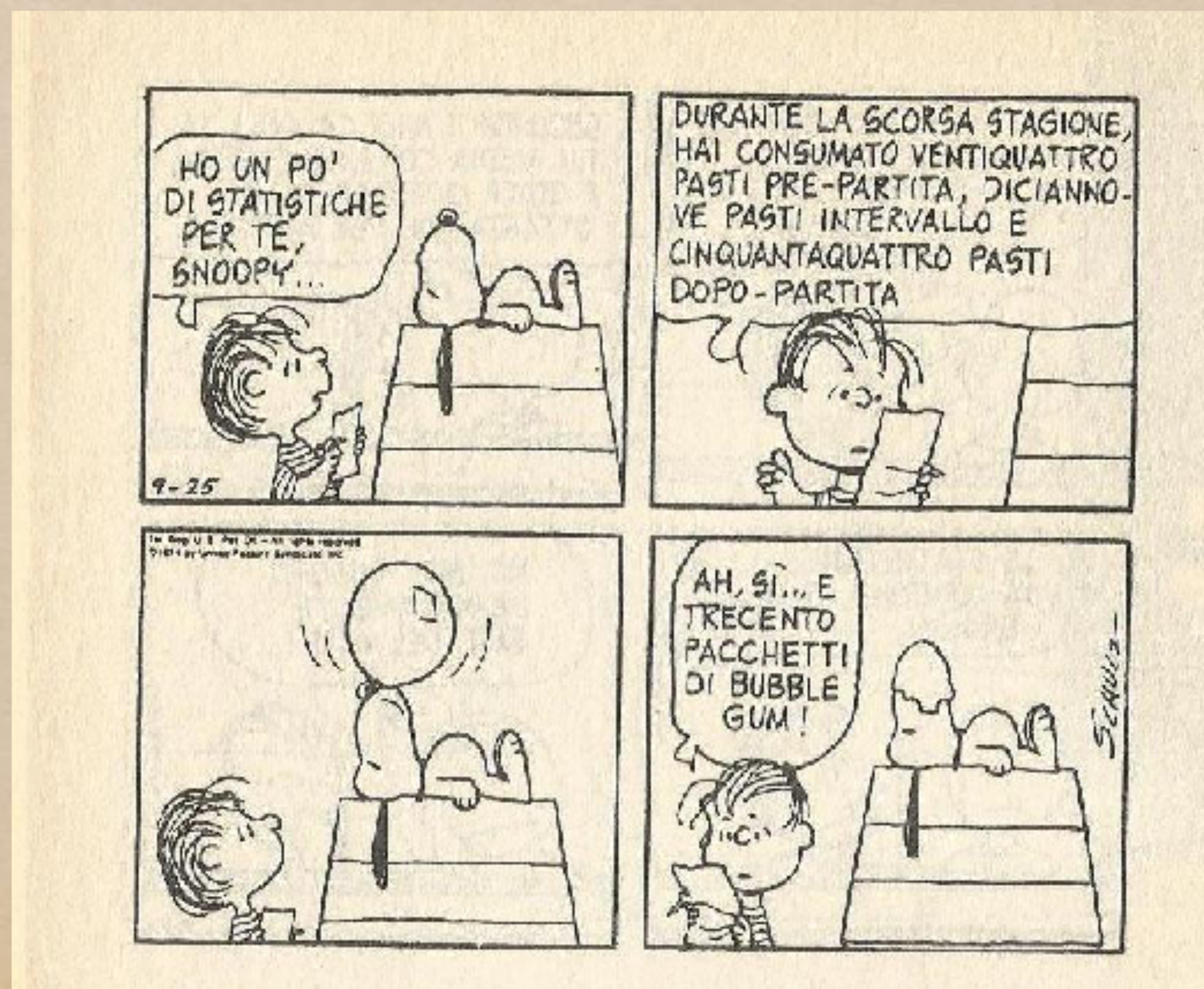
Elementi di probabilità e statistica

per principianti, specialmente se liceali…

Stefano Covino
INAF /Oss. Astronomico di Brera



Statistica e probabilità

Essenzialmente, la statistica è una disciplina che si occupa di ottenere predizioni per un determinato fenomeno quando i dati a disposizione sono limitati e/o incompleti.

Blaise Pascal (1623-1662)

È una disciplina da natali illustri. Più o meno convenzionalmente la si fa risalire agli studi di Blaise Pascal e Pierre de Fermat verso la metà del diciassettesimo secolo.

Pierre de Fermat (1601-1665)

Alcune definizioni

Evento: qualsiasi affermazione alla quale in seguito ad un esperimento o osservazione sia possibile assegnare un "grado di verità".

Ad esempio, in seguito al lancio di un dado un evento può essere: "esce un numero pari", "un numero maggiore di 7", "un numero minore o uguale a 6", ecc.

