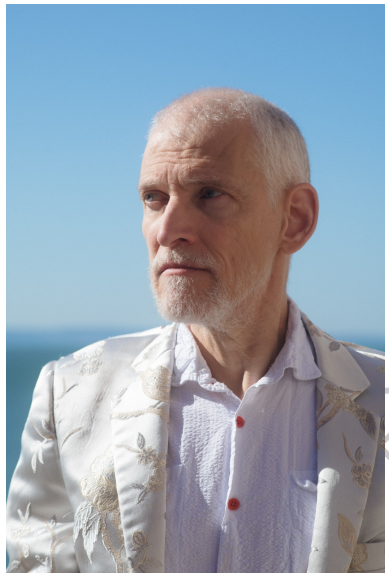
# Le Compresse di Matematica per le Scienze della Farmacia

**Lezioni 20-21**

versione 4 novembre 2026

BOZZA - DRAFT

Università degli Studi di Trieste



Alessandro Soranzo
Ricercatore universitario
Settore Scientifico Disciplinare:
MATH-03/A - Analisi matematica
Gruppo Scientifico Disciplinare:
01/MATH-03 - Analisi matematica, probabilità e statistica matematica

BOZZA - DRAFT

# Indice

**20 Quartili e box-plot** **6**
20.1 Quartili e riassunto dei 5 numeri . . . . . 6
20.2 Bibliometria delle scienze farmaceutiche . . . . . 10
20.3 Box-plot ovvero diagramma a scatola e baffi . . . . . 11
20.4 Esempio: box-plot delle prime 100 radici . . . . . 12
20.5 Definizioni formali dei box-plot . . . . . 13
20.6 Oltre i quartili: percentili in Medicina e Farmacia . . . 17
20.7 I dati? Quali dati? . . . . . 18
20.8 Complementi – Box plot riguardante il covid 19 . . . 20
20.9 ESERCIZI SULLA LEZIONE 20 . . . . . 21
20.9.1 Esercizio risolto a – Media interquartile . . . . 21

**21 Variabilità, correlazione, note finali** **24**
21.1 Indici di dispersione ovvero variabilità . . . . . 24
21.2 Covarianza; correlazione; retta di regressione . . . . . 31

BOZZA - DRAFT

21.3 L'inimicizia fra progresso e variabilità . . . . . 38
21.4 Note finali sulla Statistica Descrittiva . . . . . 43
21.5 Bugie, bugie cattive, e statistica . . . . . 44
21.6 La scelta dei parametri. Cosa misurare? . . . . . 45
21.7 Rispetto a quale standard fare le statistiche? . . . . . 51
21.8 Col ranking facciamo ben figurare il Paraguay . . . . 59
21.9 Illusioni percettive nella presentazione dei dati . . . . 60
21.10Di cosa parliamo? . . . . . 64
21.11Cherry picking – bias di selezione . . . . . 66
21.12Approfittare delle fluttuazioni statistiche . . . . . 67
21.13Capziosità verbali . . . . . 70
21.14Illustrazione di un triplice esempio reale . . . . . 71
21.15Conclusioni sulla fallace presentazione di dati veri . . 74
21.16Il valore anomalo: outlier . . . . . 78
21.17Omissione di dati ritenuti poco significativi . . . . . 80
21.18"Fatta la legge, trovato l'inganno" . . . . . 81
21.19Traslazioni e confusioni linguistiche . . . . . 82

BOZZA - DRAFT

21.20Il problema della discrezionalità . . . . . 85
21.21Esempio importante: le cause di morte . . . . . 85
21.22Cose vere in articolo titolato oppostamente . . . . . 89
21.23La falsificazione dei dati . . . . . 91
21.24Legami finanziari e risultati della ricerca farmacologica 97
21.25Chi finanzia lo studio? . . . . . 100
21.26ESERCIZI SULLA LEZIONE 21 . . . . . 105

BOZZA - DRAFT

# 20 Quartili e box-plot

Vogliamo riassumere il dataset in un modo più ricco rispetto alla sola mediana:

– aggiungeremo alla mediana 4 numeri;
– con quei 5 numeri faremo 1 diagramma.

BOZZA - DRAFT

## 20.1 Quartili e riassunto dei 5 numeri 11:10

Si abbia il dataset numerico $x_1, .., x_n$ e lo si riordini in modo crescente:

$y_1, ..., y_n$ *sono gli stessi numeri ma in ordine crescente*.

