STATISTICA

Come già discusso, si indica con statistica l'insieme delle attività, ricerche e risultati finalizzati all'analisi dei risultati e dei dati ottenuti da esperimenti che in qualche modo si possono definire casuali. La statistica è generalmente suddivisa in parte descrittiva, che si occupa dell'elaborazione dei dati, e parte inferenziale, che invece assume che i dati siano il risultato di un qualche meccanismo descrivibile in termini probabilistici, e quindi utilizza tutto quanto abbiamo visto finora per l'analisi dei dati stessi.

CAPITOLO 7

STATISTICA DESCRITTIVA

La statistica descrittiva si occupa del rapporto tra gli avvenimenti con risultati altamente non predicibili del mondo reale e la loro rappresentazione in termini scientificamente utilizzabili. Ad esempio, quando si si fanno osservazioni o si registrano i risultati di esperimenti la rappresentazione delle informazioni ottenute in termini trasmissibili ed analizzabili richiede il prendere molte decisioni spesso altamente opinabili. Per esempio, se vi trovate ad analizzare i dati relativi a migliaia di pazienti per studiare certe patologie o certe terapie, spesso questi sono stati registrati sinteticamente su fogli vari da medici diversi e vanno letti (e qui già ci vorrebbe l'aiuto di un medico), interpretati, trasferiti in forma unificata indovinando possibili differenze tra i diversi redattori e spesso rendendo numeriche varie valutazioni qualitative. Poi i dati vanno accuratamente trascritti in un data base, a quel punto analizzati per cercare delle indicazioni significative; una volta che ci siano delle osservazioni apparentemente sensate queste vanno giustificate e successivamente presentate.

E' chiaro che tutto questo è un processo lungo e delicato e che vi sono innumerevoli articoli e trattati che lo studiano. Noi qui ci occuperemo brevemente solo di un aspetto particolare: supponendo di avere dati già in forma numerica ci poniamo il problema di generare un grafico o alcuni valori rappresentativi.

Supponiamo quindi di avere un *campione* $x_1, \dots, x_n \in \mathbb{R}^n$. Per prima cosa poniamo

$$x_{min} = \min\{x_i, i = 1, \dots, n\}$$

e

$$x_{max} = \max\{x_i, i = 1, \dots, n\}$$

così che $[x_{min}, x_{max}]$ è il range del campione.

7.1. L'istogramma

L'*istrogramma* dei dati, detto semplicemente istogramma, è una rappresentazione grafica dei dati che rappresenta frequenze attraverso le aree. Descriveremo ora una modalità di realizzazione degli istogrammi, che come si vedrà prevede comunque numerose scelte arbitrarie.

Dividiamo $[\lfloor x_{min} \rfloor, \lfloor x_{max} \rfloor + 1]$ in intervallini scegliendo

$$\lfloor x_{min} \rfloor = d_0 < d_1 < \cdots < d_m = \lfloor x_{max} \rfloor + 1$$

e formando intervallini semiaperti a destra (scegliendoli semiaperti a destra si semplificano leggermente le definizioni):

$$[d_k, d_{k+1}[$$

per $k = 0, \dots, m-1$. L'istogramma è la funzione che per $x \in [d_k, d_{k+1}[$ vale

$$I(x) = \frac{|\{i : x_i \in [d_k, d_{k+1}[\}|}{n(d_{k+1} - d_k)}.$$

ESERCIZIO 25. *Con dati* $0, 1, 2, 2, 1, 1, 0, 2, 1$ *si disegnino gli istogrammi ottenuti con la scelta* $d_0 = 0, d_1 = 1, d_2 = 3$ *e con la scelta* $d_0 = 0, d_1 = 1, d_2 = 2, d_3 = 3$.

OSSERVAZIONE 9. *Attraverso l'istogramma possiamo capire e mostrare dove i dati si concentrano di più di meno, anche se l'arbitrarietà della scelta degli intervalli e del comportamento agli estremi degli intervalli lascia spazio a molti fraintendimenti e possibili rappresentazioni falsate.*

7.2. Elementi di Statistica Descrittiva

Per descrivere sinteticamente i dati si utilizzano:
(I) **valor medio empirico** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;
(II) **varianza empirica** $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$;
(III) **SD empirica** $SD = \sigma_n = \sqrt{\sigma_n^2}$;
(IV) **momento empirico k-simo** $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$;
per motivi che non discuteremo qui si utilizzano spesso σ_{n-1}^2 e σ_{n-1}, definiti come i precedenti con $n-1$ al posto di n.

