\infty$. Si noti che la funzione di potenza dipende dall'alternativa solo attraverso la quantita' $\delta = \frac{\sqrt{n}(m_1 - m_0))}{\sigma}$, che rappresenta l'effettiva "distanza" tra l'ipotesi m_0 per la media e il valore alternativo m_1, normalizzata rispetto alla deviazione standard della popolazione virtuale di tutte le medie campionarie di dimensione n.

Il livelo di significativita' osservato o p-value: Esiste un modo alternativo alla presentazione dei risultati sperimentali nella forma di rifiuto/non rifiuto dell'ipotesi ad un fissato livello desiderato di significativita'. Si puo' presentare il valore osservato $z_n = \frac{\sqrt{n}(\bar{X}_n - m_0)}{\sigma}$ sul quale si calcola $1 - \Phi(z_n)$, la probabilita' che una variabile gaussiana standard lo superi. Con un po' di riflessione si vede che questo e' il piu' piccolo livello di significativita' che, sulla base del valore osservato di $\bar{X}_n$, porta al rifiuto dell'ipotesi. Riportando questo valore come esito sperimentale lasciamo quindi all'utente finale la liberta' di scegliere il suo livello di significativita' desiderato e concludere la verifica dell'ipotesi confrontando questo livello di significativita' con quello osservato.

Alternative bilatere: Quanto sin qui esposto puo' essere facilmente adattato al caso di alternative $m \neq m_0$ da entrambi i lati rispetto al valore m_0 ipotizzato per la media. In

questo caso (se si intende trattare le alternative a sinistra e quelle a destra in modo analogo) una statistica test ragionevole e' $|Z_n| = \frac{\sqrt{n}|\bar{X}_n - m|}{\sigma}$ che, per dar luogo ad un test di livello di significativita' α, va confrontata con la soglia critica $\Phi^{-1}(1 - \frac{\alpha}{2})$. Corrispondentemente, il livello di significativita' osservato sara' $2(1 - \Phi(|z_n|)$ (se $z_n > 0$ all'area a destra di z_n va sommata quella a sinistra di $-z_n$). La funzione di potenza del test di livello di significativita' α e' una funzione pari di $\delta = \frac{\sqrt{n}(m_1 - m_0))}{\sigma}$, e risulta somma della funzione di potenza del test di livello di significativita' $\alpha/2$ per alternative a destra e di quella con lo stesso livello contro le alternative a sinistra (che e' uguale alla prima cambiando il segno all'argomento). Derivando questa funzione si puo' mostrare che essa e' minima per $\delta = 0$.

Teoria dei grandi campioni, popolazioni non gaussiane: Quanto detto fin qui si applica anche al caso in cui la distribuzione di $X_1, \ldots, X_n$ non e' necessariamente gaussiana ma ha (media e) varianza finita, in virtu' del **teorema del limite centrale** che garantisce la convergenza della distribuzione della media campionaria standardizzata alla distribuzione gaussiana standard (nel senso che gli integrali definiti convergono a quelli della distribuzione limite al tendere all'infinito della dimensione n del campione). La bonta' dell'approssimazione della distribuzione effettiva con quella limite dipende dalla distribuzione di partenza di X (in primis, l'approssimazione e' tanto migliore quanto piu' questa e' simmetrica) e dall'intervallo sul quale viene calcolato l'integrale (lontano dalla media la dimensione campionaria deve essere piu' grande). Quando il modello delle osservazioni e' una popolazione che dipende da un parametro, la varianza sara' spesso una funzione della media. Ad esempio una popolazione bernoulliana ha media p (la probabilita' di successo) e varianza $p(1-p)$. Se $p = p_0$ e' la probabilita' di successo ipotizzata, possiamo ugualmente standardizzare la media campionaria $\bar{X}_n$ (che in questo caso rappresenta la proporzione dei successi sulle n osservazioni campionarie) sotto questa ipotesi, utilizzare la soglia critica della densita' gaussiana standard (asintotica in n), pervenendo alle regioni di rifiuto di livello di significativita' (asintotico) α, rispettivamente nel caso di alternative unilatere a destra, e nel caso di alternative bilatere

$$\bar{X}_n > p_0 + \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}, \quad p_0 - \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} < \bar{X}_n < p_0 + \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}.$$