Nel primo caso abbiamo un evento casuale o aleatorio, nel secondo un evento impossibile e nell'ultimo un evento certo.

La frequenza relativa

Seguiamo un approccio "sperimentale", e definiamo frequenza relativa il rapporto osservato fra eventi "favorevoli", v, ed eventi totali, n:

$$f(E) = v/n$$

Ad esempio, lanciando una moneta un numero elevato di volte si osserva che la frequenza dell'evento "testa" tende a 0.5. O, ancora, lanciate un dado e misurerete che la frequenza di una qualunque delle facce tenderà ad essere 1/6.

La probabilità (classica)

La probabilità, $p(E)$, di un evento è il rapporto fra il numero di casi “favorevoli", f, è quello dei casi “possibili”, u, nel caso che siano tutti equiprobabili:

$$p(E) = f/u$$

Ad esempio, estraiamo una carta da un mazzo da 40, e ci domandiamo quale sia la probabilità che sia un asso:

$$p(\text{asso}) = 4/40 = 1/10 = 0.1 = 10\%$$

La probabilità è un numero fra $0 \leq p \leq 1$.

La legge empirica del caso (o dei grandi numeri)

Di fatto, si osserva che al tendere all'infinito del numero di "tentativi", la frequenza relativa tende alla probabilità classica:

$$f(E) \rightarrow p(E)$$

Si tratta di un "limite", ovvero dell'estrapolazione del comportamento di una relazione quando un parametro si avvicina asintoticamente ad un valore. Uno degli strumenti più potenti dell'analisi matematica.

I limiti

In matematica, il concetto di limite serve a descrivere l'andamento di una funzione all'avvicinarsi del suo argomento a un dato valore.

$$\lim_{x \to 8} \frac{1}{x-8} = \infty$$

$$\lim_{x \to 5} \frac{1}{x-5} = 5$$

Augustin-Louis Cauchy
(1789-1857)

$$\lim_{x \to -1^+} \frac{2x - x^2}{x+1} = -\infty$$

$$\lim_{x \to -1^-} \frac{2x - x^2}{x+1} = +\infty$$

Formulazione assiomatica della probabilità

La formulazione moderna della teoria della probabilità risale agli anni '30 del 900 e può essere espressa con riferimento alla teoria degli insiemi.

Ad esempio, sempre in riferimento al dado, lo spazio degli eventi è l'insieme {1,2,3,4,5,6}. Un sottinsieme dello stesso forma i casi "favorevoli", ad esempio i numeri pari {2,4,6}. Si possono definire gli insiemi complementari, unione ("or"), intersezione ("and"), ecc.

Andrej Nikolaevič Kolmogorov
(1903-1987)

Definizione operativa

La probabilità p(E) di un evento E è una funzione che ad ogni evento dello spazio totale U associa un numero in modo che siano verificati tre assiomi:

1. Per ogni evento $p(E) \geq 0$
2. Allo spazio totale corrisponde $p(U)=1$
3. Dati n eventi $E_1, E_2, \ldots, E_n$ a due a due incompatibili si ha: $p(E_1 \cup E_2 \cup \ldots \cup E_n) = p(E_1)+p(E_2)+\ldots+p(E_n)$

Esempi…

Lanciando un dado quale è la probabilità che esca 6 o un numero dispari?

- spazio totale $U = \{1,2,3,4,5,6\}$
- eventi favorevoli $\{1,3,5\} \cup \{6\} = \{1,3,5,6\} = E_1 \cup E_2$
- $p(E_1) = 3/6 = 1/2$, $p(E_2) = 1/6$
- $p(E_1)+(E_2) = 1/6+3/6 = 4/6 = 2/3 \approx 66.7\%$

Oppure, da un mazzo di carte da 40, quale è la probabilità di estrarre un re o una carta di fiori?

- $p(E_1) = 4/40 = 1/10$, $p(E_2) = 10/40 = 1/4$, $p(E_1 \cap E_2) = 1/40$
- $p(E_1)+(E_2)-p(E_1 \cap E_2) = 4/40+10/40-1/40 = 13/40 \approx 32.5\%$

Altri esempi…

In un sacchetto ci sono palline di tre colori, bianche rosse e blu. Sappiamo che la probabilità di estrarre una pallina rossa è 1/3, ed una blu 1/5. Quale è la probabilità di estrarre una pallina bianca?

- $p(E_1) = 1/3$, $p(E_2) = 1/5$, $p(E_1 \cup E_2) = p(E_1)+p(E_2) = 8/15$
- $p(E_3) = 1-[p(E_1)+p(E_2)] = 1-8/15 = 7/15 \approx 46.7\%$

La probabilità composta

Consideriamo due eventi, E_1 ed E_2, quale è la probabilità che si verifichino entrambi?

Ci possono essere due casi. Gli eventi sono stocasticamente indipendenti oppure no. Indipendenti significa che il verificarsi dell'uno non influenza il verificarsi dell'altro. Ad esempio il lancio di un dado non influenza il risultato del lancio successivo.

