

Supponiamo che un fabbricante stia introducendo un nuovo tipo di batteria per un'automobile elettrica. La **durata osservata** x_i delle i -esima batteria è la realizzazione (valore assunto) di **una variabile aleatoria** X_i . Si può assumere che le X_i abbiano la stessa distribuzione incognita. Per ottenere informazioni sulla distribuzione si costruiscono e mettono in funzione un certo numero n di batterie. La durata di ciascuna batteria fornisce l'insieme di dati x_1, x_2, \dots, x_n . **La media campionaria** della durata è

$$\bar{x}_n = \frac{x_1 + \dots + x_n}{n}$$

Le lettere maiuscole X_i indicano la variabili aleatorie (prima dell'esperimento), quelle minuscole x_i ne indicano le realizzazione (il numero ottenuto dopo l'esperimento).

E' naturale chiedersi alcune cose sulla **media campionaria \bar{x}_n**

- c'entra qualcosa con il **valore atteso μ** delle **variabili X_i** ?
- posso dire che legge ha, ossia come è distribuita?

Alla prima domanda risponderà la **Legge dei Grandi Numeri**. Se **n è "molto" grande**, **\bar{x}_n** sarà molto prossima a **μ** (vedremo poi quanto prossima).

Rispondere alla seconda domanda in generale non è semplice, ma se **n è "abbastanza" grande** con buona approssimazione è una realizzazione di una variabile normale per il **Teorema del Limite Centrale** (vedremo poi come determinarne i parametri).

Siano $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots$ variabili aleatorie **indipendenti** e tutte con la **stessa distribuzione**, con valore atteso $\mathbb{E}(X_1) = \dots = \mathbb{E}(X_n) = \mu$ e **stessa varianza** $\text{var}(X_i) = \sigma^2$, $i = 1, 2, \dots, n$.

Sia $\mathbf{Y}_n = \mathbf{X}_1 + \dots + \mathbf{X}_n$, per la proprietà della media:

$$\mathbb{E}(\mathbf{Y}_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) \dots + \mathbb{E}(X_n) = n\mu$$

e per le proprietà della varianza

$$\text{var}(\mathbf{Y}_n) = \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n) = n\sigma^2$$

Si ha inoltre che

$$\bar{\mathbf{X}}_n = \frac{\mathbf{X}_1 + \dots + \mathbf{X}_n}{n} = \frac{\mathbf{Y}_n}{n}$$

per cui

$$\mathbb{E}(\bar{\mathbf{X}}_n) = \mu, \quad \text{var}(\bar{\mathbf{X}}_n) = \sigma^2/n$$

Esercizio. I risultati di un test sul livello di potassio nel sangue di un individuo variano sia a causa dell'imprecisione dello strumento di misurazione, sia perchè il livello stesso varia nel tempo. Sappiamo che per un certo individuo le letture successive del livello di potassio oscillano intorno a un valore atteso μ con deviazione standard $\sigma = 0.3$. Quattro letture specifiche generano i dati

$$3.6, \quad 3.9, \quad 3.4, \quad 3.5$$

La media campionaria per il livello medio di potassio per questa persona è

$$\frac{3.6 + 3.9 + 3.4 + 3.5}{4} = 3.6$$

mentre la deviazione standard della media campionaria è

$$\frac{\sigma}{\sqrt{n}} = \frac{0.3}{2} = 0.15$$

LEGGE DEI GRANDI NUMERI

Si vede che la variabile \bar{X}_n , che ha sempre valore atteso μ , ha varianza inversamente proporzionale ad n (ossia uguale a σ/n). Nel limite $n \rightarrow \infty$ (ossia per n molto grande) la varianza diventa nulla, questo significa che la media campionaria diventa una variabile certa con valore μ .

Questo risultato è più correttamente descritto dal seguente teorema:

Teorema. Si dimostra (ma non qui) che per $n \rightarrow \infty$ e per ogni a strettamente positivo:

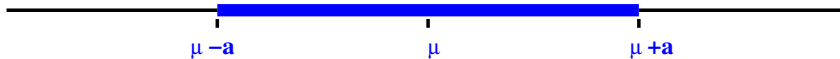
$$P(|\bar{X}_n - \mu| > a) \rightarrow 0$$

In altre parole la probabilità che \bar{X}_n sia diversa da μ diventa nulla quando n diventa molto grande. Con un certo abuso di termini e di notazione) possiamo dire che $\bar{X}_n \rightarrow \mu$ **quando** $n \rightarrow \infty$.

Se n è grande ma non infinito si avrà : $P(|\bar{X}_n - \mu| > a) \simeq 0$.

