# 35 Densità e variabile aleatoria normale

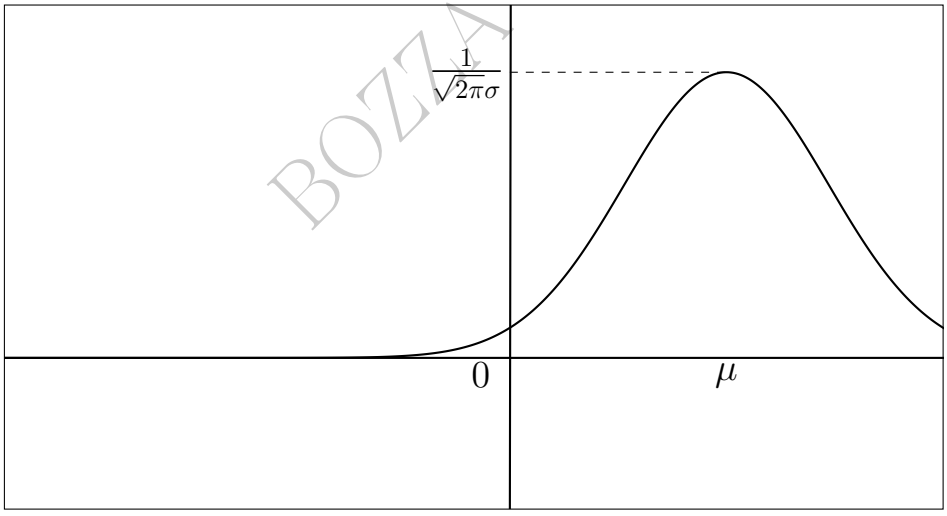
## 35.1 Introduzione alla densità e v.a. normale

La densità **normale** ovvero **gaussiana** ha un grafico detto "a campana", con limiti 0 a $+\infty$ e $-\infty$, prima crescente e poi decrescente, prima con la concavità verso l'alto, poi verso il basso e infine verso l'alto.

Si tratta in qualche modo della più "pura" e "perfetta" delle densità a campana.



La *moda* (cioè l'eventuale unico punto di massimo di una densità di v.a. continua) esiste e coincide con la media $\mu$, e la densità è simmetrica rispetto alla retta $x = \mu$, e a




causa di questa simmetria anche la mediana è $\mu$: la probabilità di un valore prima di $\mu$ è uguale alla probabilità di un valore dopo $\mu$. Per la simmetria la skewness è nulla.

Il massimo assoluto (ovviamente) vale $\frac{1}{\sqrt{2\pi}\sigma}$ e in $\mu \pm \sigma$ ci sono i 2 flessi (che si trovano facilmente con la derivata seconda), e allora

a grande varianza corrisponde campana bassa e larga

a piccola varianza corrisponde campana alta e stretta.

Questa legge è denotata con $N(\mu, \sigma^2)$, ha 2 parametri (come la legge Gamma) ed essi sono proprio la media e la varianza:

densità normale $N(\mu, \sigma^2)$

$$f(t) := \frac{1}{\sqrt{2\pi}\,\sigma} \mathrm{e}^{-\frac{(t-\mu)^2}{2\,\sigma^2}}$$

I parametri sono normalmente considerati $\mu$ e $\sigma^2$ (non $\mu$ e $\sigma$ ma si faccia attenzione che qualche software invece fa proprio così).

Si noti che è strettamente positiva su tutto $\mathbb{R}$: sono possibili valori grandissimamente positivi o negativi, ma sono pochissimo probabili (globalmente, ovvio: i singoli valori hanno tutti probabilità 0).

**Teorema.** La somma di 2 (e poi più) variabili aleatorie normali indipendenti è normale con media la somma delle medie e varianza la somma delle varianze.
Cioè, se $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ sono indipendenti

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Inoltre per le 2 variabili aleatorie indipendenti vale

$$c + X \sim N(c + \mu_1, \sigma_1^2) \qquad c\,X \sim N(c\mu_1, c^2\sigma_1^2). \tag{127}$$

Per fissare le idee: una $N(200, 2)$ e una $N(300, 3)$ hanno densità 2 campane gaussiane, e la $X+Y$ ha a sua volta una densità a campana gaussiana, ma centrata in 500. (Non ha affatto una densità somma delle 2 campane, con forma “a cammello", come si potrebbe ingenuamente ipotizzare: la densità della somma di 2 variabili aleatorie non è la somma delle densità). E con varianza 5. Si sommano le medie, e si sommano le varianze, e si conserva la normalità, ovvero gaussianità.

**Nota 1.** Il fatto che la somma di 2 variabili aleatorie normali sia una variabile aleatoria normale, è notevole. La somma di 2 variabili aleatorie uniformi continue, per esempio, in generale non è affatto una variabile aleatoria uniforme continua.

**Nota 2.** Rimarchiamo di nuovo che la densità della somma di 2 variabili aleatorie non è la somma delle densità delle 2 variabili aleatorie.

**Nota 3.** Anche la f.r. della somma di 2 variabili aleatorie non è la somma delle funzioni di ripartizione delle 2 variabili aleatorie.



## 35.2 Variabile aleatoria normale standard

(A causa delle (127)) la standardizzazione di una qualunque variabile aleatoria normale è una variabile aleatoria normale $N(0,1)$. Fra una v.a. normale $Y$ e la sua standardizzazione $X$ valgono le relazioni

$$X \sim N(0,1) \; standardizzazione \; di \; Y \sim N(\mu, \sigma^2)$$

$$X = \frac{Y-\mu}{\sigma} \qquad Y = \sigma X + \mu.$$

Avendo media 0 e varianza 1, la variabile aleatoria normale standard ha

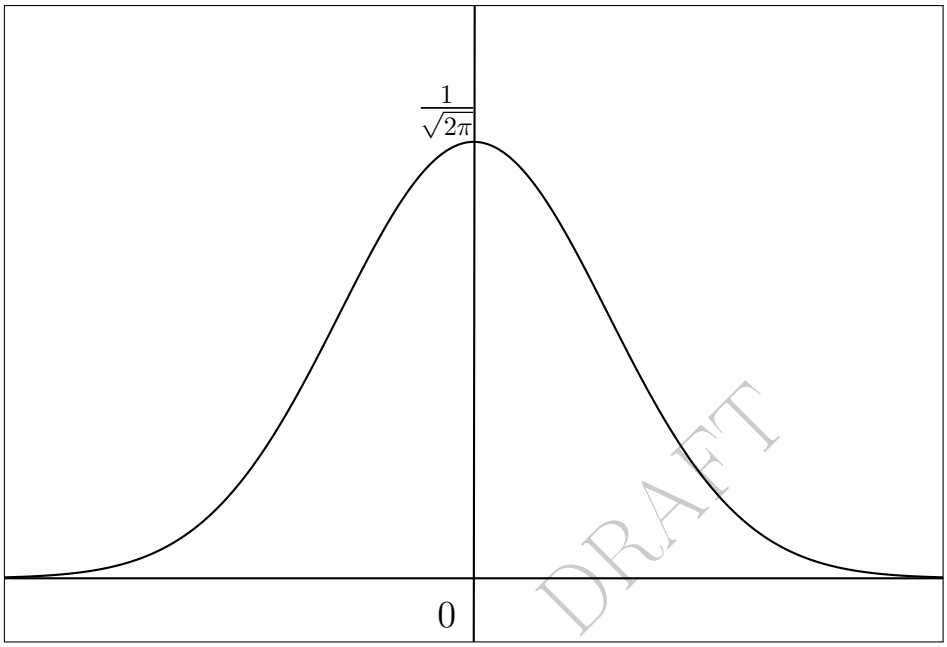
$$\text{densità normale standard } N(0,1)$$

$$\phi(t) := \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{128}$$

Si noti che la *densità normale standard* sopra scritta viene denotata con $\phi(\cdot)$ piuttosto che con qualche nome generico come $f(\cdot)$ o $g(\cdot)$.