Se gli eventi sono indipendenti si ha:

$$p(E_1 \cap E_2) = p(E_1) * p(E_2)$$

Esempio

Se ho un mazzo di carte da 40 posso domandarmi la probabilità di estrarre due volte un re se dopo la prima estrazione rimetto la carta nel mazzo.

Gli eventi sono indipendenti e la probabilità di ciascuno è 4/40=1/10.

E quindi la probabilità composta sarà:

- $p(E_1)*p(E_2) = 1/10*1/10 = 1/100 = 1\%$

La probabilità condizionata

Se gli eventi sono invece dipendenti l'avverarsi di uno condiziona la probabilità dell'altro.

In questo caso si ha:

$$p(E_1 \cap E_2) = p(E_1) * p(E_2|E_1)$$

Ad esempio la probabilità di estrarre due re <u>senza rimettere la carta nel mazzo</u> sarà:

- $p(E_1) = 4/40 = 1/10$, $p(E_2|E_1) = 3/39$
- $p(E_1 \cap E_2) = 1/10 * 3/39 = 3/390 = 1/130 \approx 0.8\%$

Le distribuzioni di probabilità

Ci si possono porre, ovviamente, domande più complesse. Ad esempio, quale è la probabilità di avere 4 teste e 6 croci dopo 10 lanci di una moneta?

La risposta a questa ed altre domande similari è data dalla distribuzione binomiale, sviluppata studiando lo svolgersi di fenomeni casuali ognuno indipendente dai precedenti con probabilità "p" di successo e "q=1-p" di insuccesso.

La distribuzione binomiale (o di Bernoulli)

Supponendo allora di avere un fenomeno ripetuto “n” volte, con “k” casi favorevoli, e con probabilità “p” per gli stessi (“q=1-p”), il risultato è:

$$P = \binom{n}{k} p^k q^{n-k}$$

Jakob Bernouille
(1654-1705)

Esempio

Come si diceva, lanciamo una moneta 10 volte e determiniamo la probabilità di avere 4 teste e 6 croci:

- p(testa)=0.5, p(croce)=1-p(testa)=0.5
- n=10, k=4
- Il coefficiente binomiale $\binom{n}{k} = 210$
- $P \approx 0.205 = 20.5\%$

Parente della distribuzione binomia è la cosiddetta distribuzione ipergeometrica che tratta il caso di "estrazione senza reinserimento".

La distribuzione di Poisson

La distribuzione binomiale per grandi numeri e probabilità piccole ($n \rightarrow \infty$, $p \rightarrow 0$) assume una forma diversa e di comune applicazione per una grande varietà di problemi:

$$P = \frac{\mu^k}{k!} \cdot e^{-\mu}$$

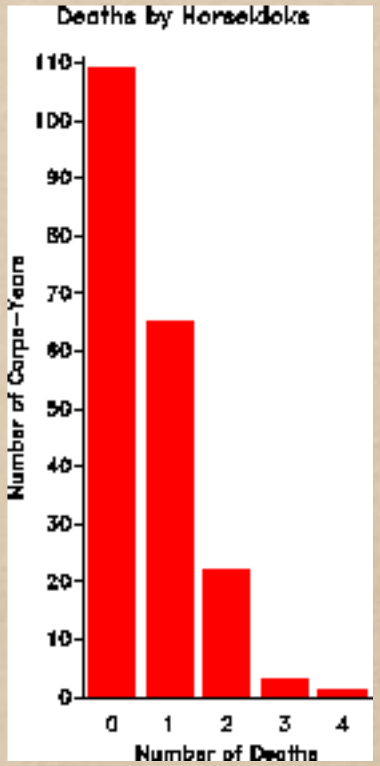
che fornisce la probabilità di avere il valor μ(=n*p) un numero k di volte per un gran numero di tentativi.

Siméon-Denis Poisson
(1654-1705)

Esempio di applicazione

La distribuzione di Poisson descrive in generale tutti i fenomeni che richiedono “conteggi”. Ad esempio, caso classico, il numero di morti per calcio da cavallo nei reggimenti di cavalleria prussiani:

