Statistica inferenziale

In un problema di statistica abbiamo un esperimento aleatorio al quale sappiamo associare uno spazio probabilizzabile, ma non sappiamo scegliere su di esso una misura di probabilità $P$: riusciamo solo a stabilire che $P$ deve far parte di una certa famiglia $\mathcal{P}$, più o meno grande, di misure di probabilità.
Il compito dello statistico è quello di raccogliere informazioni sull'esperimento aleatorio, favorendo la scelta della misura di probabilità più adeguata per descrivere l'esperimento.

<u>Definizione</u> Sia $(\Omega, \mathcal{A})$ uno spazio probabilizzabile e sia $\mathcal{P}$ una famiglia di misure di probabilità su $(\Omega, \mathcal{A})$. La terna $(\Omega, \mathcal{A}, \mathcal{P})$ è detta <u>modello statistico</u>. Esso si dice <u>parametrico</u> se si ha $\mathcal{P} = \{P^\theta, \theta \in D\}$ con $D \subseteq \mathbb{R}^k$, <u>non parametrico</u> altrimenti.

Noi ci limiteremo all'analisi di modelli statistici parametrici.
In definitiva, compito preliminare per lo studio di un problema statistico è quello di associare all'esperimento aleatorio un modello statistico.

<u>Esempi</u> (1) (controllo di qualità). Una popolazione è composta da individui di tipo A e di tipo B. Non conoscendo il rapporto effettivo tra il n° di individui di tipo A e l'intera popolazione, si sceglie un campione di N individui: le osservazioni di questo

esperimento sono allora (come abbiamo già visto) rappresentate da v.a. $X_1, \dots, X_N$, a valori in $\{0,1\}$, definite su un opportuno spazio probabilizzabile $(\Omega, \mathcal{A})$: ovviamente, $X_i = 1$ se l'i-esimo individuo del campione è di tipo A, $X_i = 0$ altrimenti. Non siamo in grado però di scegliere una misura di probabilità $P$ su $(\Omega, \mathcal{A})$, pur sapendo che le v.a. $X_i$ sono tutte indipendenti e bernoulliane con un certo parametro $\theta \in [0,1]$. Dunque avremo un modello statistico $(\Omega, \mathcal{A}, \mathbb{P})$ con $\mathbb{P} = \{P^\theta, \theta \in [0,1]\}$, ove, per ogni valore del parametro sconosciuto $\theta$, le v.a. $X_i$ hanno legge di Bernoulli $\mathcal{B}(\theta)$ secondo la misura $P^\theta$.

(2) (misura di una grandezza fisica). Per effettuare una misura di una certa grandezza con un determinato strumento, si esegue un certo numero $N$ di misurazioni, il cui risultato, secondo il teorema limite centrale, sarà descritto da v.a. $X_1, \dots, X_N$ indipendenti e gaussiane, pur non essendo note a priori né la speranza $\mu$, né la varianza $\sigma^2$ comuni a tutte le $X_i$. Si potrà quindi scegliere, su un opportuno spazio probabilizzabile $(\Omega, \mathcal{A})$, una famiglia di misure di probabilità $\{P^{\mu,\sigma} : (\mu,\sigma) \in \mathbb{R} \times \mathbb{R}^+\}$, in modo tale che, per ogni $(\mu,\sigma)$, le $X_j$ abbiano legge normale $\mathcal{N}(\mu, \sigma^2)$ secondo la misura di probabilità $P^{\mu,\sigma}$.

**Definizione** Dato un modello statistico $(\Omega, \mathcal{A}, P)$, una v.a. $X$, definita su $(\Omega, \mathcal{A})$, è detta una statistica.

**Definizione** Sia $(\Omega, \mathcal{A}, \{P^\theta : \theta \in D\})$ un modello statistico parametrico. Una successione $\{X_n\}$ di statistiche si dice <u>indipendente</u> se le $X_n$ sono indipendenti rispetto a ciascuna delle probabilità $P^\theta$. Una statistica $X$ è <u>integrabile</u> se $X$ è integrabile rispetto a ciascuna delle probabilità $P^\theta$. Scriveremo in tal caso

$$\mathbb{E}^\theta[X], \quad Var^\theta[X]$$

per indicare la speranza e la varianza di $X$ rispetto alla probabilità $P^\theta$.