Cosa significa?

L'asserzione vale per ogni a strettamente positivo. Per fissare le idee immaginiamo che sia molto piccolo, ad esempio $a = 0.01$.



La probabilità che la media campionaria osservata \bar{x}_n cada fuori dall'intervallo colorato in figura è trascurabile ($\simeq 0$) se n è abbastanza grande.

La probabilità che la media campionaria osservata \bar{x}_n cada dentro l'intervallo colorato in figura è $\simeq 1$ se n è abbastanza grande.

MEDIA TEORICA \simeq MEDIA OSSERVATA

La **legge dei grandi numeri** dice che il valore osservato $\bar{x}_n = (x_1 + \dots + x_n)/n$ della variabile aleatoria $\bar{X}_n = (X_1 + \dots + X_n)/n$ è con grande probabilità **vicino a μ** se n è grande.

Quindi se **non conosco μ** , ne posso dare una stima con \bar{x}_n che è una realizzazione della statistica campionaria \bar{X}_n . Se n è grande, la probabilità che \bar{X}_n e μ siano "molto diversi" è quasi zero.

Si noti che μ è la media teorica *a priori* di ciascuna osservazione (un parametro della distribuzione delle X_i) mentre $\bar{x}_n = (x_1 + \dots + x_n)/n$ è una media *a posteriori* delle osservazioni.

IL TEOREMA DEL LIMITE CENTRALE

Siano $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots$ variabili aleatorie **indipendenti** e tutte con la **stessa distribuzione**, con valore atteso $\mathbb{E}(X_1) = \dots = \mathbb{E}(X_n) = \mu$ e **stessa varianza** $\text{var}(X_i) = \sigma^2$, $i = 1, 2, \dots, n$.

Sia $\mathbf{Y}_n = \mathbf{X}_1 + \dots + \mathbf{X}_n$, abbiamo visto che:

$$\mathbb{E}(\mathbf{Y}_n) = n\mu \qquad \text{var}(\mathbf{Y}_n) = n\sigma^2$$

Teorema. Per le proprietà del valore atteso e della varianza, la variabile

$$\bar{\mathbf{Z}}_n = \frac{\mathbf{Y}_n - n\mu}{\sigma\sqrt{n}}$$

ha **media 0** e **varianza 1**. Inoltre si dimostra (ma non qui) che per n sufficientemente **grande** (nel limite $n \rightarrow \infty$) ha **distribuzione normale**.

In altre parole, nel limite $n \rightarrow \infty$

$$\bar{Z}_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}} \rightarrow N(0, 1) = Z$$

Quindi per n grande (ma non infinito) la probabilità $P(\bar{Z}_n < x)$ è circa uguale a $P(Z < x)$ dove Z è una **normale standard** $N(0, 1)$.

Se vogliamo calcolare probabilità relative a \bar{Z}_n possiamo utilizzare quelle relative a $N(0,1)$ come approssimazione.

Tenendo presente che $Y_n/n = \bar{X}_n$ possiamo anche scrivere

$$\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1) = Z$$

Si ricordi ancora una volta che in questo contesto \bar{X}_n è la statistica campionaria di cui la \bar{x}_n è una realizzazione (risultato di una misurazione o esperimento). Anche \bar{Y}_n e Z_n sono statistiche campionarie.

DISTRIBUZIONE DELLE MEDIA CAMPIONARIA (VARIANZA NOTA)

Anche la somma di n variabili identicamente distribuite è approssimativamente normale, infatti:

$$\mathbf{Y}_n = \sigma\sqrt{n}\bar{\mathbf{Z}}_n + n\mu \simeq \sigma\sqrt{n}\mathbf{N}(\mathbf{0}, \mathbf{1}) + n\mu = \mathbf{N}(n\mu, \sigma\sqrt{n})$$

La **media campionaria** è anch'essa approssimativamente normale:

$$\bar{\mathbf{X}}_n = \frac{\mathbf{Y}_n}{n} \simeq \frac{\sigma}{\sqrt{n}}\mathbf{N}(\mathbf{0}, \mathbf{1}) + \mu \simeq N(\mu, \sigma/\sqrt{n}).$$

Il **valore atteso** $\mu_{\bar{x}_n}$ della media campionaria è uguale a μ mentre **la varianza** $\sigma_{\bar{x}_n}$ della media campionaria è uguale a σ/\sqrt{n} e quindi decresce come l'inverso della radice quadrata della dimensione del campione. Come abbiamo visto con la legge dei grandi numeri, per un ipotetico campione infinito, la media campionaria coincide identicamente con μ .

