

Suppose a manufacturer is introducing a new type of battery for an electric car. The **observed lifetime x_i** of the i -th battery is the realization (measured value, experimental output) of a **random variable X_i** . We can assume that all **X_i** have the same unknown distribution. To obtain information about their distribution, a certain number **n** of batteries are built and operated. The lifetime of each battery provides the data set **$x_1, x_2, \dots, x_n$** . **The sample mean** of their lifetime is

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

In the following, the uppercase letters **X_i** will indicate random variables, while the lowercase letters **x_i** will indicate their realizations *i.e.*, the numerical values obtained by experiment or measurement.

INFERENCE STATISTICS

It's natural to wonder a couple of things about the **sample mean $\bar{x}$**

- does it have something to do with the **expected value μ** of **variables X_i** ?
- is it possible to know how it is distributed?

The first question is answered by the **Law of Large Numbers**. If **n is "very" large**, **$\bar{x}$** will be very close to **μ**

Answering the second question in general is not easy, but at least we can say how much **$\bar{x}$** typically differs from **μ** .

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

Sia

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

Per la proprietà della media:

$$\mathbb{E}(\bar{\mathbf{X}}) = \frac{1}{n} \left(\mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_n) \right) = \mu$$

e per le proprietà della varianza

$$\text{var}(\bar{\mathbf{X}}) = \frac{1}{n^2} \left(\text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n) \right) = \frac{\sigma^2}{n}$$

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 $\bar{X}$, 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:

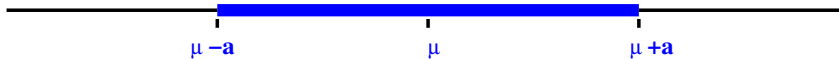
Teorema. Si dimostra (ma non qui) che per $n \rightarrow \infty$ e per ogni a strettamente positivo:

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

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

Se n è grande ma non infinito si avrà : $P(|\bar{X} - \mu| > a) \simeq 0$. 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 $\bar{x}$ cada fuori dall'intervallo colorato in figura è trascurabile ($\simeq 0$) se n è abbastanza grande.

La probabilità che la media campionaria osservata $\bar{x}$ 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 $\bar{x} = (x_1 + \dots + x_n)/n$ della variabile aleatoria $\bar{X} = (X_1 + \dots + X_n)/n$ è con grande probabilità **vicino a μ** se n è grande.

Quindi se **non conosco μ** , ne posso dare una stima con $\bar{x}$ che è una realizzazione della statistica campionaria $\bar{X}$. Se n è grande, la probabilità che $\bar{X}$ 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 $\bar{x} = (x_1 + \dots + x_n)/n$ 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 $\sigma/\sqrt{n}$ 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 .

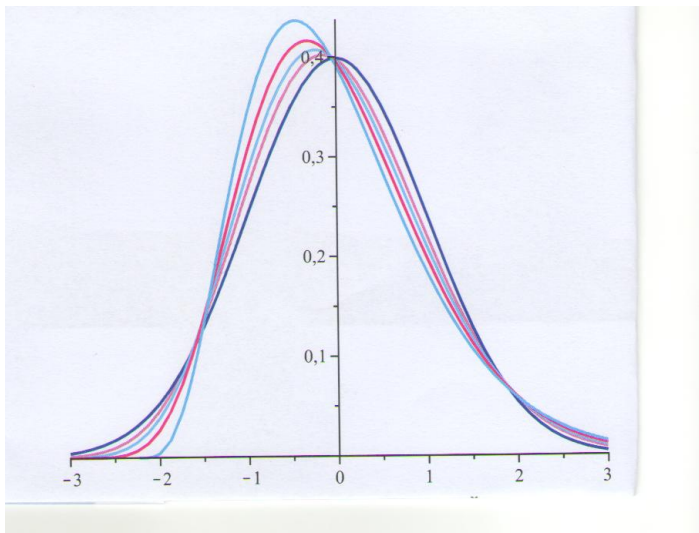
E IN PRATICA?

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**.

STIME PUNTUALI

Definizione 1

Se la stima di un parametro della distribuzione delle X_i è data da un singolo numero ottenuto dalle misurazioni relative a un campione, tale valore è detto stima puntuale del parametro.

Esempio La media campionaria osservata $\bar{x}$ è una stima del parametro μ della distribuzione delle variabili aleatorie X_i . Il numero (stima) $\bar{x}$ è una realizzazione dello stimatore X_n .

La statistica $\bar{X}$ ha valore atteso μ , quindi, in questo caso, il valore atteso dello stimatore è uguale al parametro che si vuole stimare.

Definizione 2

Uno stimatore con valore atteso uguale al parametro che si vuole stimare si dice corretto (o non distorto).

La media campionaria $\bar{X} = (X_1 + \dots + X_n)/n$ è quindi uno stimatore corretto di μ .