**Definizione** Sia $(\Omega, \mathcal{A}, \{P^\theta : \theta \in D\})$ un modello statistico parametrico, e sia $\{L(\theta), \theta \in D\}$ una famiglia di leggi di probabilità. Una sequenza finita di statistiche $X_1, \dots, X_N$ definite su $(\Omega, \mathcal{A}, \{P^\theta : \theta \in D\})$ si dice <u>campione statistico di taglia N</u>, estratto da una popolazione di legge $L(\theta)$, se le statistiche sono indipendenti e tutte dotate di legge $L(\theta)$ secondo la probabilità $P^\theta$.

Osserviamo che, fissata una famiglia di leggi di probabilità $\{L(\theta), \theta \in D\}$, si può <u>sempre</u> costruire un modello statistico e, su di esso, un campione statistico di taglia N estratto da una popolazione di legge $L(\theta)$. Esso si chiama modello statistico campionario di taglia N.

Nel seguito ci occuperemo di 3 problemi generali, che illustrano i principali metodi di inferenza della statistica.

1. Il problema della stima puntuale;
2. il problema della stima insiemistica;
3. il problema dei test d'ipotesi.

Il problema 1 consiste nella scelta di uno stimatore, cioè di una applicazione $T: \Omega \to D$, che rappresenta la strategia seguente: ci si impegna, qualunque sia la realizzazione $\omega$ dell'esperimento, ad attribuire al parametro sconosciuto $\theta$ il valore $T(\omega)$. Il problema della stima consisterà dunque nello scegliere lo stimatore T in modo da minimizzare l'errore, ossia minimizzando certe quantità legate a T, espresse per mezzo della probabilità $P\theta$.

Il problema 2 consiste nella scelta di un'applicazione $S: \Omega \to \mathcal{P}(D)$, che rappresenta la strategia seguente: ci si impegna, qualunque sia la realizzazione $\omega$ dell'esperimento, a stimare il valore di $\theta$ come appartenente all'insieme $S(\theta)$ (insieme di fiducia).

Il problema 3 consiste nella scelta di un <u>test</u>, che permetta di verificare o di confutare una fissata <u>ipotesi</u> sul parametro sconosciuto $\vartheta$, che stabilisca che esso appartiene ad un dato sottoinsieme $D_0 \subset D$. Il test consiste nella costruzione di una <u>partizione</u> $\{B, B^c\}$ di $\Omega$, e rappresenta la strategia seguente: ci si impegna, qualunque sia la realizzazione $\omega$ dell'esperimento, a rifiutare l'ipotesi se $\omega \in B$, e ad accettarla se $\omega \in B^c$.

In definitiva, lo statistico agisce seguendo una <u>regola di decisione</u> che tende a minimizzare certe conseguenze le quali, ragionevolmente, appaiono nocive; ma la regola di decisione deve essere <u>a priori</u>, vale a dire che ci si impegna a seguirla prima di compiere l'esperimento e qualunque sia il suo risultato $\omega$. Dunque l'azione dello statistico è determinata da $\omega$, ma attraverso una regola definita a priori.

## Teoria della stima

Introduciamo la nozione di stimatore. Sia $(\Omega, \mathcal{A}, \{P^\theta\}_{\theta \in D})$ un modello statistico e sia assegnato su di esso un campione statistico $X_1, \dots, X_N$ di taglia $N$. Sia poi $\psi: D \to I$ una funzione del parametro $\theta$, a valori nell'insieme $I \subseteq \mathbb{R}$.

**Definizione** Uno stimatore della funzione $\psi(\theta)$ è una statistica $T$ della forma

$$T = t(X_1, \dots, X_N)$$

con $t$ funzione da $\mathbb{R}^N$ in $I$.

Intuitivamente, assegnare uno stimatore $T = t(X_1, \dots, X_N)$ di $\psi(\theta)$ significa fissare la seguente regola: se i dati raccolti dall'osservazione di un risultato $\omega$ sono

$$(x_1, \dots, x_N) = (X_1(\omega), \dots, X_N(\omega)),$$

si stimerà la quantità sconosciuta $\psi(\theta)$ con il numero $t(x_1, \dots, x_N)$, detto appunto la stima di $\psi(\theta)$.

