Non sempre 2 è il doppio di 1

Una bugia detta con una statistica è fuorviante, ma non si può attribuirtene la colpa

Darrell Huff [1]

Paolo Ciafaloni *INFN & Dipartimento "Ennio De Giorgi" - Lecce*

La realtà ci parla anche attraverso i numeri, ma non sempre chi utilizza i numeri per dare maggiore autorevolezza alle tesi che propugna, lo fa in maniera corretta. In effetti l'utilizzo dei numeri si presta ad ogni forma di falsa interpretazione e mistificazione. In questo articolo cerco di dare qualche nozione di base sulla interpretazione dei numeri per aiutare a sviluppare il senso critico di chi legge.

Numeri e realtà

Interagiamo in molti modi con la realtà, utilizzando i nostri sensi. Certamente non tutte le interazioni necessitano di numeri: un pittore che dipinge un quadro, un musicista che suona uno strumento, un appassionato di mare che surfa le onde, sono esempi di interazioni che non coinvolgono i numeri. Ma a volte i numeri servono, quando ad esempio vogliamo sapere le dimensioni di un tavolo per sapere se entrerà in cucina, quando ci pesiamo, quando utilizziamo il denaro eccetera.

I numeri, diversamente dall'emozione che posso provare di fronte a un tramonto, hanno sempre bisogno di interpretazione. Il numero "500" acquista un significato solo se specifico, ad esempio, se si riferisce al prezzo in euro di un oggetto. Inoltre dovrei specificare di quale oggetto si tratta, e per sapere se posso permettermi di comprarlo lo devo confrontare con il mio conto in banca e il mio stipendio, e così via. In questo caso l'interpretazione che faccio del numero "500", ovvero l'elaborazione che faccio attraverso il cervello per dare un significato a questo numero, è piuttosto semplice.

Non sempre è così. Ad esempio, posso avere a disposizione una mole enorme di numeri, o dati, che riguardano pressione atmosferica, temperatura, venti di una gran quantità di località della terra in varie epoche, e tuttavia non capire nulla di cosa stia capitando al clima terrestre. In questo caso presumibilmente solo un climatologo è in grado di mettere ordine e dare una corretta interpretazione. Interpretazione che riguarda sia il passato, cioè come il clima si è evoluto nel tempo, sia il futuro, se il climatologo vuole predire l'evoluzione nei prossimi anni.

I numeri di cui parlo in questo articolo non sono di immediata interpretazione, ma non sono così complicati da richiedere un esperto per poter essere interpretati. Con qualche strumento matematico che qui introduco, tali numeri possono essere correttamente interpreta-

ti. Senza questi strumenti, l'interpretazione diventa completamente arbitraria e probabilmente sbagliata.

Parafrasando il divertente e istruttivo libretto di Darrell Huff [1], potremmo dire: "Datemi una qualsiasi tesi e troverò i dati che la confermano". I numeri in effetti sembrano dare un'aura di obiettività a quanto si asserisce, ma il loro significato può essere mistificato in mille modi: scegliendo le medie in maniera opportuna, ignorando alcuni insiemi di dati a favore di altri, ignorando il fatto che ogni numero andrebbe accompagnato dalla relativa incertezza, e così via.

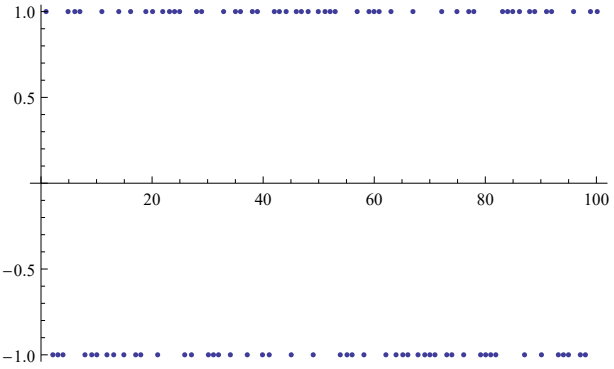

Figura 1: *Estrazione casuale di palline ripetuta 100 volte. '1' significa che è stata estratta una pallina bianca, mentre '-1' significa che è stata estratta una pallina nera.*

Palline nere e palline bianche

Piero ha davanti a sé una enorme scatola piena di palline bianche e nere, e si chiede quale sia la percentuale di palline bianche. Le palline sono troppo numerose per poterle contare tutte, diciamo un milione per fissare le idee; quindi si chiede in quale modo possa `stimare` questa percentuale.

Piero inizia estraendo una pallina; è bianca, può concludere che tutte le palline nella scatola sono bianche? Certamente no, può solo concludere che non sono tutte nere. In effetti, da una sola estrazione, si rende conto Piero, non può concludere nulla: le probabilità che sia bianca oppure nera sono sostanzialmente uguali. Intuitivamente, Piero capisce che può avere maggiori informazioni rimettendo dentro la pallina e procedendo a fare molte estrazioni successive, rimettendo dentro ogni volta la pallina estratta per non alterare le percentuali.