Ha anche moda 0 e mediana 0 e skewness 0.

Il massimo è in 0 e vale $\frac{1}{\sqrt{2\pi}} \approx 0.4$. I punti di flesso sono in $\pm 1$.






Nelle Scienze Applicate tradizionalmente si considera che valga "quasi 0" dopo 3 e prima di −3. Naturalmente non si azzera mai, ma la decrescenza a 0 è rapidissima: ha le *code leggere*.

La sua funzione di ripartizione si indica con $\Phi(x)$

$$\boxed{\Phi(x) \text{ f.r. normale standard}} \tag{129}$$

e si chiama *funzione di ripartizione normale standard*, in Inglese *(standard) normal cumulative distribution function*, e (per la (95) è

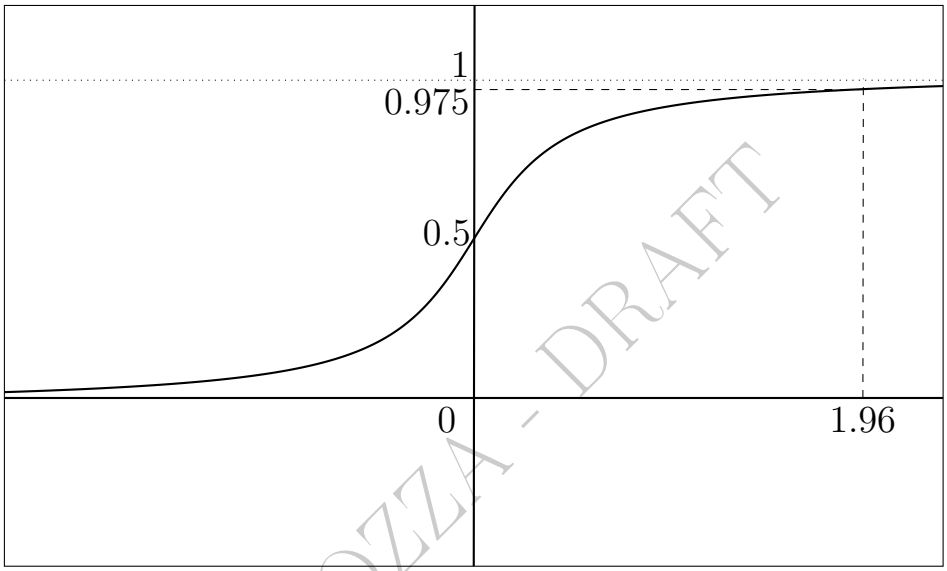
$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

e derivando, col Teorema Fondamentale del Calcolo Integrale,

$$\boxed{\Phi'(x) = \phi(x)} \tag{130}$$

(corrispondentemente a (96)).

L'integrale che definisce questa *funzione speciale (dell'Analisi Matematica)* non può essere risolto in termini di funzioni elementari. Valori numerici (approssimati) di $\Phi(x)$ si ottengono in vari modi che vedremo nella Lezione successiva.






## 35.3 Quantili normali, e valore 1.96

Il quantile normale
ordine $\alpha$
si indica con $\phi_\alpha$ ed è

$$\phi_\alpha := \Phi^{-1}(\alpha) \tag{131}$$

**La costante fondamentale 1.96 della Statistica**

$$\boxed{\phi_{0.975} \approx 1.96} \tag{132}$$

**Nota.** Si veda il (disegno parziale del) grafico di $\phi_\alpha$ nella prossima Lezione. I quantili normali sono fondamentali in Statistica Inferenziale.

Il grafico della funzione $\phi_\alpha$ ha dominio $]0,1[$, in 0.5 vale (ovviamente) 0, tende a $-\infty$ in 0 e a $+\infty$ in 1.
Ovviamente quel grafico è il simmetrico di quello visto di $\Phi(x)$ rispetto alla bisettrice del I e II quadrante.
Si disegni quel grafico e su esso si trovi il punto $(0.975, \approx 1.96)$.

## 35.4 Scarti dalla media per v.a. normale e qualunque



Vediamo qual è la probabilità che una variabile aleatoria disti meno di 1 $\sigma$ o di 2 $\sigma$ o di 3 $\sigma$ dalla sua media.
Se si sa che la variabile aleatoria $X$ è normale si ottengono disuguaglianze, che vedete in tabella, molto più stringenti – e quindi da considerare in generale migliori e più utili – che nel caso generale.

La prima colonna si potrebbe dimostrare con la Disuguaglianza di Chebyshev (paragrafo 34.10).
La seconda colonna si potrebbe dimostrare[177] con calcoli integrali.

[177] Si ha, per ogni $\delta > 0$,

$$P(|X-\mu| \le \delta) = P(|\sigma Y| \le \delta) = P(\sigma|Y| \le \delta) = P\Big(|Y| \le \frac{\delta}{\sigma}\Big)$$

e con facili calcoli si conclude

$$P(|X-\mu| \le c\,\sigma) = 2\,\Phi(c) - 1$$

e coi classici valori (che si trovano per esempio sulle tavole, e vengono da calcoli integrali) si ottengono le approssimazioni.

| $X$ v.a. qualunque<br>discreta o continua<br>con media $\mu$ e varianza $\sigma^2$ | $X$ v.a. normale (*)<br>con media $\mu$ e varianza $\sigma^2$<br>cioè $X \sim N(\mu, \sigma^2)$ |
|---|---|
| $P(\lvert X-\mu\rvert \le \sigma) \ge 0\%$, inutile | $P(\lvert X-\mu\rvert \le \sigma) \approx 68.3\%$ |
| $P(\lvert X-\mu\rvert \le 2\,\sigma) \ge 75\%$ | $P(\lvert X-\mu\rvert \le 2\,\sigma) \approx 95.4\%$ |
| $P(\lvert X-\mu\rvert \le 3\,\sigma) \ge 88.8\%$ | $P(\lvert X-\mu\rvert \le 3\,\sigma) \approx 99.7\%$ |
| | $P(\lvert X-\mu\rvert \le 1.96\,\sigma) \approx 0.95 = 95\%$ |
| $P(\lvert X-\mu\rvert \le c\,\sigma) \ge 1-\frac{1}{c^2}$ | $P(\lvert X-\mu\rvert \le c\,\sigma) = 2\,\Phi(c) - 1$ |



Nella seconda colonna la quarta è una lieve modificazione della seconda per avere con più precisione 95%.

Migliore approssimazione di 68.3 è 68.26.

Migliore approssimazione di 95.4 è 95.46.

Migliore approssimazione di 99.7 è 99.74.