Il valore intermedio,

o la media dei 2 intermedi se $n$ è pari,

sappiamo che è la mediana,

che da adesso chiameremo anche *secondo quartile*,
e lo indicheremo con

$q_{0.5}$ oppure

$q_{1/2}$ oppure

$q_{50\%}$.

Poi si divide il dataset (ordinato) in 2 parti uguali, escludendo da entrambe, nel caso di numerosità dispari, il valore centrale, e

la mediana della prima parte si chiama *primo quartile*,

e lo indicheremo con $q_{1/4}$ oppure

$q_{0.25}$ oppure

$q_{25\%}$,

e la mediana della seconda parte si chiama *terzo quartile*,

e lo indicheremo con $q_{3/4}$ oppure

$q_{0.75}$ oppure

$q_{75\%}$.

Si ha poi il quartile di indice 0, che è il minimo del dataset,

$q_0$ oppure

BOZZA - DRAFT

$q_{0\%}$

e quello di indice 1, che è il massimo del dataset:

$q_1$ oppure

$q_{100\%}$

Ecco schematizzato il caso di 11 elementi: 0 x 1 x x 2 x x 3 x 4

Si hanno così 5 quartili, che sono numeri. Ma purtroppo:

– la loro definizione non è univoca nei vari testi e[1] software;

– inoltre, di fatto, spesso si trova denominato *quartile* l'insieme di valori *fra* 2 quartili – come sopra definiti – e in questo altro senso, i quartili sono 4 dataset e non 5 numeri.

I 5 numeri detti, $q_{0\%}$,...,$q_{100\%}$, sono una statistica di sintesi (vettoriale[2]) che si chiama riassunto dei 5 numeri (o *five number summary*).

BOZZA - DRAFT

[1] Leggiamo in questo link accademico: "Molti software hanno diversi algoritmi per calcolare i quantili." In particolare la definizione di Tukey include la mediana in entrambe le metà.

[2] *Vettoriale* perché produce una cinquina di numeri e non un solo numero. È solo una questione definizionale.

Due esempi tratti da Wikipedia, l'enciclopedia libera, riscritti:

per un dataset $x_1, ..., x_{10}$, riordinato in $y_1, ..., y_{10}$,
i 5 numeri (quartili) sono nell'ordine, $y_1$, $y_3$, $\frac{y_5+y_6}{2}$, $y_8$, $y_{10}$;

per un dataset $x_1, ..., x_{11}$, riordinato in $y_1, ..., y_{11}$,
i 5 numeri (quartili) sono nell'ordine, $y_1$, $y_3$, $y_6$, $y_9$, $y_{11}$.

BOZZA - DRAFT

## 20.2 Bibliometria delle scienze farmaceutiche 11:18

Secondo Scimago, la rivista scientifica *Journal of Pharmacy and Pharmacology* ha (nel 2026) l'H index, che è un indicatore bibliometrico, di 148: è tanto o poco? Se avessimo la lista di tutte le riviste scientifiche catalogate per Pharmacology da Scimago, e ciascuna col suo H index, avremmo la risposta, salvo che sarebbe scarsamente leggibile per eccesso di numeri. Un riassunto dei 5 numeri ci permetterebbe di contestualizzare bene quel valore 107. Scimago fa una cosa del genere, magari non esattamente con l'H index (usa un suo indice interno) ma la sostanza è la stessa. Salvo che chiama primo quartile quello più alto, cioè dal terzo quartile (numero) al quarto quartile (numero). Scopriamo così che la rivista considerata sta nel secondo quartile (2025) da anni, bene ma non ottimo. Invece la rivista *British Journal of Pharmacology* risulta nel primo quartile, il migliore, dal 1999 almeno fino al 2025.

BOZZA - DRAFT

## 20.3 Box-plot ovvero diagramma a scatola e baffi 11:21

Il *box-plot* ovvero diagramma a scatola e baffi è un diagramma molto usato in ambito Biomedico per rappresentare tutti i quartili di un dataset, ovvero il riassunto dei 5 numeri, e magari anche gli outlier. Vediamo sulla rete alcuni esempi di pertinenza delle Scienze Mediche e Farmaceutiche seguendo questo `link->`, con immagini di PubMed.