Definizione 3

Se due statistiche sono entrambe stimatori corretti di un parametro, la statistica con minore varianza è detta stimatore più efficiente.

La stima ottenuta con la media campionaria è tanto più accurata tanto più il campione è numeroso. Infatti $\bar{x}$ è un valore osservato di $\bar{X}$ la cui varianza è σ^2/n . Pertanto **se si considerano due campioni di taglia m e n con $m > n$ si avrà che $\bar{X}_m$ è più efficiente di $\bar{X}_n$ nello stimare il parametro μ .**

In seguito potremo confrontare con l'efficienza di altri stimatori corretti diversi dalla media campionaria.

STIMA DELLA VARIANZA CAMPIONARIA

Abbiamo visto che una stima puntuale corretta per il valore atteso μ delle variabili aleatorie X_i è $\bar{x} = (x_1 + \dots + x_n)/n$.

Una stima puntuale della varianza σ^2 delle variabili aleatorie X_i è la varianza campionaria s^2

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots (x_n - \bar{x})^2}{n - 1}$$

Mostriamo che anche questa stima puntuale è **corretta** ossia che il valore medio dello stimatore

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots (X_n - \bar{X})^2}{n - 1}$$

è uguale alla varianza σ^2 delle variabili aleatorie X_i .

First of all, given that all X_i are equally distributed, we can write

$$\mathbb{E}(\mathbf{S}^2) = \frac{n}{n-1} \mathbb{E}[(\mathbf{X}_1 - \bar{\mathbf{X}})^2].$$

We know that $\mathbb{E}[X_1^2] = \mu^2 + \sigma^2$ and that $\mathbb{E}[\bar{X}^2] = \mu^2 + \frac{\sigma^2}{n}$. Moreover, keeping in mind that $i \neq 1$ implies $\mathbb{E}(X_1 X_i) = \mu^2$, it is easy to show that $\mathbb{E}[X_1 \bar{X}] = \mu^2 + \frac{\sigma^2}{n}$ (exercise). We will then have

$$\mathbb{E}[(\mathbf{X}_1 - \bar{\mathbf{X}})^2] = \mathbb{E}[\mathbf{X}_1^2] + \mathbb{E}[\bar{\mathbf{X}}^2] - 2\mathbb{E}[\mathbf{X}_1 \bar{\mathbf{X}}] = \frac{n-1}{n} \sigma^2$$

and ultimately

$$\mathbb{E}(\mathbf{S}^2) = \sigma^2.$$

which means that the statistical estimate is unbiased.

Se il valore atteso μ delle variabili X_i è noto è naturale prendere come stima di σ^2 il numero

$$a_n = \frac{(\mathbf{x}_1 - \mu)^2 + (\mathbf{x}_2 - \mu)^2 + \dots (\mathbf{x}_n - \mu)^2}{n}$$

In effetti a_n è una stima corretta se il valore atteso μ delle variabili X_i è noto. Sia infatti a_n una realizzazione della variabile aleatoria

$$\mathbf{A}_n = \frac{(\mathbf{X}_1 - \mu)^2 + (\mathbf{X}_2 - \mu)^2 + \dots (\mathbf{X}_n - \mu)^2}{n}$$

Verificare che $\mathbb{E}(A_n) = \sigma^2$ è semplice, infatti dalla equazione qui sopra, tenendo presente che le variabili X_i sono identicamente distribuite, si ha che

$$\mathbb{E}(\mathbf{A}_n) = \mathbb{E}[(\mathbf{X}_1 - \mu)^2] = \sigma^2$$

È molto improbabile che il valore atteso sia noto e la varianza no, in generale **anche il valore atteso è ignoto** . Anche se in questo casi sembrerebbe ragionevole stimare la varianza con

$$b_n = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots (x_n - \bar{x})^2}{n},$$

ma per quello che si è detto in precedenza **questa quantità non può essere una stima corretta della varianza**. Infatti, avendo definito

$$B_n = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots (X_n - \bar{X})^2}{n},$$

è semplice verificare che $\mathbb{E}(B_n) \neq \sigma^2$, infatti,

$$\mathbb{E}(B_n) = \frac{n-1}{n} \mathbb{E}(S^2) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

In definitiva, dato che il caso più comune è quello in cui sia il valore atteso che la varianza delle variabili aleatorie X_i sono ignoti, stimeremo μ con

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

e σ^2 con

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}.$$

Questo significa anche che l'errore della stima di μ è dell'ordine di $s/\sqrt{n-1} \simeq s/\sqrt{n}$, dove l'uguaglianza approssimata vale per grandi n . In accordo con la legge dei grandi numeri, l'errore si annulla quando $n \rightarrow \infty$. Possiamo scrivere in modo poco corretto ma espressivo:

$$\mu = \bar{x} \pm s/\sqrt{n}.$$