CAPITOLO 8

STATISTICA INFERENZIALE

Immaginiamo ora che i dati siano realizzazioni di un qualche meccanismo casuale descrivibile nell'ambito del calcolo delle probabilità. Consideriamo un caso semplice, in cui i dati del campione che stiamo considerando siano realizzazioni di variabili aleatorie indipendenti ed identicamente distribuite, discrete o continue a seconda che i dati osservati siano numeri interi o meno.

Suggerimenti sulla distribuzione delle variabili aleatorie che si prendono a modello possono venire da molte direzioni tipo teorie specifiche ed esperti o esperienze nel settore, oppure la forma della distribuzione può essere suggerita dall'istogramma dei dati stessi.

ESEMPIO 59. *Possiamo immaginare che i dati dell'esercizio 25 di cui sono stati richiesti gli istogrammi nel paragrafo precedente siano realizzazioni di variabili i.i.d. $B(2,p)$, ossia ognuno dei dati corrisponda al numero di successi in 2 prove indipendenti ognuna con probabilità di successo p; oppure che siano $B(n,p)$ o altro.*

8.1. Stima di parametri: il metodo dei momenti

Una volta identificata la distribuzione delle variabili aleatorie spesso rimangono dei parametri da determinare, come ad esempio la probabilità di successo p nell'ultimo esempio o uno o più parametri di una distribuzione continua. A questo scopo sono stati sviluppati vari metodi, e noi ne presenteremo uno chiamato metodo dei momenti.

In generale consideriamo che la distribuzione dipenda da l parametri $\theta_1, \ldots, \theta_l$. Vogliamo determinare dei valori

$$(\hat{\theta}_1, \ldots, \hat{\theta}_l) = (\hat{\theta}_1(x_1, \ldots, x_n), \ldots, \hat{\theta}_l(x_1, \ldots, x_n))$$

che siano stime ragionevoli dei valori veri dei parametri.

Si considerano allora i momenti k-simi della distribuzione $E(X^k) = E_{\theta_1,\ldots,\theta_l}(X^k)$ e si uguagliano ai momenti k-simi empirici corrispondenti, formando un numero di equazioni sufficiente a permettere di determinare la stima dei parametri ponendo $(\hat{\theta}_1, \ldots, \hat{\theta}_l)$ uguale alle soluzioni del sistema composto di tali equazioni $m_k = E(X^k)$.

ESEMPIO 60. *Con i dati dell'esercizio 25 si può stimare la probabilità p della binomiale con il metodo dei momenti. Ora se $X \sim B(2,p)$ allora $E(X) = 2p$ ed $m_1 = \bar{x} = 10/9$, per cui $\hat{p} = 10/18$.*

ESEMPIO 61. *Se invece interpretiamo i dati* $-2, 3, 1$ *come realizzazioni di variabili aleatorie i.i.d. continue* $X_i \sim U([-a, a])$ *allora* $f_{X_i}(x) = \frac{1}{2a}\mathbb{I}_{[-a,a]}(x)$. *Il tentativo di utilizzare il metodo dei momenti con il primo momento perønon funziona, perchè* $E(X_i) = 0$ *per simmetria. Tuttavia in questo caso*

$$E(X_i^2) = \frac{1}{2a}\int_{-a}^{a} x^2 dx = \frac{a^2}{3}$$

per cui, calcolando m_2 *dai dati si ha* $\hat{a} = \sqrt{14} \approx 3.74$. *Se* $\hat{a}$ *fosse risultato minore di* x_{max} *ne avremmo concluso che o il metodo di stima dei parametri non era adeguato o le assunzioni fatte non costituivano un buon modello per i dati.*

8.2. Test di ipotesi: lo z-test

Valutare se le ipotesi fatte siano un buon modello per i dati è naturalmente una questione assai complessa. Qui discuteremo uno schema che in alcuni casi offre qualche elemento di valutazione. Filosoficamente, si tratta della possibilità di evidenziare o meno dai dati elementi che contraddicano statisticamente il complesso delle ipotesi assunte. Nessuna conclusione verrà tratta in positivo: o le ipotesi sono contraddette o non lo sono (almeno sulla base dell'analisi fatta di questi dati).

Lo schema che presentiamo è detto test di ipotesi e si realizza come segue:

(0) Si fissa un livello del test, detto α, tipo $\alpha = 0.05$ oppure $\alpha = 0.01$.

(1) Si descrive l'ipotesi che si vuole testare, detta Ipotesi Nulla H_0. Tipicamente in questo si dovrebbe descrivere l'insieme di tutti i passi che costituiscono l'esperimento. Di solito si indica anche in dettaglio una ipotesi alternativa, chiamata H_1 che si testa in contrasto con H_0, ma per semplicità assumiamo che H_1 sia semplicemente la negazione di H_0.