- Si può dare versione orizzontale o verticale;
- talvolta gli outlier vengono rappresentati staccati dai baffi;
- esistono varianti, sulla lunghezza dei "baffi"[3], ma non[4] solo;
- chi pubblica un box plot dovrebbe specificare[5] i dettagli.

Vediamo ora il funzionamento con un esempio pratico.

BOZZA - DRAFT

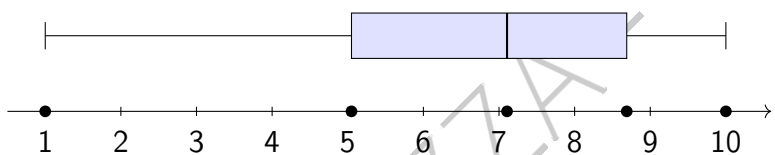
[3] Si veda per esempio questo link.

[4] Per esempio in questo link vediamo una crocetta: possiamo ragionevolmente ipotizzare che rappresenti un decesso.

[5] Per esempio in questo link leggiamo "the whiskers show the smallest and highest value within 1.5 box lengths from the box", che è una variante comune.

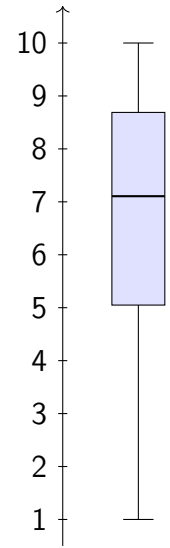
## 20.4 Esempio: box-plot delle prime 100 radici 11:24

Box-plot (qua orizzontale) delle radici quadrate dei numeri interi da 1 a 100 che sono 1, $\sqrt{2}$, $\sqrt{3}\ldots\sqrt{99}, 10$.
Si vede bene che la mediana è maggiore di 7 e i valori si addensano più verso il 10 che verso l'1. I pallini sono i valori del riassunto dei 5 numeri, ma in generale non vengono evidenziati come invece qui.




La "scatola" (box) racchiude i 50 valori centrali, dal primo quartile ($Q_1 \approx 5{,}05$) al terzo ($Q_3 \approx 8{,}69$). Il tratto verticale interno indica la mediana $\approx 7{,}11$. I baffi raggiungono il minimo $\sqrt{1} = 1$ e il massimo $\sqrt{100} = 10$. Il baffo inferiore è più lungo di quello superiore, benché ciascuno dei due intervalli rappresenti 25 osservazioni: asimmetria.

Versione verticale del box-plot dello stesso dataset




BOZZA - DRAFT

## 20.5 Definizioni formali dei box-plot 11:28

Leggiamo su Wikipedia, l'enciclopedia libera, alla voce "Diagramma a scatola e baffi":

> il diagramma a scatola e baffi (o diagramma degli estremi e dei quartili o box and whiskers plot o box-plot) è una rappresentazione grafica [...] Viene rappresentato (orientato orizzontalmente o verticalmente) tramite un rettangolo diviso in due parti, da cui escono due segmenti. Il rettangolo (la "scatola") è delimitato dal primo e dal terzo quartile, $q_{1/4}$ e $q_{3/4}$, e diviso al suo interno dalla mediana, $q_{1/2}$. I segmenti (i "baffi") sono delimitati dal minimo e dal massimo dei valori. In questo modo vengono rappresentati graficamente i quattro intervalli ugualmente popolati delimitati dai quartili.

BOZZA DRAFT

**La versione di Nature.com**, che è un sito web scientifico di altissimo livello, troviamo qui una descrizione dei box plot, in cui leggiamo

> The core element that gives the box plot its name is a box whose length is the IQR and whose width is arbitrary (...) A line inside the box shows the median, which is not necessarily central. The plot may be oriented vertically or horizontally (...) Whiskers are conventionally extended to the most extreme data point that is no more than 1.5 × IQR from the edge of the box (Tukey style) or all the way to minimum and maximum of the data values (Spear style). (...) The 1.5 multiplier corresponds to approximately (...) 99.3% coverage of the data for a normal distribution. Outliers beyond the whiskers may be individually plotted.

L'IQR sopradetto, *interquartile range*, che è $q_{0.75} - q_{0.25}$, la vedremo meglio in seguito; e la *normal distribution*, distribuzione normale, è la più tipica distribuzione delle Scienze Biomediche, e la vedremo in seguito, ma anticipiamo che la sua densità è la già introdotta campana gaussiana.