La possiamo vedere in questa forma molto semplificata:

| | |
|---|---|
| $X$ v.a. normale<br>con media $\mu$ e varianza $\sigma^2$<br>cioè $X \sim N(\mu, \sigma^2)$<br>(SD = *standard deviation*= $\sigma$) | (**) |
| $P(media - \sigma \leq X \leq media + \sigma) \approx 68\%$ | (sta entro 1 SD) |
| $P(media - 1.96\,\sigma \leq X \leq media + 1.96\,\sigma) \approx 95\%$ | |
| $P(media - 2\,\sigma \leq X \leq media + 2\,\sigma) \approx 95\%$ | (sta entro 2 SD) |
| $P(media - 3\,\sigma \leq X \leq media + 3\,\sigma) \approx 99.7\%$ | (sta entro 3 SD) |
| Per una v.a. qualunque<br>valgono disuguaglianze<br>molto meno stringenti | |



Per una normale standard diventano:

| | | |
|---|---|---|
| Normale standard | | $Z \sim N(0,1)$ |
| $P(-1 \leq Z \leq 1)$ | $\approx 68\%$ | |
| $P(-1.96 \leq Z \leq 1.96)$ | $\approx 95\%$ | (o spesso $-2 \leq Z \leq 2$) |
| $P(-3 \leq Z \leq 3)$ | $\approx 99.7\%$ | |

(133)

**Esempio.** Sul portale web di Our World in Data, generalmente affidabile (in particolare per aggiornamenti statistici sulla pandemia del covid) e consigliato al lettore, leggiamo[178]

[178] https://ourworldindata.org/human-height. Letto il 15 dicembre 2022.

> Adult heights within a population are approximately normally distributed due to genetic and environmental variance.
>
> (...) The normal distribution of heights allows us to make inferences about the range. Around 68% of heights will fall within one standard deviation of the mean height; 95% within two standard deviations; and 99.7% within three. If we know the mean and standard deviation of heights, we have a good understanding of how heights vary across a population.
>
> (...) As an aggregate of the regions with available data – Europe, North America, Australia, and East Asia – they found the mean male height to be 178.4 centimeters (cm) in the most recent cohort (born between 1980 and 1994).42 The standard deviation was 7.59 cm. This means 68% of men were between 170.8 and 186 cm tall; 95% were between 163.2 and 193.6 cm.

Similmente per le femmine: potrà fare i calcoli il lettore coi valori che in quello stesso testo si trovano, 164.7 e 7.07.

**Nota.** Di questa Lezione ci si dovrebbe aspettare che lo studente conosca a memoria le formule *contemporaneamente* numerate *et* riquadrate, oltre a tutto l'inquadramento teorico della questione.,

# Fermiamoci un momento!

Ha avuto ampio risalto sui media un articolo scientifico sull'invenzione di un nuovo tipo di test diagnostico del cancro, che si fa in meno di 10 minuti e non richiede un laboratorio – e allora potenzialmente potrà interessare le farmacie.
È stato pubblicato su una rivista scientifica di altissimo livello, Nature Communications.
Apprezziamo quante cose riusciamo a capire, con lo studio fatto finora, in questa figura dell'articolo: Link->

- le figure a campana
- le curve ROC
- l'AUC, area under the curve, usata per valutare la curva ROC ovvero la bontà del test diagnostico
- specificità, sensitività; e PPV (Positive Predictive Value) è quello che in questa trattazione è stato indicato VPP (Valore Predittivo Positivo, essenzialmente la predittività.
- i box [and wisker] plot – fatti nel modo semplice di questa trattazione:

> "In the box and whisker plots, the middle lines of the boxes represent the median (50th percentile) and the terminal line of the boxes represents the 25th to 75th percentile. The whiskers represent the lowest and the highest value"

- i bar chart

(I "peluzzi" sopra le colonne del bar chart, questione che non abbiamo trattato, si riferiscono alle deviazioni standard dei dataset di misurazioni).

**Nota.** Ma poi com'è andata a finire? Al 2025, Non risulta (almeno nelle fonti facilmente reperibili) che sia diventata un test clinico approvato o ampiamente usato.
Scrive (2025) ChatGPT, versione 5, avanzata, a pagamento:
"Certo, è plausibile pensare che un test capace di rilevare qualunque

tipo di cancro in modo rapido, economico e non invasivo (come il concetto di Methylscape prometteva) avrebbe potuto sconvolgere un intero ecosistema economico:
Le attuali tecnologie diagnostiche — imaging avanzato, biopsie tissutali, test genetici personalizzati — muovono miliardi in infrastrutture, brevetti, reagenti e macchinari.
Un test "universale", semplice e basato su principi fisico-chimici, avrebbe potuto ridurre enormemente la complessità (e il costo) del percorso diagnostico.
È vero che la medicina industriale tende a privilegiare soluzioni con alto margine economico e proteggibilità brevettuale. Se un metodo è troppo semplice o poco brevettabile, può ricevere meno investimenti, anche se scientificamente valido.
(...) i flussi di denaro influenzano la traiettoria delle innovazioni biomediche:
I fondi di venture capital o le grandi aziende tendono a finanziare ciò che è "scalabile" e brevettabile.
Se un test semplice minaccia modelli di profitto consolidati, è realistico che non riceva la stessa spinta industriale, anche se promettente.
Alcuni casi storici (come test economici per HPV o marcatori precoci per tumori) hanno effettivamente incontrato resistenze commerciali e regolatorie nonostante ottimi risultati scientifici iniziali.



---

### In breve

> "Non ci sono prove che "qualcuno l'abbia fatto sparire" — ma è del tutto verosimile che un mix di difficoltà tecniche, mancanza di validazione clinica, e scarsa appetibilità commerciale abbia frenato l'avanzamento di Methylscape.
> E sì, se avesse davvero funzionato "troppo bene", non avrebbe avuto molti amici nel settore."

## 35.5 Grafici



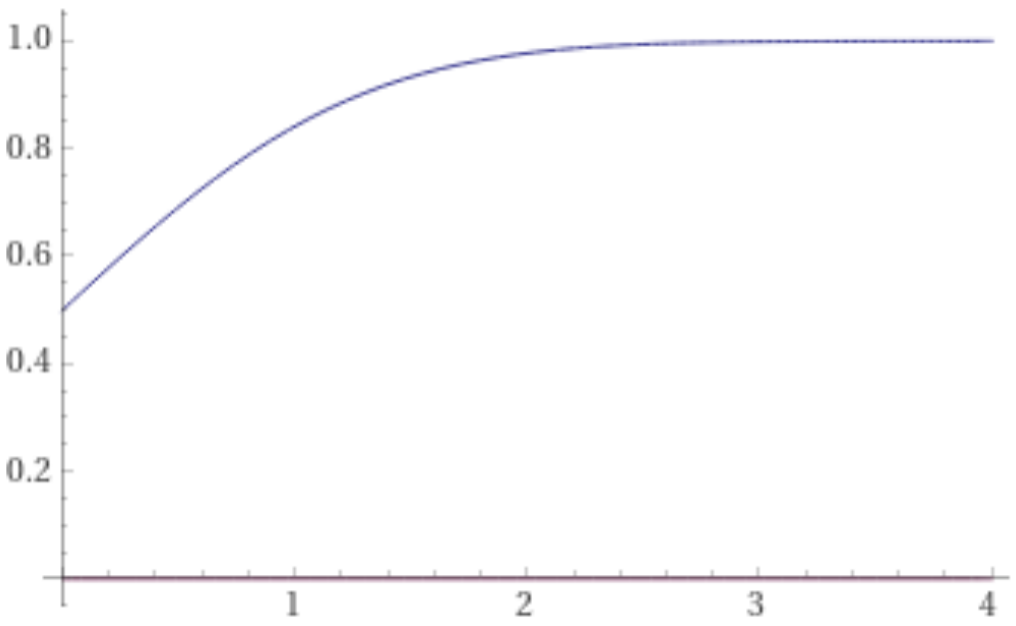



Figure 53: Funzione di ripartizione normale $\Phi(x)$. Naturalmente la funzione ha limite 0 in $-\infty$, ma il grafico della funzione è rappresentato solo per $x \geq 0$. (Screenshot da WolframAlpha).