Nella figura 1 possiamo vedere il risultato di tale serie di estrazioni: adottiamo la convenzione che il numero 1 sta ad indicare l'estrazione di una pallina bianca estratta e il numero -1 l'estrazione di una pallina nera[1]. Piero decide di fare 100 estrazioni, e trova che è stata estratta 55 volte una pallina bianca e 45 volte una pallina nera.

A questo punto però a Piero sembra probabile che la percentuale di palline bianche sia del 55%, ma si rende conto che non può esserne sicuro: per esserlo, dovrebbe contare tutte le palline, che sono molte di più di 100. Allo stesso tempo capisce che è improbabile che la percentuale sia, mettiamo, 10%. Come può quantificare il livello di incertezza con cui conosce quello che gli interessa, cioè la percentuale di palline bianche?

Per rispondere a questa domanda occorre utilizzare un po' di matematica. Supponiamo per il momento di imbrogliare, e di sapere che nella scatola ci sono metà palline bianche e metà nere. Quando estraiamo una pallina, la probabilità che sia bianca è $1/2$, così come quella che sia nera. Possiamo definire una variabile x tale che $x = 1$ se la pallina estratta è bianca, e $x = -1$ se la pallina estratta è nera. Avremo quindi $P(x = 1) = P(x = -1) = 1/2$. Possiamo definire la media di questa variabile (E sta per *Expectation value*):

$$E[x] \equiv \overline{x} = P_1 \times 1 + P_{-1} \times (-1) = 0 \ . \qquad (1)$$

Più in generale definiamo la media di una variabile casuale, cioè che non ha un valore determinato ma può assumere un certo numero di valori $v_1, v_2 \cdots v_n$, come

$$\overline{x} = \sum_{i=1}^{n} P_i v_i \ , \qquad (2)$$

dove v_i sono i valori possibili che può assumere la variabile e P_i le relative probabilità. Che significato ha la media? Supponiamo di ripetere n estrazioni, rimettendo ogni volta la pallina estratta nella scatola; ogni estrazione assume un valore definito x_i.

Definiamo la `media misurata` come

$$\bar{x}^{(n)} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i \ . \qquad (3)$$

[1] L'estrazione casuale è stata in realtà simulata da me al computer

Ci aspettiamo che se facciamo un numero grande di estrazioni (n grande) la `media misurata` sia molto vicina al valore di aspettazione di x, che possiamo anche chiamare `media vera`. Ovvero, in termini un po' più formali:

$$\lim_{n\to\infty} \bar{x}^{(n)} = E[x] \quad . \tag{4}$$

Questa aspettativa, ovvero l'avvicinarsi della media misurata, anche detta media empirica o media campionaria, alla media vera con l'aumentare del numero di ripetizioni, è nota come "legge dei grandi numeri". Nella sua versione più formale, ovvero la (4), si chiama "Teorema di Bernoulli". Per quanto l'essenza della legge possa essere intuitivamente quasi ovvia, occorsero 20 anni al matematico Jacob Bernouili per svilupparne, nel 1713, una dimostrazione sufficientemente rigorosa[2].

Possiamo visualizzare la legge dei grandi numeri facendo una piccola elaborazione sui dati rappresentati in Fig. 1. Fissato il numero di ripetizioni n, effettiamo il calcolo della media misurata fino a quel momento, in base all'equazione (3). Ripetiamo il procedimento per $n+1$ ripetizioni e così via. Le medie misurate per n che va da 1 a 100 sono riportate in Fig. (2).

Visualizziamo proprio quello che ci aspettiamo in base alla legge dei grandi numeri. Per piccoli valori di n, diciamo da 1 a 20, la media misurata si discosta molto dal valore vero $E[x] = 0$; per grandi valori di n (diciamo fra 70 e 100) la media misurata è invece vicina al valore vero. Nel linguaggio statistico si dice che le fluttuazioni, cioè le variazioni casuali della media misurata, sono grandi (di ordine 1) per n piccoli ma sono piccole (di ordine 0.1) per n grandi. Ci aspettiamo che al crescere di n oltre 100 diventino ancora più piccole.

Medie e Varianze

Quello che abbiamo detto finora, anche se interessante, è qualitativo; vogliamo sapere adesso qualcosa in più dal punto di vista quantitativo. Piero non si accontenta dell'equazione (4). Egli vuole sapere: fissato n, *quanto* la media misurata è vicina a quella vera? Ovvero, quanto è buona la stima che possiamo fare di $E[x]$?