Nr. morti/regg./anno	Nr. reg/anno
0	109
1	65
2	22
3	3
4	1



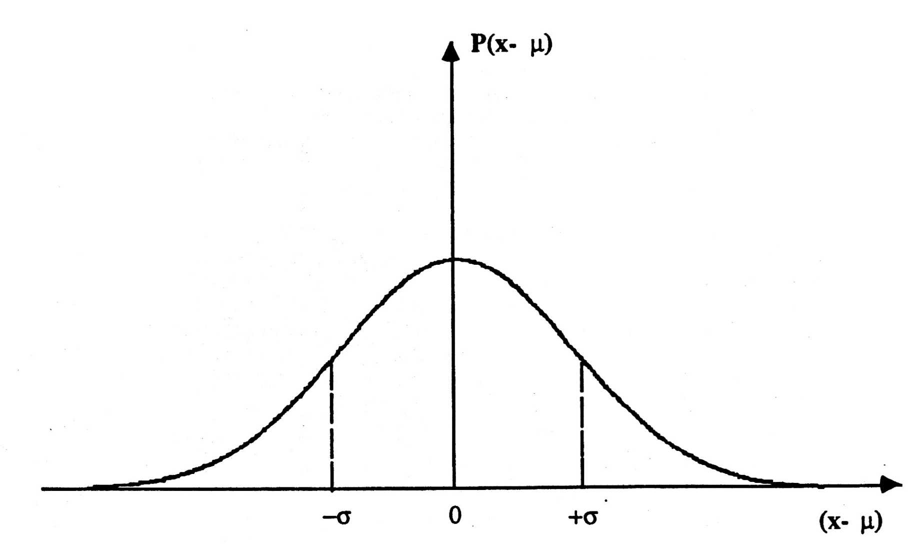
μ=0.61 morti/anno

La stessa distribuzione è alla base degli studi sui decadimenti radioattivi, le code di utilizzo, studi epidemiologici, ecc.

La distribuzione normale
(o Gaussiana)

Nel caso in cui, inoltre, μ(=n*p) è un valore "grande", e sempre con n→∞, la curva poissoniana prende una forma analiticamente molto vantaggiosa nota come curva gaussiana:

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} \mathrm{f(x)} = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} \cdot \mathrm{e}^{-1/2(\mathrm{x}-\mu)^2/\sigma^2}\, dx$$



Carl Friedrich Gauss
(1777-1855)

Esempio di applicazione

In generale la distribuzione di un gran numero di misure di una quantità fisica segue la statistica gaussiana. Questo permette di calcolare il valore più probabile, μ, ed anche i cosiddetti intervalli di confidenza (μ-σ, μ+σ).

- L'intervallo entro +/-σ corrisponde a circa il 68% di probabilità.
- L'intervallo entro +/-2σ a circa il 95%.
- L'intervallo entro +/-3σ, a circa il 99.7%.

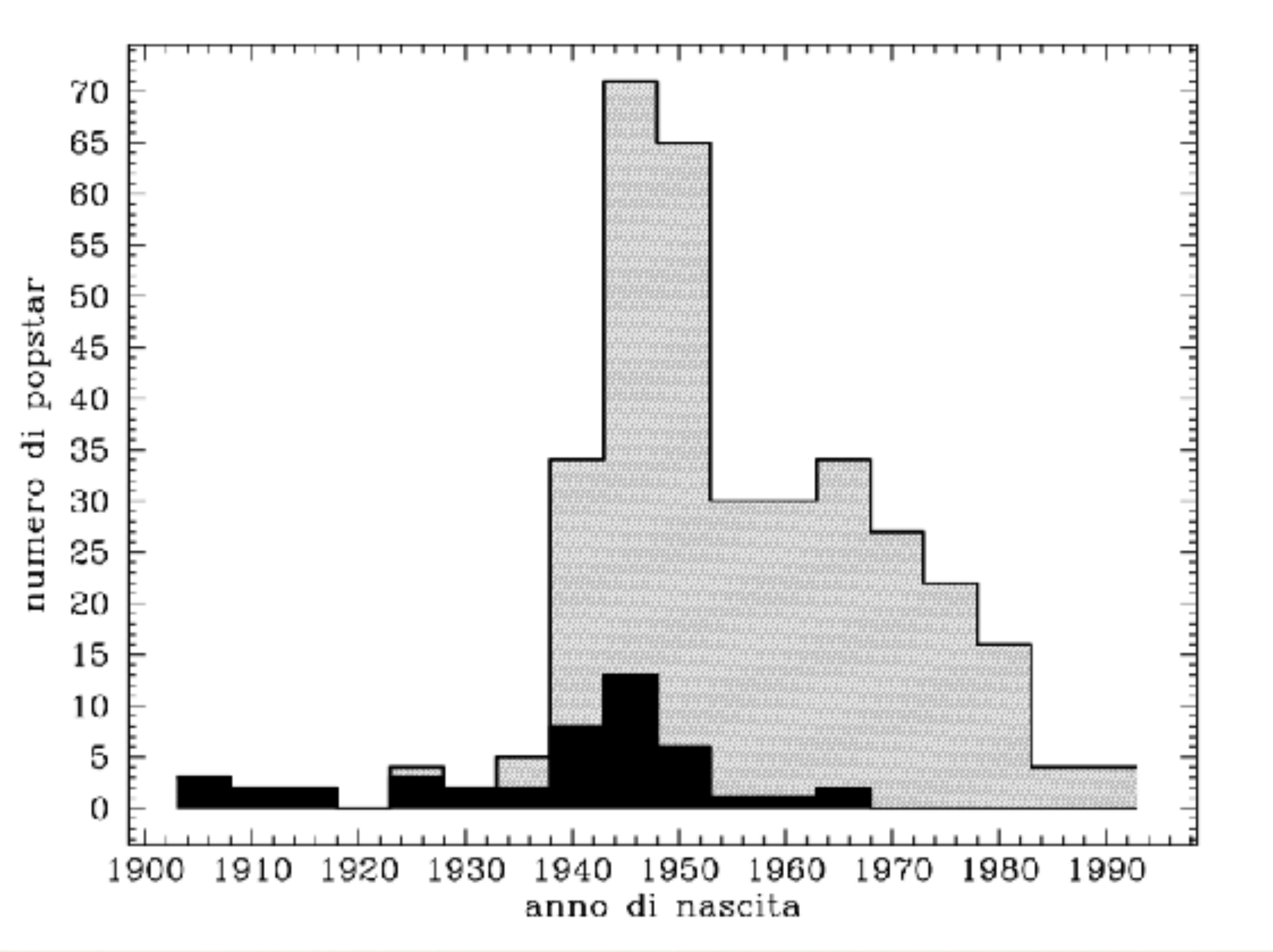
È molto probabile che la distribuzione dei vostri pesi o altezze seguano una statistica normale.