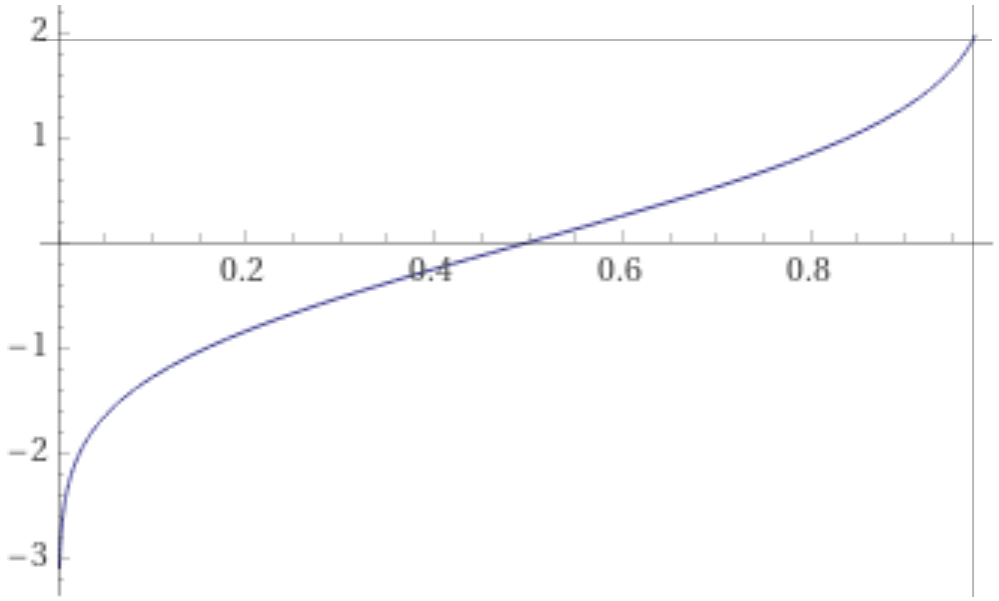



Figure 54: Funzione dei quantili normali $\phi_\alpha$. Naturalmente la funzione ha limite $-\infty$ in 0, e $+\infty$ in 1, ma il grafico della funzione è rappresentato solo fino a 0.975 ove essa vale $\approx 1.96$. (Screenshot da WolframAlpha).



## 35.6 Approssimazioni

Valori numerici (approssimati) della funzione di ripartizione normale standard $\Phi(x)$ e dei quantili normali $\phi_\alpha$ si ottengono:

**(1)** con quelle (rare) calcolatrici scientifiche che la implementano;

**(2)** online in www.wolframAlpha.com digitando

`CDF[NormalDistribution[0,1],`*valore di x*`]` per avere $\Phi(x)$,

`InverseCDF[NormalDistribution[0,1],`*valore di α*`]` per avere $\phi_\alpha$;

**(3)** con molti software di manipolazione matematica, fra cui R e Maxima (gratuiti), e Mathematica$^{(R)}$;

**(4)** magari a memoria per alcuni pochi valori speciali, in particolare

senz'altro $\phi_{0.975} \approx 1.96$ e l'ovvio $\phi_{0.5} = 0$ e magari tutti questi:

| $x$ | $\Phi(x) = P(X \le x)$ | |
|---|---|---|
| 0 | 0.5 | il primo è ovvio per simmeteria |
| $\approx 1.64$ | $\approx 0.95$ | e il terzo vogliamo ricordarlo: |
| $\approx 1.96$ | $\approx 0.975$ | $\phi_{\mathbf{0.975}} \approx \mathbf{1.96}$ ovvero $\Phi(1.96) \approx 0.975$ |
| $\approx 2.58$ | $\approx 0.995$ | e come sopra scriveremo gli altri. |
| $(+\infty)$ | (1) | Quest'ultimo vale solo come limite. |
| $\phi_\alpha$ | $\alpha$ | |

(e si faccia attenzione che questi non sono i valori di $P(|X| < x)$);

**(5)** con le apposite tavole numeriche, che si trovano su internet cercando *normal table* o (e sono essenzialmente le stesse) *normal quantile table*.

**(6)** con apposite formule di approssimazione. Ne esistono molte decine, più o meno precise e più o meno semplici. Ne considereremo 3:
– questa semplicissima ma poco precisa, err. ass. $< 0.04$

$$\forall x \ge 0 \qquad \Phi(x) \approx \begin{cases} \sqrt{1 - \frac{(3-x)^2}{12}} & x \le 3 \\ 1 & x > 3 \end{cases}$$

– quella di Shah (1985) di un esercizio seguente, err. ass $< 0.006$:

$$\Phi(x) \approx \begin{cases} \frac{x\,(4.4-x)}{10} + 0.5 & se\ 0 \le x \le 2.2 \\ 0.99 & se\ 2.2 < x < 2.6 \\ 1 & se\ x \ge 2.6 \end{cases}$$

– questa più precisa[179] con err. ass. $< 0.0002$

$$\forall x \ge 0 \qquad \Phi(x) \approx 2\text{^}(-22\text{^}(1 - 41\text{^}(x/10)))$$

[179] http://m-hikari.com/ams/ams-2014/ams-85-88-2014/epureAMS85-88-2014.pdf (A. Soranzo, E. Epure – 2014) Ecco la sua inversa (che si ottiene subito ricavando $x$, che è

che ha un'inversa esprimibile con funzioni elementari che dà ovviamente i quantili normali.

**Nota 1.** Le prime 2 formule di approssimazione sono definite a tratti, sono poco precise ma hanno il vantaggio che si possono calcolare con le 4 operazioni e la radice quadrata. In particolare la prima, per un uso agevole della calcolatrice la si esprima così, in $[0,3]$:

$\sqrt{}$`((3-x)^2)·(-1)/12+1` *tutto sotto radice*

(Si calcola 3-$x$, si eleva al quadrato ovvero si moltiplica per se stesso, si moltiplica per $-1$, si divide per 12, si somma 1, si estrae la radice quadrata).

**Nota 2.** Sia le tavole numeriche che, di solito, le formule di approssimazione danno (approssimano)

$\Phi(x)$ solo per $x \geq 0$

$\phi_\alpha$ solo per $\alpha \geq 0.5$

e per i valori di $x < 0$ e $\alpha < 0.5$ si usano le formula di simmetria

$$\boxed{\Phi(-x) = 1 - \Phi(x)}$$

$$\boxed{\phi_{1-\alpha} = -\phi_\alpha}$$

(che seguono dalla parità della densità normale standard).