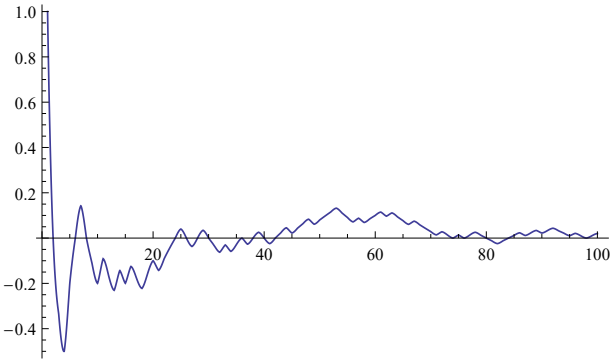

Figura 2: *Nel grafico è riportato, per ogni valore di n e in base ai dati di Fig. 1, il valore della media misurata, che è definita in Eq. (3). Le fluttuazioni attorno alla media vera (che è 0) diminuiscono in maniera evidente all'aumentare di n.*

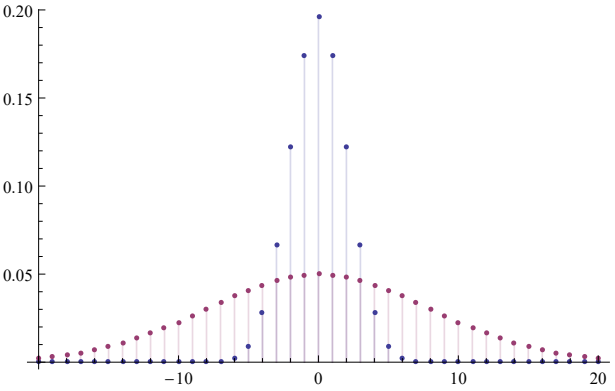

Figura 3: *Due distribuzioni di probabilità con la stessa media ($\mu = 0$) ma diversa deviazione standard. La distribuzione coi puntini blu ha deviazione standard $\sigma = 4$, mentre quella coi puntini rossi ha $\sigma = 16$.*

Per rispondere a questa domanda introduciamo la `deviazione standard`, detta anche `scarto quadratico medio` σ:

$$\sigma \equiv \sqrt{E[(x-\mu)^2]} = \sqrt{\sum_i P_i (x_i - \mu)^2} \quad . \tag{5}$$

La deviazione standard è una misura della probabilità che una certa quantità devìi dal valor medio μ: nella (5) vengono calcolate le deviazioni sia a destra che a sinistra (conta il modulo della differenza $x_i - \mu$) pesate dalle relative probabilità. La deviazione standard è quindi una misura di quanto sia probabile che una data variabile si discosti dal proprio valor medio. In Fig. 3 vediamo due distribuzioni con stessa media ($\mu = 0$ per entrambe) e diversa deviazione standard. È evidente che le probabilità sono significa-

tivamente diverse da 0 all'incirca in un intervallo $\mu - \sigma < x < \mu + \sigma$. La variabile relativa alla distribuzione con $\sigma = 4$ è quindi determinata con maggior accuratezza rispetto alla variabile relativa alla distribuzione con $\sigma = 16$. Idealmente vogliamo diminuire la deviazione standard per avere una stima più accurata. Procediamo quindi a calcolare la deviazione standard nel caso della prima estrazione di pallina (si ha $\mu = 0$ in questo caso):

$$\sigma^2 = E[x^2] = \sum P_i x_i^2 = \frac{1}{2}(1)^2 + \frac{1}{2}(-1)^2 = 1 \quad . \tag{6}$$

La deviazione standard in questo caso è quindi 1, un numero grande in quanto dello stesso ordine dei due possibili valori assunti dalla variabile. Ma andiamo avanti e calcoliamo la deviazione standard della media di due estrazioni, $\bar{x}^{(2)}$:

$$E[(\frac{x_1 + x_2}{2})^2] = \frac{1}{4}(E[x_1^2] + E[x_2^2] + 2E[x_1 x_2]) \quad . \tag{7}$$

Abbiamo già calcolato $E[x_1^2] = E[x_2^2] = 1$. Per quanto riguarda la variabile $x_1 x_2$, essa può assumere valore 1 con probabilità 1/2, e -1 con probabilità 1/2; si ha quindi $E[x_1 x_2] = 0$.

Questo è un risultato del tutto generale quando due variabili sono incorrelate, cioè quando il risultato dell'una non influenza il risultato dell'altra: il valore di aspettazione del loro prodotto è 0 (se la media non è nulla, ad essere 0 è $E[(x_1 - \mu)(x_2 - \mu)]$).