Esempio: il 2016 e le pop stars

È vero che il 2016 ha visto una moria peculiare di musicisti di successo?

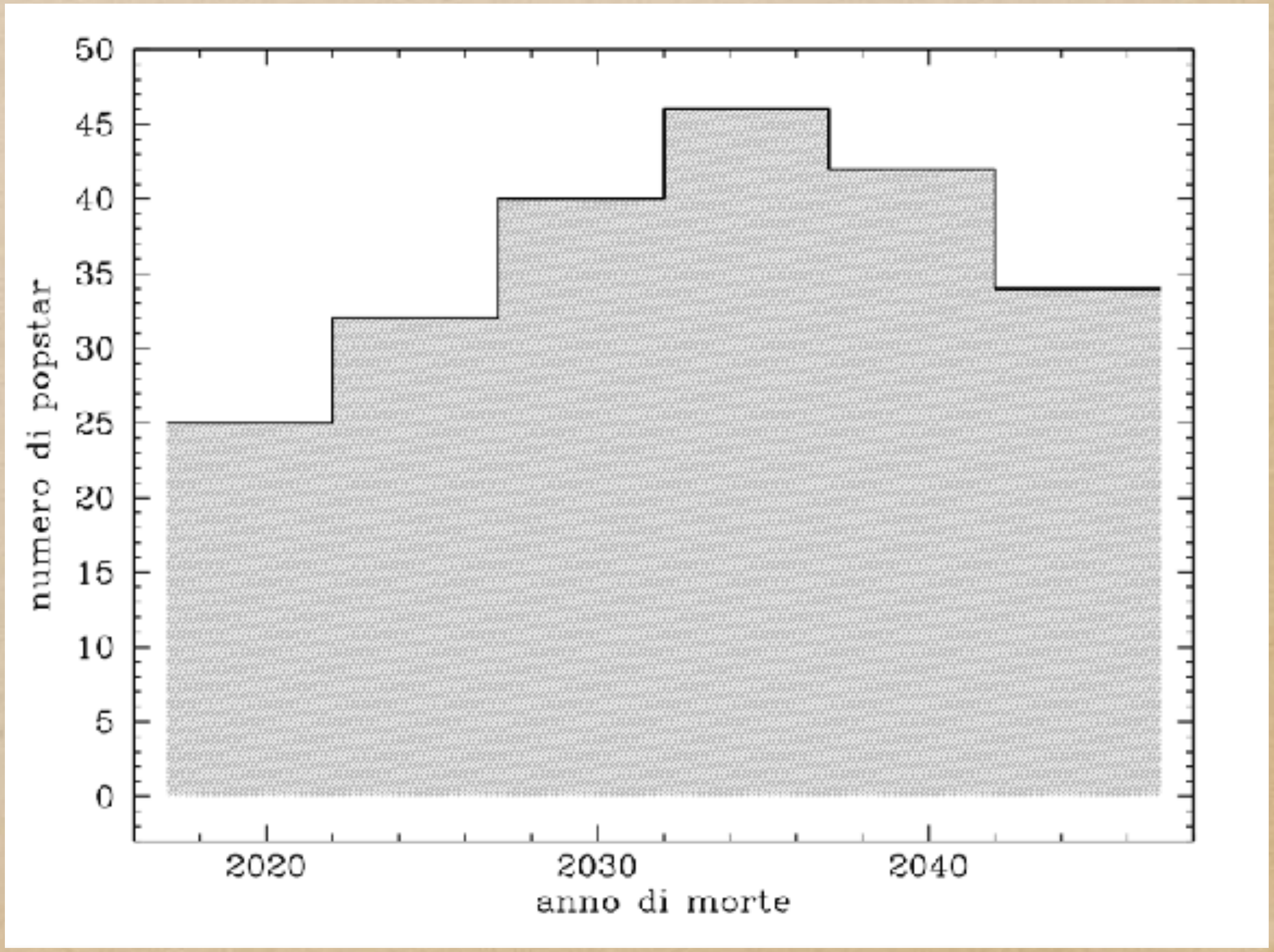
- I decessi sono tanti, si è detto, ma tanti rispetto a cosa?

