UN'APPLICAZIONE DEL TEOREMA LIMITE CENTRALE

Supponiamo di lanciare una moneta e di ottenere 45 teste in 100 lanci: è un risultato plausibile, oppure devo ritenere che la moneta non sia equilibrata? E se ottenessi 450 teste in 1000 lanci? O 4500 teste in 10000 lanci?

La risposta nei tre casi è molto diversa: se la moneta fosse davvero equilibrata, ottenere 45 teste in 100 lanci sarebbe più che plausibile, 450 teste in 1000 lanci sarebbe estremamente improbabile, mentre 4500 teste in 10000 lanci sarebbe quasi impossibile. Vediamo di capire perché.

Costruiamo un modello per il nostro fenomeno aleatorio, introducendo una successione $\{X_i\}_{i\in\mathbb{N}}$ di variabili aleatorie indipendenti e identicamente distribuite, con la seguente legge:

$$P(X_i = 0) \;=\; P(X_i = 1) \;=\; \frac{1}{2}\,.$$

Non è importante dire precisamente quale sia lo spazio di probabilità $(\Omega, \mathcal{A}, P)$ su cui le variabili sono definite: uno qualunque va bene ai nostri scopi (si può ad esempio prendere l'intervallo $[0, 1]$ con la misura di Lebesgue, e costruire le X_i come opportune funzioni).

L'interpretazione è la seguente: l'evento $\{X_i = 1\}$ significa che all'i-esimo lancio esce testa, mentre l'evento $\{X_i = 0\}$ significa che all'i-esimo lancio esce croce. Allora il numero totale di teste ottenute nei primi n lanci è dato dalla variabile aleatoria

$$S_n \;:=\; X_1 + X_2 + \ldots + X_n\,.$$

Vogliamo capire se sia ragionevole ottenere 45 teste in 100 lanci. Uno potrebbe essere tentato di calcolare la probabilità di tale evento, cioè $P(S_{100} = 45)$, ma questa non è una buona idea. In effetti, quando il numero di lanci è grande, la probabilità di ottenere un *qualunque numero fissato di teste* risulta piccola: ciò è semplicemente dovuto al fatto che ci sono molti esiti possibili.

Un'idea più furba per capire se 45 teste in 100 lanci sia un risultato ragionevole è di calcolare la probabilità di ottenere esiti "peggiori", cioè di ottenere *45 teste o meno*. Vogliamo dunque calcolare la probabilità

$$P\big(\{S_{100} \le 45\}\big) \;=\; P\big(S_{100} - 50 \le -5\big)\,.$$

Se tale probabilità risulta piccola, concluderemo che l'evento osservato è improbabile. Analogamente calcoleremo

$$P\big(S_{1000} - 500 \le -50\big) \qquad \text{e} \qquad P\big(S_{10000} - 5000 \le -500\big)\,,$$

e più in generale

$$P\big(S_n - n/2 \le -k\big)\,.$$

Date: 15 novembre 2007.

Si noti che la variabile S_n ha una distribuzione nota (Binomiale): $S_n \sim B(n, \frac{1}{2})$, per cui esiste una formula esplicita per la probabilità che cerchiamo:

$$P\big(S_n - n/2 \le -k\big) \;=\; \sum_{\ell=0}^{\lfloor n/2 \rfloor - k} \binom{n}{\ell} \cdot \frac{1}{2^n}\,. \tag{1}$$

Benché non esista un'espressione chiusa per tale somma, è certamente possibile calcolarne il valore usando un computer. Vogliamo però mostrare come stimare il valore di tale probabilità usando il Teorema Limite Centrale. Specializzato al nostro contesto, tale teorema afferma che per ogni $x \in \mathbb{R}$

$$\lim_{n\to\infty} \; P\left(\frac{S_n - n/2}{\sqrt{n}/2} \le x \right) \;=\; \Phi(x)\,, \tag{2}$$

dove abbiamo usato il fatto che $\mu := E(X_1) = \frac{1}{2}$ e $\sigma^2 := \mathrm{Var}(X_1) = E(X_1^2) - (E(X_1))^2 = \frac{1}{4}$, per cui $\sigma = \frac{1}{2}$. Naturalmente $\Phi(x)$ è la funzione di ripartizione della legge Normale standard:

$$\Phi(x) \;:=\; P\big(\mathcal{N}(0,1) \le x\big) \;=\; \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,\mathrm{d}t\,.$$