Siamo in grado ora di calcolare la deviazione standard, che è la radice quadrata dell'equazione (7): $\sigma(\bar{x}^{(2)}) = 1/\sqrt{2}$. Quindi la deviazione standard è diminuita rispetto al valore ($\sigma = 1$) della singola variabile. Cosa succede quando si calcolano le fluttuazioni della variabile `media dopo` n `estrazioni`

$$\bar{x}^{(n)} = \sum_i^n \frac{x_i}{n} \quad ?$$

Otteniamo

$$\begin{aligned} \sigma_n &= \sqrt{\frac{1}{n^2}(\sum_i^n E[x_i]))^2} \\ &= \sqrt{\frac{1}{n^2}\sum_i^n (E[x_i^2])} = \frac{1}{\sqrt{n}} \quad . \end{aligned} \tag{8}$$

avendo tenuto conto del fatto che, essendo le variabili incorrelate, $E[x_i x_j] = 0 \, \forall i, j$.

La (8) dà un significato quantitativo alla `legge dei grandi numeri`: non solo le fluttuazioni (deviazione standard) diminuiscono con il numero di ripetizioni, ma adesso sappiamo che diminuiscono come l'inverso della radice di n.

Possiamo fare un discorso analogo per il numero di palline bianche estratte, $b^{(n)}$ (nel nostro esempio, $b^{(100)} = 55$). La variabile $\bar{x}^{(n)}$ conta le palline bianche con peso 1, e quelle nere con peso -1; inoltre il numero delle palline nere è $n - b^{(n)}$, per cui:

$$\bar{x}^{(n)} = 2\frac{b^{(n)}}{n} - 1; \qquad b^{(n)} = (\bar{x}^{(n)} + 1)\frac{n}{2} \quad . \tag{9}$$

A una variazione $\Delta\bar{x}^{(n)}$ della media corrisponde quindi una variazione delle palline bianche $\Delta b^{(n)} = \Delta\bar{x}^{(n)}\frac{n}{2}$, per cui abbiamo

$$\sigma_b = \frac{n}{2}\sigma_n = \frac{1}{\sqrt{n}}\frac{n}{2} = \frac{\sqrt{n}}{2}. \tag{10}$$

In definitiva, le fluttuazioni della nostra variabile $\bar{x}^{(n)}$ scalano con n come $1/\sqrt{n}$ mentre le fluttuazioni del numero di palline bianche scalano come $\sqrt{n}$.

Piero può quindi dare un significato più preciso alla sua stima del numero di palline bianche. In effetti Piero può prendere come misura dell'incertezza nel numero di palline bianche il valore di σ_b, quindi ritenere ragionevolmente che la media vera sia compresa fra il valore $\mu - \sigma_b$ e il valore $\mu + \sigma_b$. Avendo effettuato 100 ripetizioni ed essendo quindi $\sigma_b = \sqrt{n}/2 = 5$, Piero conclude che la sua stima della frazione di palline bianche è:

$$\frac{55 \pm 5}{100} = (55 \pm 5)\% \Leftrightarrow E[x] = \mu = 0.1 \pm 0.1 \quad . \tag{11}$$

Limite Centrale

Che significato ha l'equazione (11)? Più precisamente, con quale probabilità la stima $50\% \leq \mu \leq 60\%$ contiene la media vera, sconosciuta a Piero? Per rispondere a questa domanda quantitativa è utile introdurre un'approssimazione che utilizza una funzione dovuta al matematico e fisico Carl Friederich Gauss [3], scritta in termini dei

parametri μ, σ:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad . \tag{12}$$

Questa distribuzione suppone l'uso di una variabile x continua, ed è quindi diversa dalle distribuzioni discrete finora discusse, in cui si assegna la probabilità P_i che una data variabile assuma un certo valore x_i (vedi l'equazione (6)). Per le distribuzioni continue di probabilità, il significato è che $G(a)da$ è la probabilità che la variabile x assuma un valore compreso fra a e $a + da$. Nel caso della Gaussiana otteniamo i valori di media e deviazione standard:

$$E[x] = \int xG(x)dx = \mu; \quad \sqrt{E[(x-\mu)^2]} = \sigma \tag{13}$$

L'importanza della gaussiana come distribuzione di probabilità è data dal `Teorema del limite centrale` [4], che possiamo enunciare nel modo seguente:

Una variabile casuale che sia somma di n variabili casuali, purché n sia sufficientemente grande, ha una distribuzione di tipo gaussiano indipendentemente dalle distribuzioni delle n variabili.