Affrontiamo la questione correttamente, e definiamo il campione dei dati. Selezioniamo le pop stars internazionali basandoci sui dischi venduti in carriera, ecc. e vediamo la distribuzione degli anni di nascita…



La gran parte delle pop stars sono nate negli anni '40 e '50. Probabilmente il risultato composto dell'andamento demografico e della crescita economica.

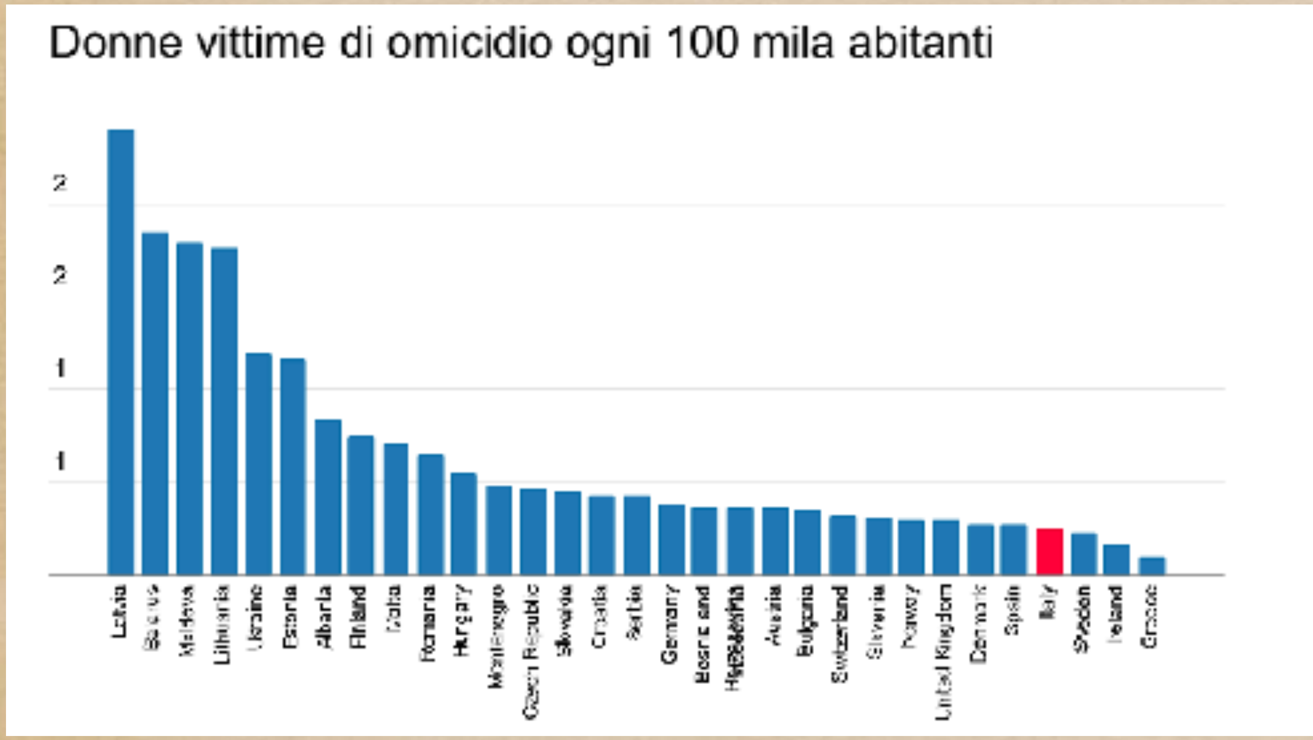
Esistono naturalmente statistiche sulla "speranza di vita" di persone che vivono nei paesi occidentali a seconda dell'anno di nascita. Se le applichiamo al campione delle pop stars possiamo prevedere i tassi di mortalità medi per i prossimi anni...