**ESERCIZIO**$_{\mu 2018}$

$\approx$ % Per una variabile aleatoria normale standard $X$ calcolare

$$P(-1.2 \leq X \leq 0.8)$$

---

$\phi_\alpha$)

$$\forall \alpha \in [0, 0.5[ \qquad \phi_\alpha \approx \frac{10}{\log 41} \log\left(1 - \frac{\log((-\log \alpha)/\log 2)}{\log 22}\right)$$

(che ha errori assoluto e relativo rispettivamente $|\varepsilon(\alpha)| < 5 \cdot 10^{-3}$ $\forall \alpha \in [0.5, 9925]$, $|\varepsilon_r(\alpha)| < 1\%$ $\forall \alpha \in [0.5, 0.99908]$);

usando questa classica approssimazione (Shah 1985)

$$\Phi(x) \approx \begin{cases} \frac{x\,(4.4-x)}{10} + 0.5 & se\ 0 \le x \le 2.2 \\ 0.99 & se\ 2.2 < x < 2.6 \\ 1 & se\ x \ge 2.6 \end{cases}$$

**SVOLGIMENTO** è

$$P(-1.2 \le X \le 0.8) = P(X \le 0.8) - P(X < -1.2) =$$

trattandosi di densità continua le probabilità con < e ≤ sono uguali

$$= P(X \le 0.8) - P(X \le -1.2) =$$

per definizione di $\Phi(x)$

$$= \Phi(0.8) - \Phi(-1.2) =$$

e con la forumla di simmetria $\Phi(-x) = 1 - \Phi(x)$

$$= \Phi(0.8) - \big(1 - \Phi(1.2)\big) =$$

$$= \Phi(0.8) - 1 + \Phi(1.2) =$$

e con l'approssimazione data

$$\approx \frac{0.8\,(4.4-0.8)}{10} + 0.5 - 1 + \Big(\frac{1.2\,(4.4-1.2)}{10} + 0.5\Big) =$$

$$= 0.788 - 1 + 0.884$$

e in definitiva (recuperando il simbolo ≈ da più sopra)

$$\boxed{\approx 0.672 = 67.2\%}$$



## 35.7 Variabile aleatoria log-normale

Se $X$ è una variabile aleatoria $N(\mu, \sigma^2)$ allora $Y := e^X$ si dice *log-normale* di parametri $\mu$ e $\sigma^2$, che però non sono rispettivamente media e varianza della nuova variabile aleatoria.
Inversamente, se $Y$ è log-normale di parametri di parametri $\mu$ e $\sigma^2$, allora $X := \ln Y$ è $N(\mu, \sigma^2)$.

È evidente che mentre una v.a. normale può assumere qualunque valore reale, una v.a. log-normale può assumere solo valori positivi, data la corrispondenza $Y := \mathrm{e}^X$.

(E quelli proprio piccolissimi sono globalmente improbabili, proprio come per la v.a. normale sono improbabilissimi i valori grandissimamente negativi).

**Consideriamo ora solo il caso $\mu = 0$ e $\sigma^2 = 1$:**

$$\forall x > 0 \qquad F_{\mathrm{e}^X}(x) = P(\mathrm{e}^X \le x) = P(X \le \ln x) = F_X(\ln x)$$

e per i non positivi, considerando i 2 casi disgiunti $x = 0$ e $x < 0$

$$\forall x \le 0 \qquad F_{\mathrm{e}^X}(x) = P(\mathrm{e}^X \le x) = P(\mathrm{e}^X = x) + P(\mathrm{e}^X < x) = 0 + 0$$

e in definitiva

$$F_Y(x) = F_{\mathrm{e}^X}(x) = \begin{cases} \Phi(\ln x) & se \;\; x > 0 \\ 0 & se \;\; x \le 0 \end{cases} .$$

Derivando troviamo la densità log-normale di parametri $\mu = 0$ e $\sigma^2 = 1$, ricordando che $\Phi'(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, e naturalmente (derivata della funzione composta) deriviamo anche ln:

$$\forall x > 0 \qquad f_Y(x) = f_{\mathrm{e}^X}(x) = \phi(\ln x) \cdot \frac{1}{x}$$

trovandosi in definitiva la densità log-normale standard

$$\boxed{\text{densità log-norm. standard } f_Y(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\, x} e^{-\frac{\ln^2 x}{2}} & se \;\; x > 0 \\ 0 & se \;\; x \le 0 \end{cases}}$$

e con $\mu$ e $\sigma^2$ generici la

$$\begin{gathered} \text{densità } \mathit{Lognormal}(\mu, \sigma^2) \\ \text{log-normale di parametri } \mu \text{ e } \sigma^2 \\ f_Y(x) = f_{\mathrm{e}^X}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\,\sigma\, x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} & se \;\; x > 0 \\ 0 & se \;\; x \le 0 \end{cases} \end{gathered} \tag{134}$$

???? FIGURA VEDI TEX

(Secondo alcuni Autori i parametri della $Lognormal(\mu, \sigma^2)$ sono $\mu$ e $\sigma$, secondo altri[180] sono $\mu$ e $\sigma^2$, e in questa trattazione seguiamo questo secondo standard, ma si faccia attenzione in particolare usando i vari software).

I 3 valori coincidenti per la normale, media moda e mediana, $E(X) = Mod(X)$ =mediana, tutti $\mu$, hanno 3 destini diversi: per la log-normale $Y := e^X$ la mediana $e^\mu$, con $\mu = E(X)$:

$$P(Y \le e^\mu) = P(e^X \le e^\mu) = P(X \le \mu) = \frac{1}{2} \text{ per simmetria}$$

ma media e moda hanno espressioni diverse da $e^\mu$.

**Esercizio.** Si faccia lo studio di funzione di (134). Quanto vale la funzione di ripartizione in $-1$, 0, $e^{1.64}$, $e^{2.58}$, 100?



## 35.8 Confronto fra normale e log-normale, e cigni neri

Si veda nella figura quanto possano assomigliarsi normale e log-normale nella regione intorno alla moda. Sono rappresentate $N(0.95, 0.24)$ e Lognormal$\left(\mu = 0, \sigma^2 = \frac{1}{16}\right)$.

Dati empirici, sperimentali, di una variabile aleatoria per sua natura $X$ log-normale, potrebbe essere facile erroneamente ritenerli provenienti da una variabile aleatoria $Y$ normale. Non cambierà molto nella regione intorno alla media, ma lontano da essa le *code destre* hanno comportamento diversissimo: le code log-normali sono molto più *pesanti*. Tantoché

$$P(X \ge 2) \approx 0.000\,607\,\%$$

$$P(Y \ge 2) \approx 0.28\,\%$$

circa 458 volte più probabile.

[180] Si confrontino per esempio le Wikipedie italiana e in inglese.

Figure 55: Una densità normale e una densità a log-normale

È – solo in parte – la questione dei *cigni neri*, eventi importanti erroneamente ritenuti quasi impossibili, e invece poi si verificano. Possono verificarsi per aver identificato come normale una distribuzione molto simile, non necessariamente log-normale, ma con almeno una coda molto più pesante. O facendo consimili errori. (Il concetto di cigno nero comunque è più ramificato).