Osservazione 1: lo stimatore $T$ dipende dalla taglia $N$ del campione: perciò, di solito, si costruisce una successione $\{T_n\}_{n \in \mathbb{N}}$ dove, per ogni $n$, la statistica $T_n$ è uno stimatore di $\psi(\theta)$

tramite un campione di taglia $n$. 

<u>Osservazione 2</u> Uno stimatore, essendo una v.a., non assumerà mai il valore $\psi(\theta)$, ma naturalmente si spera che prenda valori non troppo distanti da esso. Poichè qualunque funzione delle osservazioni è uno stimatore, sarà necessario fissare qualche criterio per stabilire quali di esse sono "buoni" stimatori e quali no. Noi ci limiteremo a studiare buoni stimatori solo in casi molto semplici, legati alla media e alla varianza della legge del campione.

<u>Definizione</u> La statistica $T$ è uno <u>stimatore corretto</u> (o <u>non distorto</u>) del parametro $\psi(\theta)$ se risulta

$$E^{\theta}[T] = \psi(\theta) \qquad \forall \theta \in D.$$

In caso contrario, $T$ si dirà <u>stimatore distorto</u>.

Se lo stimatore si discosta da $\psi(\theta)$, la sostituzione del valore $\psi(\theta)$ con $T$ comporta un errore, e quindi un <u>costo</u>, che si esprime mediante una funzione positiva $C(\theta,a)$, che misura la perdita proveniente dal sostituire $\psi(\theta)$ con il valore $a$. Dunque, se $T(\omega)$ si discosta da $\psi(\theta)$, il costo sarà la v.a.

$$\omega \mapsto C(\theta, T(\omega)).$$

<u>Definizione</u> Il <u>rischio</u> dello stimatore $T$ è il suo costo medio, ossia la funzione

$$R_T(\theta) = E^\theta\left[C(\theta,T)\right], \quad \theta\in D.$$

Generalmente, si sceglie come costo la funzione quadratica

$$C(\theta,a) = |\psi(\theta)-a|^2,$$

detta appunto <u>costo quadratico</u>. Con questa scelta, il rischio di $T$ è la funzione

$$R_T(\theta) = E^\theta\left[(\psi(\theta)-T)^2\right],$$

detta <u>rischio quadratico</u>.

Si noti che se $T$ è uno stimatore corretto, allora

$$R_T(\theta) = E^\theta\left[(\psi(\theta)-T)^2\right] = \psi(\theta)^2 - 2\psi(\theta)E^\theta[T] + E^\theta[T^2] =$$
$$= E^\theta[T^2] - E[\theta]^2 = \mathrm{Var}^\theta[T].$$

Diciamo infine, se $S, T$ sono stimatori di $\psi(\theta)$, che $T$ è <u>preferibile</u> a $S$ se risulta $R_T(\theta) \leq R_S(\theta)$; se $\mathcal{T}$ è una famiglia di stimatori di $\psi(\theta)$, diremo che $T\in\mathcal{T}$ è uno stimatore <u>ottimale</u> (rispetto a $\mathcal{T}$) se $T$ è preferibile a $S$ per ogni $S\in\mathcal{T}$.

Passiamo ora dal generale al particolare, supponendo di avere un campione statistico $X_1, \dots, X_N$ di taglia $N$ su un modello statistico $(\Omega, \mathcal{A}, \mathbb{P})$. Supponiamo che il campione sia costituito da v.a. gaussiane con media $\mu$ e varianza $\sigma^2$, entrambe finite ma sconosciute. Dunque

$$\mathbb{P} = \{ P^{\mu,\sigma} \}_{(\mu,\sigma) \in \mathbb{R} \times \mathbb{R}^+} .$$

<u>Definizione</u> La <u>media empirica</u>, o <u>media campionaria</u>, è lo stimatore $\bar{X}$ di $\mu$ così definito:

$$\bar{X} = \frac{X_1 + \dots + X_N}{N} .$$

(dunque $\bar{X} = t(X_1 \dots X_N)$ con $t(\underline{x}) = \frac{1}{N} \sum_{i=1}^{N} x_i$).