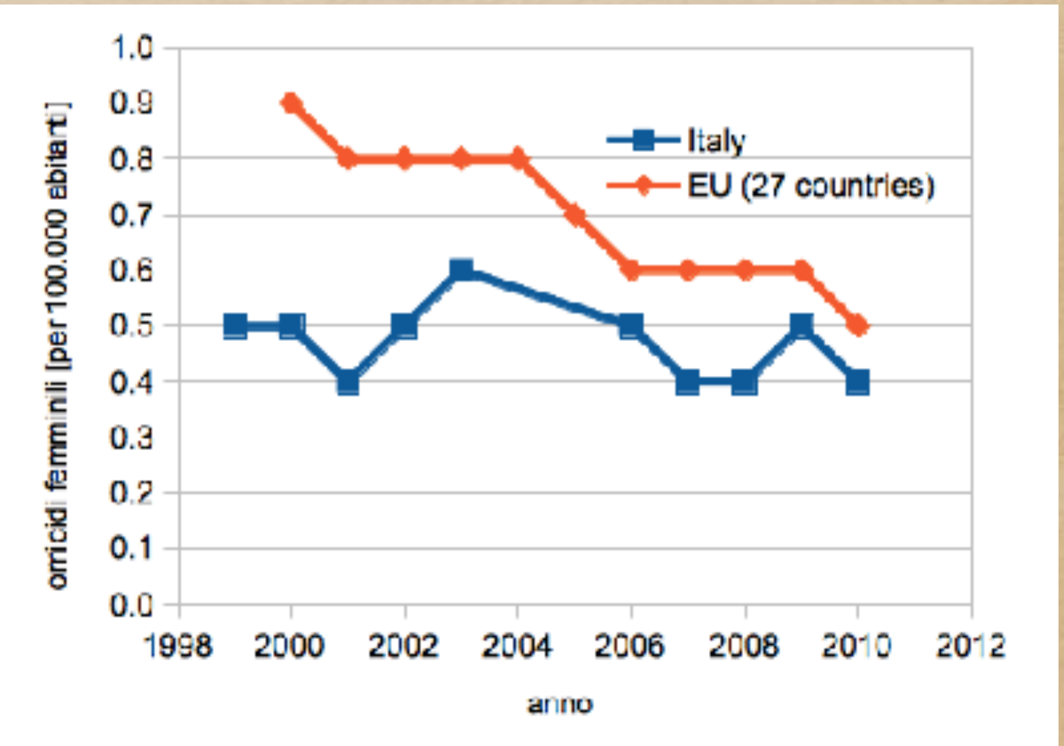


È, non dimenticate, la statistica fornisce risposte corrette se le domande sono ben poste...

Esempio: il femminicidio

Argomento, per ovvi motivi, di grande risonanza mediatica. Ma esiste realmente un'emergenza femminicidio?