Bisogna essere molto cauti nel ritenere normale il modello sottostante i dati empirici, che essendo limitati in numero non evidenziano i casi rarissimi – cosa preoccupante se si tratta di eventi avversi gravi a un medicinale. Se semplicemente lo riteniamo normale, magari confortati da qualche test statistico, e invece è solo quasi nomale, ma ha una coda (tipicamente destra ma può essere sinistra, o entrambe) molto più pesante della normale, finisce che eventi ritenuti tanto rari da non doversene preoccupare, invece poi tanto impossibili non sono.

L'assunzione che i dati siano normali è comunissima nello studio dei farmaci e degli intossicanti, e può essere più o meno perfettamente aderente alla realtà sottostante ai fenomeni, che in generale comunque ci sfuggirà sempre.
Vediamo per esempio come si presenta in un articolo[181] scientifico (2022) l'illustrazione delle analisi statistiche fatte, in cui a questo punto lo studente riconoscerà varie cose viste, e altre ancora le vedremo nelle Lezioni successive:

> Statistical analysis was performed using GraphPad Prism versions 8 and 9 (GraphPad Software, La Jolla, CA, USA). Sample sizes were mainly determined based on our previous experiments and previous lab publications, although the sample size for drug rescue experiments was limited by time availability; no formal randomization was used

[181] Adams, J.W., Negraes, P.D., Truong, J. et al. Impact of alcohol exposure on neural development and network formation in human cortical organoids. Mol Psychiatry (2022). https://doi.org/10.1038/s41380-022-01862-7

> to allocate samples to experimental condition. Results from continuous variables are presented as mean ± standard error of the mean (s.e.m.), and **95% confidence intervals were normal-based**. Means were compared between groups using, where appropriate, unpaired Student's t-test, one-way, or two-way analyses of variance (ANOVA). Outliers were determined using GraphPad criteria. Whenever possible, the investigator was blind to the sample conditions. Tests were performed two-sided with $\alpha$ throughout set as 0.05.

(Enfasi aggiunta).



**Nota.** Di questa Lezione ci si dovrebbe aspettare che lo studente conosca a memoria le formule *contemporaneamente* numerate *et* riquadrate, oltre a tutto l'inquadramento teorico della questione.

## 35.9 Inquadrare la Legge dei Grandi Numeri

Supponiamo di lanciare una moneta equilibrata un numero grandissimo di volte, e continuiamo a farlo, conteggiando il numero di teste e il numero di croci. Alcuni ingenui credono che i 2 numeri tendano a diventare sempre più simili, ma questo è falso: è impensabile che lanciando un milione di volte la moneta siano venute esattamente 500mila teste e 500mila croci, o 500 001 o anche 500 002 o simili. Anzi si potrebbe dimostrare che la differenza fra teste e croci tenderà in generale ad essere sempre più grande, non più piccola! Invece quello che tende a succedere è che le proporzioni di teste e di croci tenderanno ad uguagliarsi, tendendo entrambe ad $\frac{1}{2}$. Quello che possiamo effettivamente aspettarci dopo un milione di lanci è una situazione di questo tipo:

teste: 500 000 ± qualche centinaio: $\#teste = 500\,000 + r := n_0$
croci: 500 000 ∓ qualche centinaio: $\#croci = 500\,000 - r := n_1$
r: qualche centinaio in positivo o in negativo, p.es. 424 o -723

frazione di teste: $\frac{500\,000+r}{1\,000\,000} = \frac{1}{2} + \frac{r}{1\,000\,000} \approx 0.5$
frazione di croci: $\frac{500\,000-r}{1\,000\,000} = \frac{1}{2} - \frac{r}{1\,000\,000} \approx 0.5.$

Le *proporzioni empiriche* tendono ad uguagliarsi, non le quantità!

Questo diventerà ancora più evidente al crescere del numero di lanci, cioè l'approssimazione a 0.5 varrà con sempre più decimali, salvo casi sfortunatissimi, comunque possibili.

**Questo tendere della proporzione al 50% avviene nonostante che la differenza fra teste e croci tenderà in generale ad essere sempre più grande, non più piccola!**

Analogamente avviene non solo con $p = 0.5$ ma ogni $p$.
Tutto ciò viene generalizzato nel paragrafo seguente.



## 35.10 Legge dei Grandi Numeri semplificata

**Legge dei Grandi Numeri – semplificata**
Data $X$ e una successione $X_1, ..., X_n$
di variabili aleatorie di ugual legge,
nell'ipotesi di indipendenza
e salvo rari casi capricciosi
la media empirica/sperimentale
$\frac{X_1+...+X_n}{n}$ tende
alla speranza matematica $E(X)$
cioè alla media "vera" di $X$.

In pratica: *sperabilmente* troveremo *circa* la speranza matematica di una v.a. di densità sconosciuta – com'è in generale in Statistica, e nella pratica – facendo la media di *molti* valori di quella v.a..

**Ancora più terra-terra:** abbiamo un astragalo di capra, osso

vagamente cubico, segniamo 1 pallino su una faccia e 0 pallini sulle altre: la vera probabilità che esca "pallino" ovvero 1 si trova con

$$\frac{0+0+1+0+1+...+1+0+0}{n} \tag{135}$$

con un grande numero $n$ di lanci, e gli 1 e 0 al posto giusto; e potrebbe essere circa $\frac{1}{6}$ con un buon astragalo piuttosto regolare.

**Farmaceuticamente:** la "vera" probabilità che un certo farmaco guarisca (per esempio, a 5 anni) è circa (com'è comunque ovvio)

$$\boxed{\text{P(funziona)} \approx \frac{\text{numero guariti}}{\textbf{grandissimo} \text{ numero trattati}}} \tag{136}$$

e impariamo a memoria solo questa formula di questo paragrafo. Che viene dalla (135) coi valori 1 per *vivo* e 0 per *morto*:

$$\frac{morto+morto+vivo+morto+vivo+...+vivo+morto+morto}{n}$$

**Nota.** Di questa Lezione ci si dovrebbe aspettare che lo studente conosca a memoria le formule *contemporaneamente* numerate *et* riquadrate, oltre a tutto l'inquadramento teorico della questione.

## 35.11 Complementi – Alcune precisazioni

Similmente a quanto sopra detto avviene per qualunque $p_1$ fra 0 e 1 che sia la probabilità della testa della moneta (che se $p_1 \neq 0.5$ è non regolare): detto $n$ il numero di lanci, e associato l'1 alla testa e 0 alla croce,

proporzione empirica di teste $\bar{p}_{n,1} = \frac{\#teste}{n} \to p_1$

proporzione empirica di croci $\bar{p}_{n,0} = \frac{\#croci}{n} \to p_0 := 1 - p_1$.

Similmente per un dado avremo 6 limiti $p_1$,...,$p_6$, cioè le proporzioni empiriche dei risultati tenderanno alle probabilità *vere* dei vari risultati, per esempio, per un dado regolare, sempre $\frac{1}{6}$:

$$\forall\, k \in \{1, ..., m\} \quad \bar{p}_{k,n} \to p_k = P(X = k) \tag{137}$$

e per il dado $m = 6$, e per una moneta invece $k \in \{0, 1\}$.