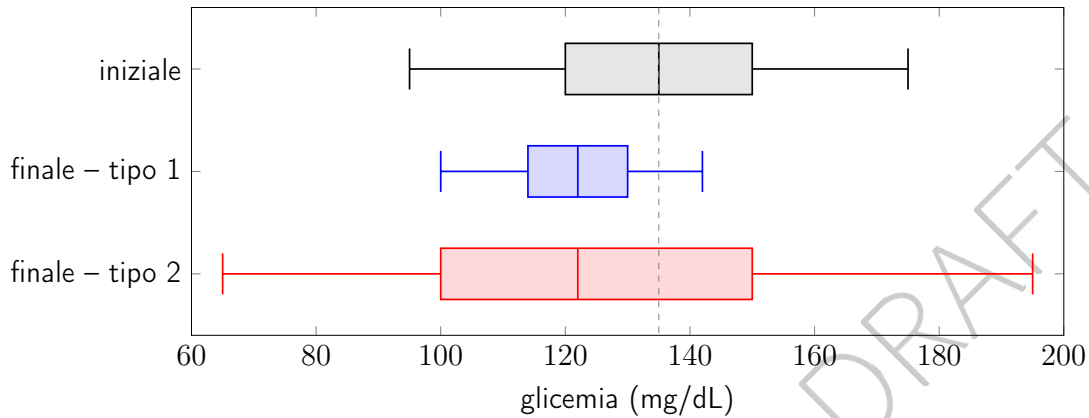
BOZZA - DRAFT

**Applicazione.** Abbiamo 1000 soggetti ciascuno con 1 valore fisiologico – glicemia, peso, qualunque cosa – su cui vogliamo intervenire, e gli diamo un farmaco. Diciamo, la glicemia.
Dopo una settimana abbiamo 1000 valori nuovi, o piuttosto, in generale – nella realtà concreta – un po' meno (qualcuno può essere morto, o irraggiungibile, o non ne vuole più sapere di noi), diciamo 980.
La mediana dei valori iniziali e la mediana dei valori finali, certo, sono un'indicazione importante.
Per fissare le idee supponiamo che si volesse diminuire il parametro.
Supponiamo che la mediana sia diminuita: bene!
Gli altri 4 numeri ci indicheranno più chiaramente come si è evoluta la situazione. (Del solo parametro controllato, ovvio, e questo è un problema generale della Medicina e della Farmacologia: *cosa* stiamo misurando? La glicemia, la mortalità, il benessere, i soldi... possono variare *indipendentemente*).

Per esempio potremmo avere una situazione iniziale e una finale di uno di questi 2 tipi:

con mediana ridotta e valori addensati nel tipo 1,
oppure con mediana ridotta ma valori diradati nel tipo 2.

BOZZA - DRAFT




BOZZA - DRAFT

Non faremo qua Medicina, ma è ovvio che la seconda può essere pericolosa, nonostante il miglioramento della mediana.

Si noti – e questo è un problema generale – che molta parte di verità può sfuggirci nascosta nei soggetti che non abbiamo potuto misurare la seconda volta. Si potrebbe dire: bene, nel primo diagramma rappresentiamo solo i soggetti di cui abbiamo anche la seconda misurazione. Sì, ma il problema resta, perché la mancanza della seconda misurazione potrebbe essere causata in tutto o in parte proprio dal farmaco: morte, irreperibilità per cure all'estero, grave delusione del soggetto. (In parte, è il problema dei *morti per altra causa*).

## 20.6 Oltre i quartili: percentili in Medicina e Farmacia



L'uso più familiare è quello delle *curve di crescita* pediatriche:

"il bambino è al 90° percentile per il peso"

significa che, fra i bambini "come lui" – della stessa età e dello stesso sesso della popolazione di riferimento, per esempio quella dell'OMS – il $90\%$ pesa meno di lui o quanto lui, e il $10\%$ pesa di più.

**Tecnicamente.** Generalizzando i quartili, il *percentile* $k$-esimo $q_{k\%}$ è il valore che lascia alla sua sinistra il $k\%$ dei dati (ordinati). Così il primo quartile, la mediana e il terzo quartile sono il 25°, il 50° e il 75° percentile. (Più in generale si parla di *quantili* $q_p$, con $0 \leq p \leq 1$.)