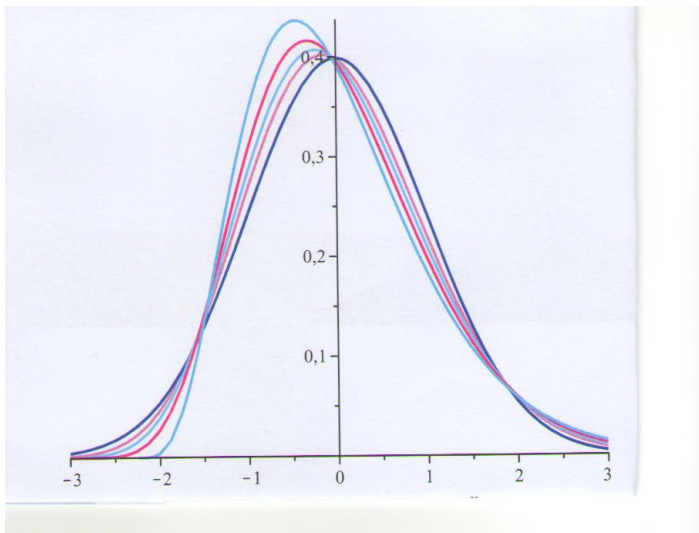
E IN PRATICA?

Il Teorema del limite centrale lascia aperta la questione di **quanto il campione debba essere numeroso** affinché **l'approssimazione sia valida**.

Se le variabili X_i sono variabili gaussiane il problema non si pone perché qualsiasi combinazione lineare di variabili gaussiane è anch'essa gaussiana. In tal caso, la distribuzione di X_n è gaussiana per ogni n (quella di \bar{Z}_n è normale standard). Altrimenti si accetta la seguente regola pratica.

- Se la legge delle X_1, \dots, X_n non è troppo asimmetrica, a livello empirico si è stabilito che **$n \geq 30$** va bene.

Per convincerci dell'asserzione consideriamo il caso in cui la distribuzione delle X_i è esponenziale (non importa qui sapere la definizione). **Confrontiamo i grafici delle densità delle somme standardizzate** di queste variabili indipendenti per **$n = 5$, $n = 10$, $n = 20$ e $n = 50$** con il grafico di una $N(0, 1)$.



In blu la $N(0, 1)$, in azzurro la \bar{Z}_5 , in magenta la \bar{Z}_{10} , in celeste la \bar{Z}_{20} e in violetto la \bar{Z}_{50} .

Esercizio. Il **livello di colesterolo** nel sangue di una popolazione di lavoratori ha una **media 202** e una **deviazione standard 14**. Viene selezionato un **campione di 36 lavoratori**, si approssimi la **probabilità che la media campionaria** dei loro livelli di colesterolo sia **compresa tra 198 e 206**.

\bar{X}_{36} ha distribuzione approssimativamente normale con valore atteso $\mu = 202$ e deviazione standard $\sigma/\sqrt{n} = 14/\sqrt{36} = 7/3$, quindi

$$\bar{Z}_{36} = \frac{\bar{X}_{36} - 202}{7/3} \simeq N(0, 1) = Z$$

$$\begin{aligned} P(198 < \bar{X}_{36} < 206) &= P\left(\frac{198 - 202}{7/3} < \bar{Z}_{36} < \frac{206 - 202}{7/3}\right) \\ &\simeq P(-1.714 < Z < 1.714) = \phi(1.714) - \phi(-1.714) = 0.913 \end{aligned}$$

dove ϕ è la solita funzione di distribuzione della normale standard.

RIASSUMENDO

Per un campione di numerosità n estratto da una popolazione con distribuzione di media μ e varianza σ si ha che

- ▶ $\mu_{\bar{X}_n} = \mu$ per ogni n ,
- ▶ $\sigma_{\bar{X}_n} = \sigma/\sqrt{n}$ per ogni n ,
- ▶ \bar{X}_n è distribuita **normalmente** per ogni n se le X_i **sono gaussiane**,
- ▶ \bar{X}_n è distribuita **quasi normalmente** se n è grande ($n > 30$) anche se le X_i **non sono gaussiane**.
- ▶ $\bar{X}_n \rightarrow \mu$ **quando** $n \rightarrow \infty$ (non realizzabile in pratica).

Si noti che la media campionaria è approssima il parametro μ ma la sua precisione ($\sigma_{\bar{X}_n} = \sigma/\sqrt{n}$) viene definita in base ad una **varianza σ nota**. Vedremo cosa fare quando la **varianza è incognita**.