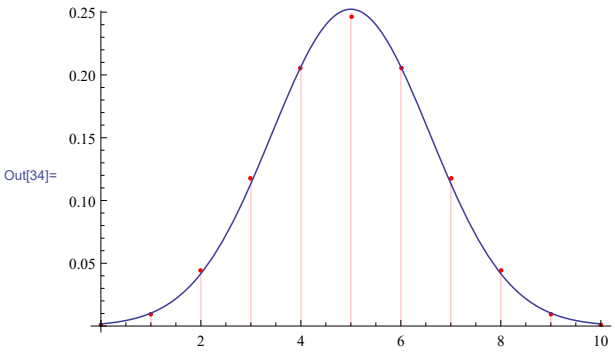

Figura 4: *Il teorema del limite centrale: distribuzione della variabile $\bar{x}^{(n)}$ con n=10 (puntini rossi) e distribuzione gaussiana (linea blu).*

Questo teorema è piuttosto sorprendente, in quanto vale qualsiasi siano le distribuzioni delle n variabili che concorrono a formare la variabile finale; questo significa che la distribuzione gaussiana è presente in moltissimi casi in cui l'effetto finale sia la somma di tanti piccoli effetti.

Ad esempio, la nostra variable $\bar{x}^{(n)}$ definita dall'equazione (3) è la somma di tante variabili x_i, e quindi possiamo ritenere che la sua distribuzione sia gaussiana con $\mu = 0$ e $\sigma = 1/\sqrt{n}$. Ma quando possiamo dire che n è **sufficientemente grande**?

In figura 4 appare il confronto fra la distribuzione di palline bianche con $n = 10$ (puntini rossi) e la distribuzione di Gauss; come si vede l'approssimazione data dalla gaussiana è buona, e ci si aspetta che sia ancora migliore per il nostro $n = 100$.

Come usiamo la distribuzione gaussiana per dare un significato quantitativo alla stima di Piero, Eq. (11)? Supponiamo di ripetere il nostro esperimento di estrazione di 100 palline per 1,2,....N volte. Quante volte il risultato che otteniamo per la media cade in intervallo di $\pm 5\%$ intorno alla media vera? Ovvero, quanto volte la nostra stima $50\% < \mu < 60\%$ contiene la media vera? Rispondere a questa domanda è piuttosto semplice: tale probabilità P è data dall'integrale

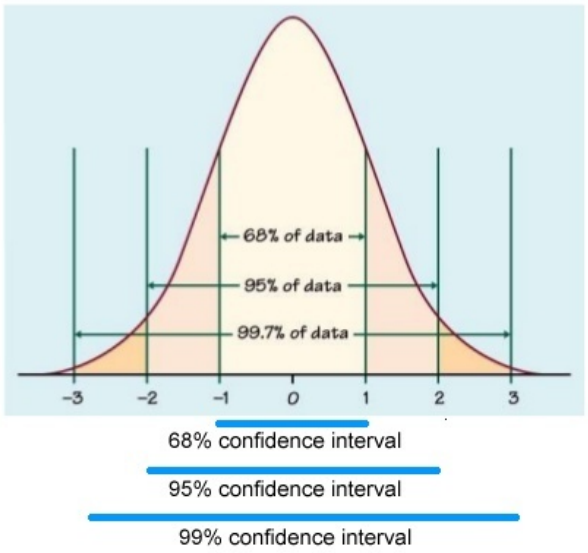

Figura 5: *Gli intervalli di confidenza sono connessi all'area sotto la curva gaussiana: da -σ a σ tale area vale circa 0.68, da -2 σ a 2 σ vale circa 0.95, da -3 σ a 3 σ vale circa 0.99*

fra 50 e 60 della Gaussiana con media 50 e deviazione standard $\frac{\sqrt{n}}{2} = 5$, e vale circa 0.68, e se N è molto grande il numero di volte in cui la media misurata μ sta nell'intervallo prescelto è $PN = 0.68N$.

Piero sintetizza questa informazione dicendo che:

$$\frac{n_b}{n} = 55\% \pm 5\% \qquad C.L. = 68\% \ , \tag{14}$$

dove $C.L.$ sta per confidence level, ovvero inter-

vallo di confidenza, ed ha il significato discusso sopra.

In genere si scelgono intervalli più ampi di intervallo di confidenza, come 2σ (=10 nel nostro caso) ad esempio:

$$\frac{n_b}{n} = 55\% \pm 10\% \qquad C.L. = 95\% \qquad (15)$$

in quanto l'integrale della distribuzione gaussiana fra $\mu - 2\sigma$ e $\mu + 2\sigma$ vale circa 0.95.

Le equazioni (14) e (15) sono infine ciò che Piero cercava. Esse quantificano, una volta scelto, un certo intervallo di confidenza, l'intervallo in cui riteniamo stia la media vera.