Il 90° percentile non è un "voto" come un 9 a scuola: un bambino al 90° percentile non sta necessariamente meglio di uno al 50°. Più che il singolo valore, al pediatra interessa se nel tempo il bambino resta su una stessa curva o la attraversa.

BOZZA - DRAFT

## 20.7 I dati? Quali dati? 11:39

*In Italia si usa raccogliere pochi dati, e parecchi di quei pochi nasconderli – per poi decidere su impressioni, ideologicamente.*

Scrive il direttore Marco Cattaneo in un editoriale (settembre 2023) di Le Scienze (edizione italiana di Scientific American):

> era nata la campagna #DatiBeneComune, «per chiedere al Governo italiano di pubblicare in maniera aperta i dati sulla gestione della pandemia di COVID-19». Oggi la campagna, che ha superato i 60.000 firmatari, si rivolge ancora alle istituzioni per disporre di dati trasparenti sui principali temi di interesse dei cittadini, a partire dal Piano nazionale di ripresa e resilienza (PNRR).
>
> I dati, sempre i dati, croce di questo paese. Non dico che si debba avere l'attenzione spasmodica per i numeri degli Stati Uniti, dove c'è una maniacale ossessione per le statistiche persino nel baseball. Ma i numeri, i dati trasparenti e con il massimo grado di dettaglio possibile, dovrebbero essere lo spirito guida delle decisioni pubbliche, soprattutto quando ci sono in gioco investimenti importanti e strategie per il futuro.

BOZZA - DRAFT

> (...) L'analisi delle fonti a disposizione, a livello nazionale, mostra lacune che si sarebbe tentati di definire sconfortanti.
>
> «Il problema, peraltro, non è solo di mancanza del dato, ma anche, nei casi in cui esiste, di una ritrosia a rilasciare i dati», scrivono Faiella e Lavecchia. (Enfasi aggiunta)

La "sconfortante" "ritrosia italiana a rilasciare i dati" sulla mortalità e lo stato vaccinale impedisce (2026) a tutt'oggi (!) a più di 5 anni dalle vaccinazioni di calcolare quanti giorni o mesi o anni di vita gli italiani hanno (sperabilmente) guadagnato con la vaccinazione covid-19 coloro che si sono vaccinati, rispetto ai no-vax – il che genera sospetti nei *malfidenti*.