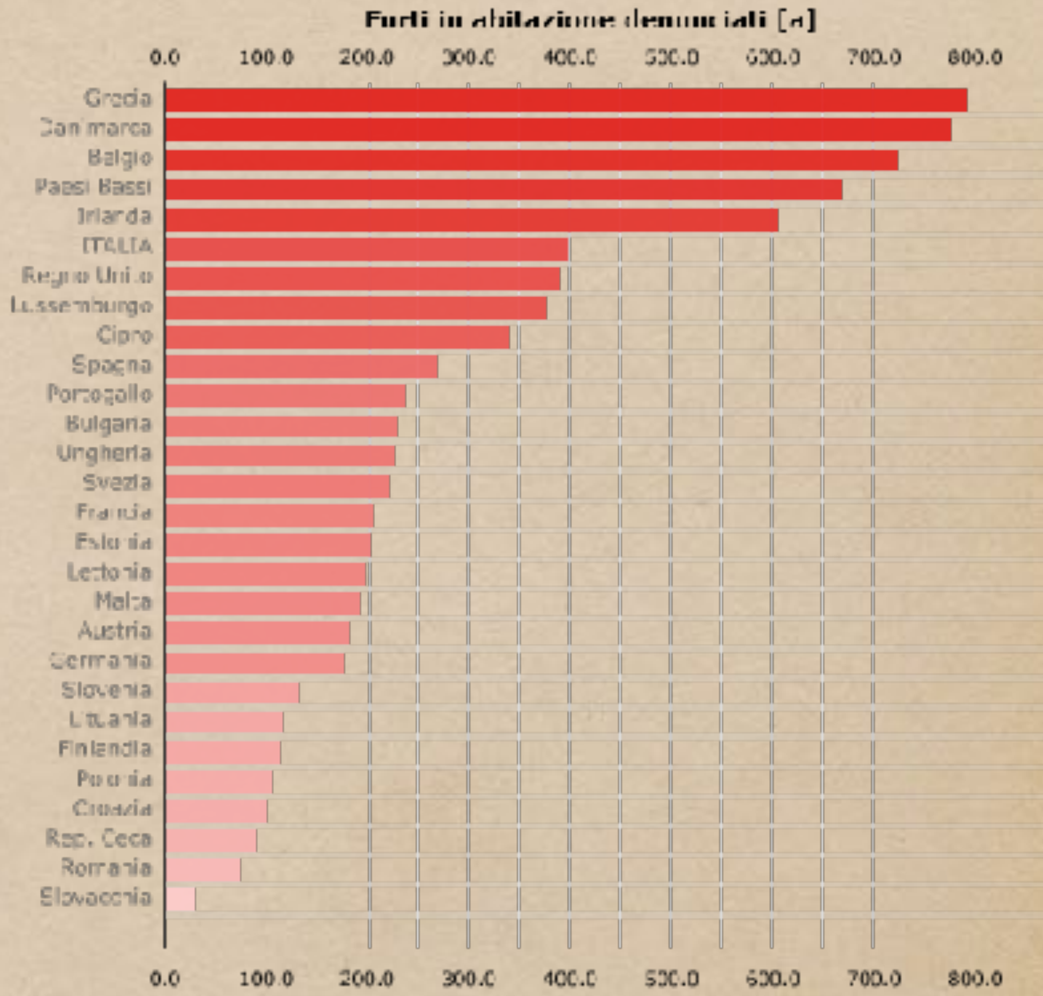




Il problema esiste, ma certamente una insufficiente predisposizione a verificare i dati nel mondo dell'informazione, incluso gli utenti, genera notizie distorte.

Esempio: reati predatori

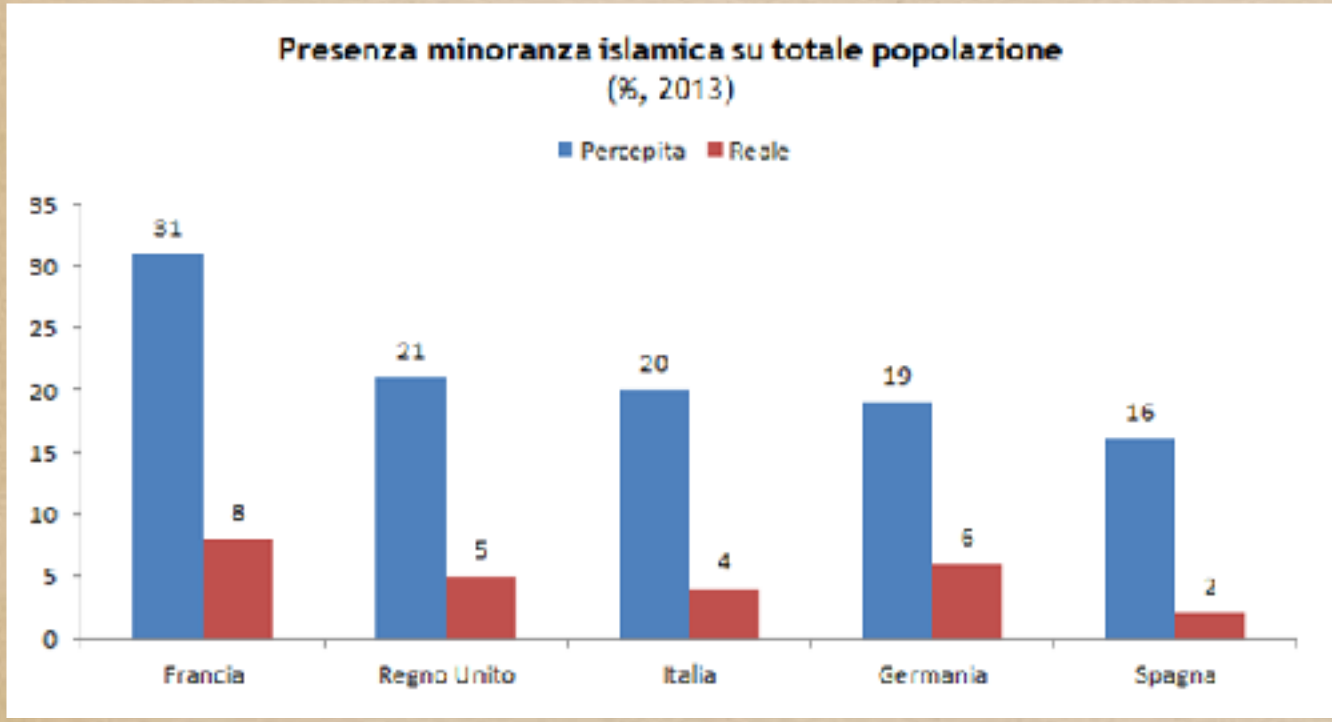
Altro argomento di grande rilevanza sociale. Come stanno veramente le cose?



Ancora una volta il problema esiste, ma appare alquanto diverso da come viene presentato. Un eccesso di letture ideologiche non favorisce l'analisi corretta.

Esempio: terrorismo

Oppure ancora, quanto è drammatica la situazione dovuta ai moderni terrorismi?



Qui è interessante vedere come la percezione pubblica del fenomeno sia slegata dalla realtà fattuale. Gli anni '70 ed '80 del secolo passato sono stati caratterizzati da un'intensità dei fenomeni terroristici molto superiore all'attuale.