Queste sono informazioni di tipo probabilistico, e non potrebbe essere altrimenti dato che stiamo effettuando una stima. Ad esempio, se scelgo 68% C.L., esiste una probabilità del 32 % che la media vera **non** stia nell'intervallo definito dall'equazione (14). Posso diminuire tale probabilità fino al 5 %, ma l'intervallo della mia stima diventa più ampio (vedi l'equazione (15)).Per finire, abbiamo fin qui assunto di conoscere a priori la deviazione standard, ma in molti casi non è così. Si pensi ad esempio al caso in cui vogliamo stimare l'altezza delle donne adulte italiane ad esempio: non possiamo sapere quale sia la deviazione standard della vera distribuzione. Possiamo però stimare dal nostro campione la deviazione standard utilizzando l'espressione usuale:

$$\sigma^2 = \sum_i P_i (x_i - \mu)^2 \ , \qquad (16)$$

in cui μ è stimato dal campione nel modo usuale e P_i, invece che essere ricavato da una distribuzione nota a priori, è stimato in base alle frequenze dei valori: $P_i = n_i/n$ dove n_i è il numero di volte che ricorre un dato valore di altezza e n il numero totale. Ad esempio, nel caso di Piero $n_{bianche} = 55, n_{nere} = 45, n = 100$ per cui per la singola variabile x_i abbiamo $P_1 = .55, P_{-1} = .45$, da cui stimiamo media e varianza:

$$\mu = 0.1$$
$$\sigma^2 = (1-\mu)^2 0.55 + (-1-\mu)^2 0.45 = 0.99 \ .$$

In base all'equazione (10), questo si ripercuote in una fluttuazione del numero di palline bianche pari a

$$\sigma_b = \sqrt{0.99}\frac{\sqrt{n}}{2} \approx 4.97 \ ,$$

un po' più piccola (ma non molto) della varianza vera pari a 5.

La stima di Piero per la media delle palline bianche a $2\sigma_b \approx 9.9\%$ diventa quindi:

$$\frac{n_b}{n} = 55\% \pm 9.9\% \qquad 95\% C.L. \qquad (17)$$

È bene rimarcare che il valore vero del numero di palline bianche può essere al di fuori dell'intervallo $45.1\% < \frac{n_b}{n} < 64.9\%$; la probabilità che questo accada è però piccola, circa del 5%.

Andamento col tempo

Spesso quello che ci interessa non è una fotografia della situazione, come nel caso del numero di palline bianche in una scatola, bensì capire come evolve nel tempo una data quantità. Ad esempio, potremmo essere interessati a come cambiano da 100 anni fa ad oggi l'altezza media e l'aspettativa di vita in Italia, o l'andamento negli ultimi sei mesi del numero di contagiati di `Covid-19`.

Va subito detto che nel fare questo dovremo necessariamente tener conto delle fluttuazioni statistiche (deviazione standard). Ad esempio, supponiamo che ci siano circa 10.000 tamponi positivi in un dato giorno durante la pandemia di `Covid-19`. Come abbiamo visto, le fluttuazioni sono dell'ordine di $\sqrt{10.000} = 100$. Queste fluttuazioni, che possiamo chiamare statistiche, non sono però le uniche. Se confrontiamo i numero di contagiati in un giorno con quelli del giorno successivo, ci saranno molte altre cause di fluttuazione: in un dato giorno ad esempio si sono effettuati molti più tamponi del giorno precedente. Oppure in un dato giorno si sono fatti più tamponi in una regione maggiormente colpita, e ancora si sono fatti più tamponi a una data fascia di età (gli anziani ad esempio, maggiormente a rischio). Per non parlare del fatto che si sono fatti tamponi solo a chi lo chiede oppure a chi mostra sintomi, e non sappiamo nulla di tutti coloro che sono positivi ma non mostrano sintomi; e così via. In definitiva, le fluttazioni, cioè le variazioni nel conteggio dei positivi che nulla hanno a che vedere con il vero numero di positivi, saranno numericamente molto più grandi di 100.

Questo significa che se un giorno si misurano 10.000 nuovi positivi e il giorno dopo 11.000, non è corretto dire che i positivi sono aumenta-

ti. Quello che stiamo osservando sono piuttosto fluttuazioni non significative intorno alla media 10.500. In generale, possiamo dire che non è sensato dare i numeri, cioè analizzare le variazioni del numero di positivi da un giorno all'altro, cosa che peraltro i media fanno di continuo e non solo per il `Covid-19`. Più sensato sarebbe diminuire il peso delle fluttuazioni considerando un numero più elevato di tamponi. Si potreppe valutare la differenza del numero di tamponi positivi da una settimana all'altra ad esempio, anziché da un giorno all'altro.

Una prima informazione che vogliamo estrarre (tenendo conto delle fluttuazioni) è se la quantità in oggetto sia sostanzialmente invariata con il tempo, oppure se aumenti (o diminuisca). Dopodiché può interessare con con quale velocità la quantità stia aumentando, cioè quale tipo di funzione descriva meglio l'andamento temporale. Spesso, almeno nel caso delle epidemie, si usano approssimazioni di tipo lineare oppure esponenziale. Una data quantità segue una curva lineare se la crescita (o decrescita) da un giorno all'altro è `costante`. Se invece la crescita (o decrescita) giornaliera è una `percentuale costante` del valore del giorno precedente, allora l'andamento è detto esponenziale.