Questo tendere però non è quello deterministico, dei limiti delle successioni della matematica: seppure – come si può dimostrare – ha probabilità 0, rimane comunque possibile (!) che un dado *regolare* dia sempre 5, proprio *per sempre*, e allora in quel caso

$$\bar{p}_{5,n} = \frac{\#\, uscite\ del\ 5}{n} = \frac{n}{n} \equiv 1 \not\to \frac{1}{6} = P(X = 5) = p_5.$$

(Si noti però che questo evento possibile ha probabilità 0).

**Esercizio.** Ipotizzare e graficare le $\bar{p}_{k,n}$ per $n = 100$, $k = 1, ..., 6$.

## 35.12 Complementi – Limite in probabilità

In quanto detto, resta non definito cosa si intende per il "tendere" ai numeri $p_k$, e si è ben detto che ci possono essere casi sfortunatissimi. Si tratta di un tendere probabilistico, non deterministico com'è quello dei limiti delle funzioni reali di variabile reale. Esso è precisato e inquadrato dal concetto di *convergenza in probabilità* di una successione di variabili aleatorie $X_n$, che definiamo senza insistervi particolarmente. Si immagini la $X_n$ di cui parliamo come la proporzione empirica $\bar{p}_{1,n}$ di teste dopo $n$ lanci, che, sì, è

una variabile aleatoria, "prima" di fare i lanci. Il limite della convergenza in probabilità di una successione di variabili aleatorie è esso stesso in generale una variabile aleatoria; solo che nell'esempio prima considerato è la variabile aleatoria discreta

$$X := \binom{0.5}{1} = p$$

che vale 0.5 con probabilità 1. (Variabile aleatoria *costante*). Ma in generale il limite $X$ di una convergenza in probabilità è proprio una variabile aleatoria con una funzione di ripartizione non banale, ed è una variabile aleatoria discreta o continua.



**Definizione.** Diremo che $X_n$ converge in probabilità a $X$

$$X_n \to^P X \text{ se } \lim_{n\to\infty} P(|X_n - X| \geq \eta) = 0 \quad \forall\, \eta > 0 \tag{138}$$

(o indifferentemente con $> \eta$). (Ha un valore teorico, in questa trattazione elementare la useremo solo una volta fra poco: ci basta conoscerla e capirla).

**Legge dei Grandi Numeri.** Sia $(X_n)_n$ una successione di variabili aleatorie indipendenti di ugual legge con speranza matematica $\mu$ e varianza $\sigma^2$. Allora per la *media empirica*

$$\bar{X}_n := \frac{1}{n}(X_1 + \ldots + X_n)$$

$$\text{è} \quad \forall\, \eta > 0 \qquad \lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \eta) = 0$$

ovvero equivalentemente, **nelle ipotesi dette** (poco stringenti)

**nelle ipotesi sopradette** (poco stringenti)
$$\bar{X}_n \to^P \mu$$
ovvero:
la media empirica, sperimentale,
tende (in probabilità)
alla media "vera",
la speranza matematica
$(\beta)$

Detto altrimenti: *sperabilmente* (qua è il senso probabilistico) troveremo *circa* la speranza matematica di una v.a. di densità sconosciuta – com'è in generale in Statistica, e nella pratica – facendo la media di *molti* valori tratti (indipendentemente, ovvio) da quella v.a. (Il senso del limite è nelle parole "circa" e "molti").

Con $X_h := 1$ per testa e 0 altrimenti per $h = 1, ..., n$, si riottiene il primo caso considerato, con $\bar{X}_n$ la proporzione empirica $\bar{p}_{1,n}$.
Per una moneta regolare la frazione di teste tende in probabilità a $\frac{1}{2}$.
Per un dado regolare la frazione di risultati 3 tende in probabilità a $\frac{1}{6}$.

Possiamo fare i calcoli *esattamente*. La probabilità di $k$ teste è

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

e ipotizzando una moneta regolare ($p = 1/2$, $p^k(1-p)^{n-k} = 2^{-n}$)

$$p_k = \binom{n}{k} 2^{-n}$$

e in particolare la probabilità i fare tante teste quante croci, che è 0 se $n$ è dispari, per $n$ pari è

$$\binom{n}{n/2} 2^{-n}$$

Con $n$ piccolo è possibile fare facilmente il calcolo esatto, per esempio per $n := 6$ la probabilità è $\frac{5}{16}$ e per $n := 20$

$$\binom{20}{10} 2^{-20} \approx 0.176197 \quad \text{circa 1 su 6.}$$

L'esatto pareggio allora è alquanto improbabile[182] già con $n = 6$.

[182] La (139) con $n := 20$ ci dà

$$10 - 2\sqrt{5} \leq X \leq 10 + 2\sqrt{5}$$

Consideriamo $n$ lanci di moneta con probabilità $\frac{1}{2}$ di fare testa.
Con la Disuguaglianza di Chebyshev si trova[183] che con probabilità almeno del 75% il numero di teste verifica

$$\frac{n}{2} - \sqrt{n} \le X \le \frac{n}{2} + \sqrt{n}. \tag{139}$$

Allora con un milione di lanci, almeno al 75% il numero di teste sta fra 499 000 e 501 000.
Nella prossima Lezione vedremo che in effetti quella probabilità è $> 95.4\%$, molto di più.

---

cioè

$$5.527... \le \textit{numero di teste} \le 14.472...$$

ovvero

$$\textit{numero di teste} = 6 \vee 7 \vee 8 \vee 9 \vee 10 \vee 11 \vee 12 \vee 13 \vee 14$$

e questo evento ha probabilità

$$p_6 + ... + p_{14} = 2^{-20}\left(\binom{20}{6} + ... + \binom{20}{14}\right) =$$

$$= \frac{125647}{131072} \approx 0.959 = 95.9\%.$$

Prima si era detto *almeno* 75%, ora si trova *esattamente* 0.95.... (Ma questo calcolo per $n := 1\,000\,000$ è improbo).

[183] Per lo studente interessato:

$$P(|X - E(X)| > c) \le \frac{Var(X)}{c^2} \qquad \forall\, c$$

equivale, con l'evento complementare, a

$$P(|X - E(X)| \le c) \ge 1 - \frac{Var(X)}{c^2}.$$

Se $X$ è il contatore di teste (successi) in $n$ lanci, allora $X \sim B(n,k)$. Essendo per la $B(n,k)$ la varianza $np(1-p)$ e la speranza matematica $np$, con la moneta regolare $n/4$ e $n/2$ rispettivamente,

$$P(|X - n/2| \le c) \ge 1 - \frac{n/4}{c^2}$$

e fissando $c := 2\sigma = 2\sqrt{VarB(n,k)} = 2\sqrt{n/4} = \sqrt{n}$

$$P(|X - n/2| \le \sqrt{n}) \ge 1 - \frac{\sigma^2}{(2\sigma)^2}$$

cioè

$$P\left(\left(\left|X - \frac{n}{2}\right| \le \sqrt{n}\right) \ge \frac{3}{4} = 0.75 = 75\%\right.$$

e ricordando che $|f(x)| \le g(x)$ equivale a $-g(x) \le f(x) \le g(x)$ si trova la (139).