Il test chi quadrato di addatamento: Il problema della verifica di una ipotesi sulla probabilita' di successo in una popolazione bernoulliana e' di importanza cosi' grande che merita di menzionare la sua generalizzazione al caso di piu' di 2 alternative. Per trattare questo argomento riscriviamo il quadrato della statistica test standardizzata

$$\frac{n(\bar{X}_n - p_0)^2}{p_0(1-p_0)} = (n(\bar{X}_n - p_0))^2(\frac{1}{np_0} + \frac{1}{n(1-p_0)} == \frac{(n\bar{X}_n - np_0)^2}{np_0} + \frac{(n\bar{Y}_n - n(1-p_0))^2}{n(1-p_0)}$$

dove $\bar{Y}_n = 1 - \bar{X}_n$ e' la proporzione di insuccessi nel campione. Questa statistica ha come distribuzione (asintotica per $n \to +\infty$) quella del chi quadrato con 1 grado di liberta', che e' la densita' del quadrato di una gaussiana standard. Nel caso di esperimenti con K modalita' mutuamente escludentisi con probabilita' ipotizzate uguali a $p_0^i, i = 1, \ldots, K$, la naturale generalizzazione di quella scritta sopra e' la somma su tutte le modalita' delle quantita' (numero di osservato di successi-numero atteso di successi per la modalita' al quadrato)/numero atteso di successi, essendo il numero di atteso di successi per la i-esima modalita' sotto l'ipotesi pari a np_0^i. Si puo' dimostrare che la distribuzione asintotica di questa statistica e' quella del chi quadrato con $K - 1$ gradi di liberta', e quindi utilizzare quando n e' grande la soglia critica pari al quantile $1 - \alpha$ della suddetta densita', per avere un test con livello di significativita' (asintotico) pari a $1 - \alpha$.

Popolazioni gaussiane, varianza ignota: In questo caso la deviazione standard che compare nella statistica test va stimata sul campione:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}.$$

La densita' della variabile $T_n = \frac{\sqrt{n}(\bar{X}_n - m)}{\hat{\sigma}}$ quando le osservazioni sono prese da una densita' gaussiana di media m e di varianza qualunque, non dipendono da tale varianza (facile da verificare, svisto che se gli scarti dalla media vengono moltiplicati per una costante, la statistica T_n rimane invariata), ma solo dalla dimensione campionaria n, o meglio dal numero dei gradi di liberta' $m = n-1$. La densita' in questione e' la t di Student con m gradi di liberta'. Innanzi tutto questa e' una densita' pari. Poi, qualunque sia il numero dei gradi di liberta' la funzione di distribuzione per valori positivi e' piu' piccola che per la gaussiana (valori grandi sono piu' probabili a causa della variabilita' del denominatore), quindi la funzione quantile e' piu' grande (per valori maggiori di $\frac{1}{2}$). Di conseguenza le soglie critiche aumentano, a parita' di livello di significativita' (come ci si aspetta dato che stiamo rimpiazzando il valore vero di σ con una stima), mentre fissato lo stesso valore per T_n o per Z_n, a parita di dimensione campionaria, il livello di significativita' osservato e' piu' grande nel primo caso (se la varianza non e' nota i dati sono meno significativi). Le differenze tendono a scomparire quando il numero dei gradi di liberta' cresce, per la convergenza della t di Student alla gaussiana standard al tendere dei gradi di liberta' a $+\infty$. Per popolazioni non gaussiane, dato che abbiamo bisogno di grandi campioni per utilizzare il teorema del limite centrale, si tende comunque ad usare le soglie critiche gaussiane anche in presenza di varianze stimate.

Dalla verifica delle ipotesi agli intervalli di confidenza: E' chiaro che nei problemi di verifica delle ipotesi su di un parametro il valore ipotizzato viene privilegiato rispetto ai valori alternativi, alla luce del fatto che il livello di significativita' e sempre un numero piuttosto piccolo, che ci cautela contro rifiuti ingiustificati dell'ipotesi. Ma, se mettiamo sotto test tutti i possibili valori del parametro, non ne privilegiamo nessuno e ci muoviamo verso problemi di stima parametrica, piuttosto che di verifica delle ipotesi. Riferiamoci di nuovo ad un problema che riguarda osservazioni gaussiane con media ignota e varianza nota, e mettiamo sotto test ciascun valore m contro alternative bilatere, ad un livello di significativita' fissato e uguale a α. Il valore m non e' rifiutato quando

$$m - \frac{\sigma}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2) < \bar{X}_n < m + \frac{\sigma}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2)$$

il che avviene se e solo se

$$\bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2) < m < \bar{X}_n + \frac{\sigma}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2)$$