(2) Si determina il valore di una certa funzione dei dati e dei parametri ottenuti da H_0 che si possa considerare come la realizzazione di una variabile aleatoria di cui sia nota, almeno approssimativamente, la distribuzione. Nel test che discutiamo qui questo valore sarà indicato con z.

(3) Si stima la probabilità che la variabile aleatoria Z di cui z è una realizzazione devii da quanto atteso per più di quanto osservato. Tale probabilità è detta valore p del test.

(4) Si confrontano p e α. Se $p < \alpha$ si rigetta H_0 ed altrimenti non si rigetta.

Come esempio consideriamo lo z-test in cui si valuta l'ipotesi attraverso il confronto tra la differenza della media campionaria da quella attesa con la deviazione standard per dati di variabili aleatorie normali. La restrizione alle variabili aleatorie normali, tuttavia, si può aggirare

utilizzando il test in forma approssimata sulla base del Teorema Centrale del Limite. Da H_0 ricaviamo il valore atteso $E(X)$ e la deviazione standard $SD(X)$ di X ed il CLT ci dice che

$$Z = \frac{\sum_{i=1}^{n} X_i - nE(X)}{\sqrt{nSD(X)}}$$

ha una distribuzione approssimativamente $N(0,1)$ quando le X_i siano i.i.d. tali che $E(X_i^2) < \infty$. Quindi conosciamo approssimativamente la distribuzione di Z e nel test valutiamo quindi la sua realizzazione

$$z = \frac{\sum_{i=1}^{n} x_i - nE(X)}{\sqrt{nSD(X)}}.$$

ESEMPIO 62. *Ritornando all'esempio 1, se esce testa 459 volte su 1000 lanci di una moneta, allora si esegue un test di ipotesi relativamente alla correttezza della moneta ponendo:*

(0) $\alpha = 0.02$

(1) H_0*= la moneta è equilibrata e ben mescolata ad ogni lancio; i risultati sono correttamente registrati.*

(2) Da H_0 *i risultati sono relativi a prove i.i.d.* $B(1000, 1/2)$*. Per cui* $E(\sum_{i=1}^{1000} X_i) = 1/2$ *e* $SD(\sum_{i=1}^{1000} X_i) = \sqrt{1000}/2$ *e*

$$z = \frac{459 - 500}{\sqrt{1000}/2} \approx -2.59$$

(3) Il valore p si determina dall'approssimazione normale come

$$\int_{-\infty}^{-2.59} \frac{\sqrt{2\pi}}{e^{-\frac{x^2}{2}}} \approx 0.0048$$

(4) Poichè $0.0048 = p < \alpha = 0.02$ *si rigetta* H_0*, ossia si conclude che l'insieme delle ipotesi fatte non è consistente con i dati osservati.*

OSSERVAZIONE 10. *Avevamo già visto che questo numero di teste non era accettabile; la terminlogia del test di ipotesi ha permesso di chiarire bene la procedura e le conclusioni.*

Ora possiamo concludere l'analisi dell'esempio 2 attraverso uno z-test (approssimato).

ESEMPIO 63. *(0)* $\alpha = 0.02$

(1) Un modello adeguato è assumere che il tempo di funzionamento di ogni pezzo i *sia una variabile aleatoria* X_i *indipendente con distribuzione* $exp(\lambda)$*. Quindi poniamo*

$H_0 =$ *I tempi di vita sono realizzazioni di variabili aleatorie i.i.d. esponenziali di parametro* λ *tale che il tempo medio di funzionamento dichiarato corrisponde al valor medio della distribuzione, ossia* $\frac{1}{\lambda} = 1000$*, quindi* $\lambda = \frac{1}{1000}$.

(2) Da H_0 si ottiene che, per ogni i, $E(X_i) = 1000$ e $SD(X_i) = 1000$ e si può applicare lo z-test poichè $(EX_i^2) < \infty$. Si ha

$$z = \frac{90200 - 100 \cdot 1000}{\sqrt{100} \cdot 1000} \approx -0.98$$

(3) Dall'approssimazione normale

$$p \sim \int_{-\infty}^{-0,98} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \sim 16,35\%. \tag{8.28}$$

(4) Poichè $p \geq \alpha$ ne concludiamo che l'ipotesi nulla non è invalidata dai dati.

Quindi un tempo totale di funzionamento di 90000 ore era perfettamente in linea con la durata media prevista.

ESERCIZIO 26. *Verificare che con una soglia dell'1% si dovrebbe considerare falsa la dichiarazione sulla durata media se il tempo totale di funzionamento fosse minore di 76700 ore.*

ESERCIZIO 27. *Perchè questa volta si tende a considerare inappropriato un risultato ancora perfettamente adeguato?*