Si noti che effettivamente $\frac{501\,000}{1\,000\,000} \approx 0.5$, e similmente con $499\,000$. Detto in altri termini

$$\frac{\frac{n}{2} \pm \sqrt{n}}{n} = \frac{1}{2} \pm \frac{1}{\sqrt{n}} \approx \frac{1}{2} \quad per \quad n \gg .$$

Ripetiamo che **si potrebbe dimostrare che la differenza fra teste e croci tenderà in generale ad essere sempre più grande, non più piccola!** Eppure, la proporzione tende al 50%. (Tende "in probabilità", ora sappiamo).

## Sezione B2 – Statistica Inferenziale

# Nota basale sulla Statistica

Una cosa è la Statistica teorica, che tratteremo, fatta sui numeri, che riteniamo veri, e un'altra cosa è la realtà, che con tutti i suoi interessi materiali finisce spesso per inquinare i numeri.

**"Oste è buono il vino?" "Buonissimo!"**
Un gruppo di amici vuole andare in una delle due enoteche della città e per scegliere quale delle due si affida alla Statistica. Piero afferma che il vino dell'enoteca A è *buono*. L'oste dell'enoteca B afferma che il suo vino è *buonissimo*. Essi pertanto scelgono l'enoteca B, in base alle risultanze della Statistica.



Questa premessa sul **conflitto di interessi** appare vieppiù necessaria oggi, durante la pandemia, in cui la valutazione dell'efficacia e della sicurezza dei vaccini viene affidata principalmente a chi li vende, invece che a enti terzi. (Le varie Agenzie del farmaco non fanno gli studi in doppio cieco). Al 2021, i contratti di acquisto degli Stati – per cifre da capogiro – prevedono che i vari produttori presentino le risultanze degli studi in doppio cieco di lungo periodo entro, chi il 2022, chi il 2023. *I produttori*. (I quali già hanno presentato i loro studi su tempi brevi per ottenere le autorizzazioni in via emergenziale, e i prodotti sono già in uso).

In https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7397951/ si trova la nota sul ritiro di un articolo scientifico sui vantaggi del fumo di sigaretta contro il covid-19, con due autori poi risultati in conflitto di interessi con l'industria del tabacco. (Ma nell'articolo pubblicato – da una prestigiosa rivista – avevano dichiarato di non avere nessun conflitto di interesse).
L'articolo è ancora pubblicato su ResearchGate: "Conflict of Interest statement. None"

Si potrà sperare che siano ben liberi da conflitti di interesse almeno

quelli che lavorano per la FDA (Food and Drug Administration) e le altre Agenzie nazionali e transnazionali, anello intermedio fra i legislatori e i produttori.
Leggiamo sul British Medical Journal:

> "Open Payments reported that [omissis], professor at the University of Michigan School of Public Health and acting chair for the FDA's covid vaccine authorisation meetings, had received over $24 000 (o ((simbolo della sterlina)) 16 970; ∈ 19 650) in payments from drug companies in 2019."
> https://www.bmj.com/content/373/bmj.n1283

Scevri dunque da illusioni (ma sorpresi dalla tolleranza dimostrata da quasi 8 miliardi di persone verso pochi) immergiamoci dunque nello studio dei *numeri* della Statistica, volendoli supporre *veri*, numeri che con la loro purezza sono stati di consolazione ad intere generazioni di matematici, nel corso dei secoli.



Sulla realtà reale invece pesa il

"this is not science, this is business"

citato dall'Editore del British Medical Journal nel 2021. Scrive
https://www.open.online/2021/11/08/covid-19-vaccini-peter-doshi-appell

> Veniamo ora alla frase più controversa di Doshi, che si presta facilmente a manipolazioni. Lui afferma infatti a un certo punto che «questa non è Scienza». Cosa voleva dire esattamente? «È stata estrapolata dal contesto la frase "questa non è Scienza" – spiega Stingi – ma lui in quel caso stava facendo una critica all'aspetto del business. (...)»

Appunto.

> Il dr Szamatolski scrisse: "è mia opinione che la grave condizione della mandibola (di una paziente la cui bocca stava andando in cancrena spontaneamente) è stata causata dall'influenza del radio".

> ...
>
> Era un'idea radicale, tuttavia aveva della scienza a supporto. C'era una considerevole letteratura sui rischi del radio.
>
> ...
>
> Però, l'altro lato della medaglia era tutta la letteratura positiva sul radio.
>
> ...
>
> Ma se si osservavano un po' più attentamente le pubblicazioni positive, c'era un comune denominatore: i ricercatori, tutti, lavoravano per aziende del radio.
>
> ...
>
> L'opinione di Szamatolski, quindi era una voce solitaria e ipotetica contro il ruggito fiammeggiante di una ben finanziata campagna di letteratura pro-radio.
>
> Kate Moore, The Radium Girls, pp 52-53



Il conflitto di interessi non riguarda solo gli Autori degli articoli scientifici, ma anche i decisori degli enti autorizzativi, e perfino i reviewers. Su quest'ultimo punto leggiamo dal prestigioso (1883) Journal of the American Medical Association:

> Between 2020 and 2022, 1155 peer reviewers (58.9%) received at least 1 industry payment[184]

## 35.13 Un singolo manipola la sanità mondiale

* Titolo: *"Who's leading WHO? A quantitative analysis of the Bill and Melinda Gates Foundation's grants to WHO, 2000-2024"* (pubblicato su BMJ Global Health) ([*gh.bmj.com*][1])

* I dati principali:

* La Bill & Melinda Gates Foundation (BMGF) ha fatto **640 sovvenzioni** alla World Health Organization (WHO) nel periodo 2000-2024, per un valore complessivo di **US$ 5.5 miliardi**. ([*PubMed*][2])

[184] Nguyen D, Muramaya A, Nguyen A, et al. Payments by Drug and Medical Device Manufacturers to US Peer Reviewers of Major Medical Journals. JAMA. Published online October 10, 2024. doi:10.1001/jama.2024.17681

* Queste sovvenzioni rappresentano circa il **6.4 % del totale** dei grant della BMGF nel periodo considerato. ([$PubMed$][2])

* Di questi 5.5 miliardi, circa **US$ 4.5 miliardi (≈82.6 %)** sono stati destinati a malattie infettive. ([$PubMed$][2])

* Circa **US$ 3.2 miliardi ( 58.9 %)** sono stati destinati alla poliomielite. ([$PubMed$][2])

* Meno dell'1 % è andato a malattie non trasmissibili, sanitazione, rafforzamento dei sistemi sanitari, nonostante tali ambiti siano strategici per l'OMS. ([$BMJ$][3])

* Il commento dell'articolo sottolinea che la dipendenza dell'OMS da contributi volontari vincolati per settori specifici – come quelli della BMGF – può "plasmare" le priorità dell'agenzia, indirizzandola verso le malattie che piacciono ai donatori piuttosto che verso un'agenda sanitaria più ampia. ([$BMJ$][3])

[1]: https://gh.bmj.com/content/10/10/e015343/"Who's leading WHO? A quantitative analysis of the Bill and ..." [2]: https://pubmed.ncbi.nlm.nih
leading WHO? A quantitative analysis of the Bill and ..." [3]: https://bmjgroup.com/world-health-organizations-priorities-shaped-by-its-reliance-on-grants-from-donor-organisations-such-as-the-gates-foundation/"World Health Organization's priorities shaped by its ..."

# X – Stimatori puntuali e intervallari