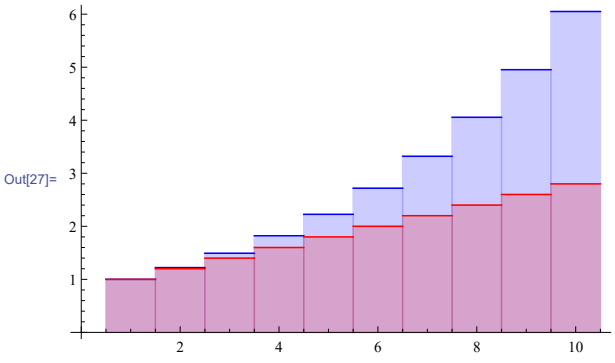

Figura 6: *Confronto fra andamento lineare (in rosso) ed esponenziale (in blu). Nel caso lineare, la quantità varia di 0.2 ogni giorno, per un totale di 10 giorni. Per l'esponenziale l'aumento è del 20% ogni giorno.*

In figura 6 sono riportati i grafici di un andamento lineare (in rosso) e uno esponenziale (in blu). Il valore iniziale della variabile è 1; nel caso lineare tale variabile cresce di 0.2 ogni giorno mentre nel caso esponenziale cresce del 20% ogni giorno. Inizialmente le due curve sono vicine ed è difficile distinguerle, ma poi l'esponenziale cresce molto di più in quanto la quantità di cui aumenta ogni giorno non è costante bensì aumenta a sua volta col tempo. Da queste curve dovrebbe apparire chiara l'assurdità di certe conclusioni che si sentono trarre: se una quantità triplica da un giorno all'altro non ha nessun senso dire che "è aumentata esponenzialmente". Per capire se una crescita è di tipo esponenziale oppure no occorre seguire l'andamento per diversi giorni.

Supponiamo di fare, ad esempio, l'ipotesi che una data quantità x sia costante nel tempo. Naturalmente, a causa delle fluttuazioni, questa quantità non assumerà esattamente lo stesso valore ogni giorno. Come possiamo fare per verificare la nostra ipotesi quindi? Per prima cosa determiniamo il valore della costante, chiamiamola k, in modo da minimizzare la somma degli scarti quadratici relativi ad ogni valore x_i corrispondente al giorno i, con n giorni totali:

$$\frac{\partial}{\partial k}\sum_{i=1}^{n}(x_i-k)^2 = -2\sum_{i=1}^{n}x_i + 2kn = 0$$

$$\Rightarrow \quad k = \frac{1}{n}\sum_{i=1}^{n}x_i \quad . \tag{18}$$

Determinare la costante k (che poi coincide, e non è una sorpresa, con la usuale media μ) ammonta a fare un *fit* dei dati con una funzione costante; il metodo qui utilizzato si chiama `metodo dei minimi quadrati`. Ma come possiamo giudicare la bontà del *fit*?

Definiamo per prima cosa una variabile che quantifichi le fluttuazioni, sommando gli scostamenti dalla media:

$$\chi^2 \equiv \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\mu} \quad . \tag{19}$$

Gli scostamenti quadratici vengono qui divisi per la media in quanto ci si attende che $\sigma_i^2 \sim \mu$; ogni punto x_i contribuisce circa 1 al χ^2. Se le distribuzioni degli x_i sono di tipo gaussiano allora la distribuzione del χ^2 è nota [6].

In maniera analoga a quanto visto in precedenza, possiamo stabilire un $C.L.$. Se il χ^2 viene costruito a partire da n variabili indipendenti $x_1 ... x_n$, ci aspettiamo che il suo valore sia vicino a n. Se invece il valore del χ^2 è molto più grande di n, allora diciamo che la qualità del *fit* è cattiva e la nostra ipotesi che, ad esempio, i dati vengano

descritti da una funzione costante, è sbagliata.

Supponiamo di avere ad esempio 4 variabili indipendenti. Allora, dall'integrale della distribuzione del χ^2 [5] otteniamo che la probabilità di avere un valore maggiore o uguale a 7.78 è del 90 %. Quindi, con il $C.L.$ del 90%, se il valore che otteniamo dai nostri dati è maggiore di 7.78 rigettiamo l'ipotesi che i dati siano costanti nel tempo.