Questo stimatore è corretto: infatti, per ogni $\theta = (\mu, \sigma)$ si ha

$$E^\theta[\bar{X}] = E^{\mu,\sigma}\left[\frac{X_1 + \dots + X_N}{N}\right] = \frac{E^{\mu,\sigma}[X_1] + \dots + E^{\mu,\sigma}[X_N]}{N} =$$

$$= \frac{N\mu}{N} = \mu .$$

Inoltre, si osservi che (essendo in particolare le $X_j$ indipendenti)

$$Var^\theta[\bar{X}] = \frac{Var^\theta[X_1] + \dots + Var^\theta[X_N]}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} .$$

Dunque la media empirica è uno stimatore corretto della media

del campione, ma la sua varianza è ridotta, rispetto a quella del campione, di un fattore N.
Inoltre, per il teorema limite centrale, la statistica $Z=\sqrt{N}\,\frac{\bar{X}-\mu}{\sigma}$ ha approssimativamente legge $\mathcal{N}(0,1)$ rispetto a ciascuna $P_{\mu,\sigma}$.
Dunque la legge di $\bar{X}$ è approssimativamente $\mathcal{N}(\mu,\frac{\sigma^2}{N})$:

$$P(\bar{X}\le c)= P\left(Z\le \frac{\sqrt{N}}{\sigma}(c-\mu)\right) = \Phi\left(\frac{\sqrt{N}}{\sigma}(c-\mu)\right),\quad c\in\mathbb{R}.$$

<u>Definizione</u> La <u>varianza empirica</u>, o <u>varianza campionaria</u>, è lo stimatore $S^2$ di $\sigma^2$ così definito:

$$S^2=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2,$$

mentre la statistica $S=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2}$ è detta invece <u>deviazione standard empirica</u> (o <u>campionaria</u>).

Perché dividere per $N-1$ e non per $N$? Perché, così facendo, $S^2$ è uno stimatore corretto (e altrimenti no). Ciò si vede, un po' faticosamente, così: partiamo dalla relazione

$$N\bar{X}=\sum_{i=1}^{N}X_i.$$

Allora

$$\sum_{i=1}^{N}(X_i-\bar{X})^2=\sum_{i=1}^{N}X_i^2-2\bar{X}\sum_{i=1}^{N}X_i+N\bar{X}^2=\sum_{i=1}^{N}X_i^2-N\bar{X}^2.$$

Ne segue

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\overline{X})^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right),$$

da cui

$$(N-1)S^2 = \sum_{i=1}^{N} X_i^2 - N\overline{X}^2.$$

Prendiamo la speranza di entrambi i membri e ricordiamo che $\mathbb{E}^{\mu,\sigma}[Y^2] = \mathrm{Var}^{\mu,\sigma}[Y] + \mathbb{E}^{\mu,\sigma}[Y]^2$ per ogni v.a. integrabile $Y$. Quindi

$$(N-1)\,\mathbb{E}^{\mu,\sigma}[S^2] = \mathbb{E}^{\mu,\sigma}\left[\sum_{i=1}^{N} X_i^2\right] - N\,\mathbb{E}^{\mu,\sigma}[\overline{X}^2] =$$

$$= N\,\mathbb{E}^{\mu,\sigma}[X_1^2] - N\,\mathbb{E}^{\mu,\sigma}[\overline{X}^2] =$$

$$= N\,\mathrm{Var}^{\mu,\sigma}[X_1] + N\,\mathbb{E}^{\mu,\sigma}[X_1]^2 -$$

$$- N\cdot\mathrm{Var}^{\mu,\sigma}[\overline{X}] - N\,\mathbb{E}^{\mu,\sigma}[\overline{X}]^2 =$$

$$= N\sigma^2 + N\mu^2 - N\frac{\sigma^2}{N} - N\mu^2 = (N-1)\sigma^2,$$

da cui, finalmente, $\mathbb{E}^{\mu,\sigma}[S^2] = \sigma^2$.

Il calcolo di $\mathrm{Var}^{\mu,\sigma}[S^2]$ è assai intricato; lo vedremo dopo.