Essendo la probabilita' che avvenga il primo evento uguale a $1-\alpha$, quando le osservazioni vengono da una gaussiana di media m e varianza σ^2, lo stesso avviene per la probabilita' del secondo. Ora osserviamo che mentre nel primo caso gli estremi dell'intervallo sono fissati e ci chiediamo la probabilita' che una variabile aleatoria prenda valori interni a questo, nel secondo caso sono gli estremi dell'intervallo ad essere variabili aleatorie, mentre il valore m e' fisso. Diciamo quindi che l'intervallo aleatorio $(\bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2), \bar{X}_n - \frac{\sigma}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2))$ e' un **intervallo di confidenza** per la media della popolazione gaussiana, il che significa che qualunque sia il valore vero di questo parametro, esso sara' contenuto nell'intervallo con

probabilita' $1-\alpha$ (cioe' una percentuale $(1-\alpha)100$ per cento delle volte; quando $\alpha = 0.05$, 19 volte su 20). Si noti che il modo in cui l'ampiezza dell'intervallo dipende da σ, n e α e' esattamente in accordo con la nostra intuizione (per la dipendenza da α si noti che al decrescere di α la soglia $\Phi(^{-1}(1-\alpha/2)$ tende ad essere sempre piu' grande, e infatti tende a $+\infty$ per $\alpha \to 0+$). Quando la varianza e' ignota (e il campione piccolo) la varianza σ va rimpiazzata con la stima $\hat{\sigma}$ di cui abbiamo gia' parlato e il quantile della gaussiana standard va rimpiazzato con il corrispondente quantile della t di Student con $n-1$ gradi di liberta' e quindi a parita' di confidenza l'intervallo si allarga. Quando n e' grande questi quantili tendono a quelli della gaussiana standard ed ecco quindi che quando n e' grande anche per popolazioni non gaussiane possiamo utilizzare i test e gli intervalli di cui sopra, con le varianze stimate sul campioni e i quantili della densita' gaussiana standard.

La stima di una probabilita' di successo: Per popolazioni bernoulliane il test (asintotico) dell'ipotesi $p = p_0$ di livello α contro alternative bilatere, tenendo conto che la varianza stimata $\hat{sigma}_n$ differisce da $\hat{X}_n(1-\hat{X}_n)$ per un fattore che tende ad 1 al tendere di $n \to +\infty$ (e che quindi puo' essere ignorato) ha la regione di rifiuto complementare di

$$p_0 - \frac{\bar{X}_n(1-\bar{X}_n)}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2) < \bar{X}_n < p_0 + \frac{\bar{X}_n(1-\bar{X}_n)}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2)$$

e quindi all'intervallo di confidenza per p_0

$$\bar{X}_n - \frac{\bar{X}_n(1-\bar{X}_n)}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2) < p_0 < \bar{X}_n + \frac{\bar{X}_n(1-\bar{X}_n)}{\sqrt{n}}\Phi(^{-1}(1-\alpha/2).$$

con il suo valore massimo che si ottiene per $\bar{X}_n = 1/2$, e che vale $1/4$. Una possibilita' alternativa e' di rimpiazzare la stima della varianza $\bar{X}_n(1-\bar{X}_n)$. In questo modo e' chiaro che l'intervallo si allarga rispetto a quanto occorre per garantire il livello di confidenza desiderato e quindi l'intervallo diventa "conservativo", include cioe' piu' valori di quanto necessario per garantire il livello "nominale" di confidenza.

Dall'intervallo di confidenza al test, il concetto di significativita' pratica: E' evidente che una volta disponibile (come sovente nei lavori scientifici) un intervallo di confidenza al livello di confidenza $1-\alpha$, la verifica di un'ipotesi su di uno specifico valore per la media, contro alternative bilatere, con livello di significativita' α consiste nel verificare se tale valore e' compreso nell'intervallo o meno: nel primo caso non si arriva a rifiutare l'ipotesi, cosa che fa nel secondo caso. Tuttavia puo' darsi che, anche se il valore di α e' molto piccolo, l'intervallo non contenga esattamente il valore ipotizzato ma ne contenga altri molto vicini. In questo caso i dati anche se significativi contro l'ipotesi, potrebbero non esserlo nei confronti di ipotesi indistinguibili, ai fini pratici, da questa ipotesi. Nei problemi scientifici i parametri di interesse rappresentano gli effetti di certe variabili su altre, e quindi il valore 0 per parametri di questo tipo assume un interesse particolare. Ora puo' succedere che, sebbene il valore 0 venga rifiutato con un livello di significativita' molto piccolo, valori piccoli del parametro (che indicano un effetto molto limitato di una variabie sull'altra) non possano essere rifiutati, anche incrementando in modo consistente il livello di significativita' del test (e quindi riducendo l'ampiezza dell'intervallo). I dati sperimentali quindi suggeriscono che una variabile influenzi l'altra ma non sono in grado di escludere che questa influenza possa essere molto bassa.