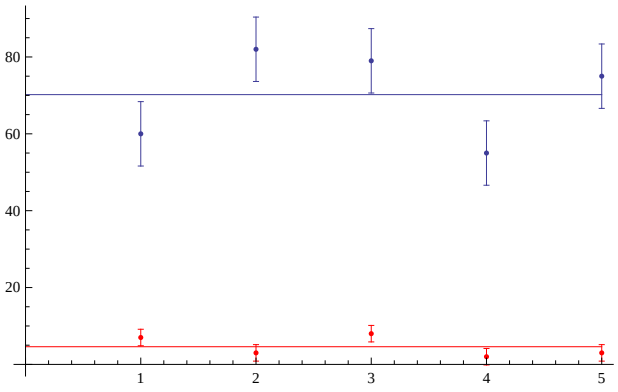

Figura 7: *Due insiemi di dati al variare del giorno (da 1 a 5)*

In figura 7 sono rappresentati, in rosso e in blu, due insiemi di dati insieme alle rispettive medie, in un arco temporale di 5 giorni. Nella figura sono rappresentate anche le incertezze, o fluttuazioni ($\pm 1\sigma$), come segmenti verticali per ogni dato. In entrambi i casi il numero di variabili indipendenti (o gradi di libertà) è 4: infatti, una volta fissata la media si ha un vincolo che permette di ricavare uno dei dati (ad esempio quello relativo al quinto giorno) dagli altri.

Il valore di χ^2 che si ottiene per i dati in rosso è 6.35, per quelli in blu è 8.19. Siccome il valore critico al livello di confidenza del 90% per 4 gradi di libertà è 7.78, rigettiamo l'ipotesi che i dati siano descritti da una costante per il caso blu mentre la accettiamo per il caso rosso. Notare che nel caso dei dati in rosso, il valore numerico è 3 nel giorno 2 e 8 nel giorno 3: malgrado questo i dati sono ben descritti da una costante. Questo fa capire l'importanza di tener conto delle fluttuazioni nell'analizzare i dati e capire gli andamenti.

In maniera analoga a quanto fatto finora per la funzione costante, si possono fare *fit* dei dati con qualsiasi tipo di funzione, come ad esempio funzioni lineari (retta), esponenziale, polinomi ecc. Cruciale è l'importanza del numero di gradi di libertà (o variabili indipendenti).

È evidente che dato un insieme qualsiasi di punti posso sempre tracciare una curva che passi per ognuno di essi, ma in questo caso il *fit* è privo di significato. Ad esempio, se ho solo due dati posso sempre fare un *fit* con una retta, ma questo non mi dice nulla nè sull'andamento dei dati nè sulle possibili estrapolazioni (previsioni).

Supponiamo di fare un *fit* con un polinomio di grado n; la funzione in questo caso è definita da $n+1$ parametri. Il numero di gradi di libertà rilevante in questo caso è

gradi di libertà = numero di variabili - (n+1).

Se faccio un fit di n punti con un polinomio di grado $n-1$, il numero di gradi di libertà è 0: questo corrisponde al fatto che posso sempre determinare i parametri della mia funzione in modo che passi per gli n punti. In questo caso il test del χ^2 non è significativo. Per aver un buon *fit*, oltre ad avere un valore basso di χ^2, la mia curva deve essere definita da un numero limitato di parametri.

Conclusioni

Se si dovesse estrapolare un unico messaggio da questo articolo sarebbe questo: **qualsiasi numero, per essere correttamente interpretato, va accompagnato dalla relativa incertezza**. Quindi, ha poco significato dire ad esempio "il numero attuale di contagiati da Covid-19 in Italia è 100.000", mentre ha senso dire "il numero di contagiati è compreso fra 93.000 e 105.000 con una probabilità (livello di confidenza) del 90%".

Oltre a questa consapevolezza, dovremmo sempre tener presente che se si fa un conteggio, le fluttuazioni statistiche sono più grandi nel caso di numeri piccoli. Per cui, non è sempre vero che 2 è il doppio di 1: i due numeri sono di fatto equivalenti una volta tenuto conto delle fluttuazioni. Quello di cui non ho parlato qui invece è di come la scelta di quello che si chiama un campione non significativo possa portare a false interpretazioni e mistificazioni: invito alla lettura di [1] per questo tema.

[1] Darrell Huff: *How to Lie with Statistics*, Norton, New York (1954).

[2] https://en.wikipedia.org/wiki/Law_of_large_numbers

[3] https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

[4] https://en.wikipedia.org/wiki/Central_limit_theorem

[5] https://math-wiki.com/images/5/5a/Chi-table.pdf

[6] https://en.wikipedia.org/wiki/Chi-square_distribution#Probability_density_function

Paolo Ciafaloni: è dipendente dell'Istituto di Fisica Nucleare (INFN). Svolge attività di ricerca nel campo della fisica teorica delle particelle e della Cosmologia, e insegna teoria delle interazioni forti.