Qual è la legge della varianza empirica? C'è un teorema a questo proposito.

<u>Teorema (di Cochran)</u> Se $\overline{X}$ e $S^2$ sono la media empirica e la varianza empirica di un campione statistico di taglia $N$ e legge

normale $\mathcal{N}(\mu, \sigma^2)$, allora:

(i) $\overline{X}$ e $S^2$ sono statistiche indipendenti;

(ii) $\overline{X}$ ha legge normale $\mathcal{N}(\mu, \frac{\sigma^2}{N})$;

(iii) $W = \frac{N-1}{\sigma^2} S^2$ ha legge $\chi^2(N-1)$,

(iv) $T = \frac{\overline{X}-\mu}{S}\sqrt{N}$ ha legge $t(N-1)$.

<u>dim.</u> omessa. □

Dunque possiamo fare stime probabilistiche sulla speranza e sulla varianza di campioni statistici con legge gaussiana.
Ma c'è di più: dal teorema di Cochran si sa che $W$ ha legge $\chi^2(N-1)$, e dunque $Var[W] = 2(N-1)$; come abbiamo verificato qualche lezione fa. Dunque

$$Var[S^2] = Var\left[\frac{\sigma^2}{N-1} W\right] = \frac{\sigma^4}{(N-1)^2} Var[W] = \frac{2\sigma^4}{N-1},$$

e dunque, come accade per la media empirica, la varianza empirica ha varianza che tende a 0 per $N \to \infty$, ossia la stima di $\sigma^2$ mediante $S^2$ è sempre più precisa all'aumentare della taglia del campione.

<u>Esempio</u> Per misurare una grandezza fisica si eseguono $N$ misurazioni indipendenti ottenendo $N$ risultati $X_1, \dots, X_N$. Poiché i dati ottenuti formano un campione di taglia $N$, per $N$ grande la media e la varianza

empiriche sono stimatori corretti abbastanza precisi della media del campione e della varianza del campione. Quindi una stima della grandezza misurata è semplicemente

$$\bar{x} = \frac{1}{N}(x_1 + \dots + x_N)$$

e una stima della varianza del campione è

$$s^2 = \frac{1}{N-1}\left[(x_1 - \bar{x})^2 + \dots + (x_N - \bar{x})^2\right].$$

<u>Esempio</u> Il tempo di vita di un tipo di lampadine è, in media, 500 ore (h), con deviazione standard 80 h. Comprate 16 lampadine, e assumendo che questo campione abbia legge normale, qual è la probabilità che la media empirica sia maggiore di 525 h?

Siano $X_j$, $j = 1 \dots 16$, v.a. che rappresentano la durata della $j$-sima lampadina. Le $X_j$ hanno legge $N(500, 80^2)$. Dunque $\bar{X}$ ha legge $N\left(500, \frac{80^2}{16}\right) = N(500, 400)$. Perciò

$$P(\bar{X} > 525) = P\left(\frac{\bar{X} - 500}{20} > \frac{25}{20}\right) = 1 - \Phi(1.25) \simeq 0.11.$$

<u>Esempio</u> Il tempo impiegato da un microprocessore per svolgere certi processi è descritto da una v.a. normale con media 30 ns (nanosecondi) e deviazione standard di 3 ns.

Se si osserva l'esecuzione di 16 processi, qual è la probabilità che la varianza empirica $S^2$ sia maggiore di 15 ns?

Se $X_j$, $1 \leq j \leq 16$ è la v.a. che rappresenta la durata del j-simo processo, $X_j$ ha legge $N(30,9)$. Per il teorema di Cochran, $S^2$ ha legge $\chi^2(15)$. Si ha

$$P(S^2 > 15) = P\left(\frac{n-1}{\sigma^2} S^2 > \frac{15 \cdot 15}{9}\right) = P\left(\frac{n-1}{\sigma^2} S^2 > 25\right) =$$

$$= 1 - P\left(\frac{n-1}{\sigma^2} S^2 \leq 25\right).$$

Dalla tavola della legge $\chi^2(15)$, il valore 25 corrisponde circa al quantile $\chi^2_{0.95}(15)$. Perciò

$$P(S^2 > 15) = 1 - 0.95 = 0.05.$$