Possiamo riscrivere l'equazione (2) *per n grande* come

$$P\big(S_n - n/2 \le x\,\sqrt{n}/2\big) \;\approx\; \Phi(x)\,. \tag{3}$$

Torniamo al nostro problema di stimare $P\big(S_n - n/2 \le -k\big)$. Applichiamo innanzitutto la correzione di continuità: questo significa semplicemente notare che per k intero

$$P\big(S_n - n/2 \le -k\big) \;=\; P\big(S_n - n/2 \le -k + 0.5\big)\,,$$

dal momento che S_n assume solo valori interi (questo permette di migliorare l'approssimazione normale). Quindi possiamo applicare l'equazione (3): ponendo $-k = x\sqrt{n}/2$, cioè $x = -2k/\sqrt{n}$, e ricordando che $\Phi(-z) = 1 - \Phi(z)$, si ottiene

$$\boxed{P\big(S_n - n/2 \le -k\big) \;\approx\; 1 - \Phi\left(\frac{2(k-0.5)}{\sqrt{n}}\right)}\,. \tag{4}$$

Usando le tavole con i quantili della distribuzione Normale, possiamo dunque applicare questa stima con $n = 100, 1000, 10000$ e $k = 5, 50, 500$ rispettivamente, ottenendo

$$P\big(\text{45 teste o meno in 100 lanci}\big) \;\approx\; 1 - \Phi\left(\frac{2\cdot 4.5}{\sqrt{100}}\right) \;=\; 1 - \Phi(0.9) \;\approx\; 18\%$$

$$P\big(\text{450 teste o meno in 1000 lanci}\big) \;\approx\; 1 - \Phi\left(\frac{2\cdot 49.5}{\sqrt{1000}}\right) \;=\; 1 - \Phi(3.13) \;\approx\; 0.087\%$$

$$P\big(\text{4500 teste o meno in 10000 lanci}\big) \;\approx\; 1 - \Phi\left(\frac{2\cdot 499.5}{\sqrt{10000}}\right) \;=\; 1 - \Phi(9.99) \;\approx\; 0\%\,.$$

Il calcolo esatto di tali probabilità, mediante la formula (1) (e un computer), conferma la bontà di queste stime: infatti i valori esatti sono

$$\begin{aligned} P\big(\text{45 teste o meno in 100 lanci}\big) &= 0.184 \pm 10^{-3} \\ P\big(\text{450 teste o meno in 1000 lanci}\big) &= 0.000865 \pm 10^{-6} \\ P\big(\text{4500 teste o meno in 10000 lanci}\big) &= 7.75 \cdot 10^{-24} \pm 10^{-26}\,. \end{aligned}$$

Concludiamo con alcune osservazioni. Riscriviamo per semplicità la formula (4) senza correzione di continuità:

$$P\big(S_n - n/2 \le -k\big) \;\approx\; 1 - \Phi\left(\frac{2k}{\sqrt{n}}\right).$$

Cerchiamo di "leggere" questa equazione: il membro destro è una funzione di $k/\sqrt{n}$. Questo significa che, variando n e k, *la probabilità rimarrà (all'incirca) la stessa purché* $k/\sqrt{n}$ *resti costante*. Ricordiamoci che n è il numero di lanci, mentre k è lo scostamento dal numero medio di teste $(n/2)$: quindi questa equazione ci dice che, per n grande, la differenza k tra il numero di teste osservato e il numero medio $n/2$ sarà tipicamente dell'ordine di $\sqrt{n}$.

Infine, si osservi che abbiamo scritto l'equazione approssimata (3) affermando che fosse una buona stima *per* n *grande*. In questo caso, potendo calcolare le probabilità esattamente, abbiamo verificato *a posteriori* che è davvero così. Tuttavia uno vorrebbe avere informazioni più precise su tale approssimazione *a priori*. In questa direzione, possiamo citare il seguente risultato che rafforza il Teorema Limite Centrale, detto *stima di Berry-Esséen*: se le variabili $\{X_i\}_{i\in\mathbb{N}}$ sono i.i.d. con $E(X_1) =: \mu \in \mathbb{R}$, $\mathrm{var}(X_1) =: \sigma^2 \in (0,\infty)$ e inoltre $E(|X_1|^3) =: \rho < \infty$, allora per ogni $n \in \mathbb{N}$ e $x \in \mathbb{R}$

$$\left| P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) \;-\; \Phi(x) \right| \;\le\; \frac{3\rho}{\sigma^3\sqrt{n}}\,.$$

Una dimostrazione si può trovare nel capitolo XVI del libro di William Feller *An Introduction to Probability Theory and Its Applications*, Vol. II, Second edition, John Wiley & Sons (1971).