BOZZA, DRAFT

## 20.8 Complementi – Box plot riguardante il covid 19

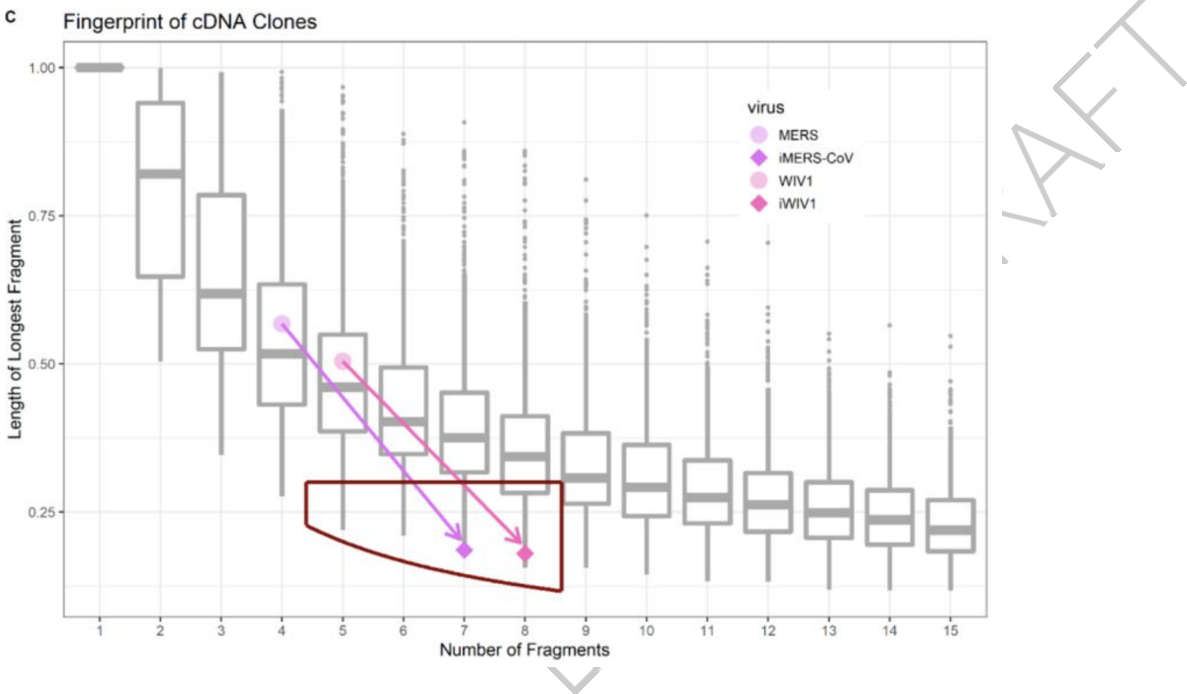



Figura 1: Figura 2 c in "Endonuclease fingerprint indicates a synthetic origin of SARS-CoV-2", by Bruttel, Valentin and Washburne, Alex and VanDongen, Antonius: in bioRxiv (2022), DOI 10.1101/2022.10.18.512756

## 20.9 ESERCIZI SULLA LEZIONE 20 11:43

### 20.9.1 Esercizio risolto a – Media interquartile

$_{\mu 2023}$ ≈ Calcolare la media interquartile del dataset
54,071 117,512 19,097 2,345 282,012 0,112 0,023 0,016 2,150 630,013 2.848 3,410

BOZZA DRAFT

#### SVOLGIMENTO

Viene usato lo standard della virgola decimale. Questo si vede dal numero 0,112 e altri. (Allora il punto in 2.848 è separatore delle migliaia, e questo della virgola decimale e del punto separatore delle migliaia è esattamente lo standard imposto in Farmacia dal Ministero della Salute italiano).

Il dataset ordinato in modo crescente, e scritto senza separatore delle migliaia, è
0,016 0,023 0,112 2,150 2,345 3,410 19,097 54,071 117,512 282,012

630,013 2848 Esso ha $n = 12$ cioè $4 \cdot 3$ elementi

●●● ●●● ●●● ●●●

ed eliminando il primo e ultimo "quartile" dà

2,150 2,345 3,410 19,097 54,071 117,512

e la media aritmetica dei soprastanti 6 numeri

$$\frac{1}{6}(2{,}150 + 2{,}345 + 3{,}410 + 19{,}097 + 54{,}071 + 117{,}512) =$$

$$= \frac{198{,}585}{6} \approx$$

è la media interquartile cercata:

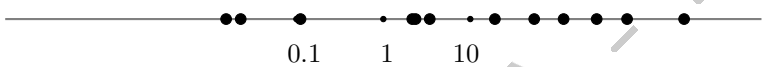
$$\boxed{\approx 33{,}0975}$$

o con minor precisione

BOZZA - DRAFT

$$\boxed{\approx 33{,}1}$$

**Nota 1.** I dati coprono 5 ordini di grandezza; in scala logaritmica si leggono molto meglio. Ecco un diagramma cartesiano in scala logaritmica dei 12 valori (ma 2 sono quasi sovrapposti). Il pallino più a destra corrisponde al 2848, il massimo del dataset.




Salvo l'ordine il dataset è stato ottenuto da una variabile aleatoria log-normale di parametri $\mu = 1$ e $\sigma^2 = 25$, che non sono assolutamente media e varianza della variabile aleatoria. Precisamente sono stati tratti gli esponenziali dei valori della variabile aleatoria normale usata nell'esercizio 5: gli esponenziali dei valori di una variabile aleatoria normale sono distribuiti log-normalmente.

Si noti l'uguaglianza, salvo i numeri scritti all'asse delle ascisse, col diagramma dell'esercizio **??**, che non è in scala logaritmica.

BOZZA - DRAFT

# 21 Variabilità, correlazione, note finali

## 21.1 Indici di dispersione ovvero variabilità 11:49

*[...] seconno le statistiche d'adesso*
*risurta che te tocca un pollo all'anno:*
*e, se nun entra nelle spese tue,*
*t'entra ne la statistica lo stesso*
*perché c'è un antro che ne magna due.* [Trilussa]

BOZZA - DRAFT

**Gli indici di dispersione ovvero variabilità cercano di quantificare con 1 numero la non omogeneità di dati numerici.**

Sono statistiche di sintesi $f(x_1, ..., x_n)$ da $\mathbb{R}^n$ in $\mathbb{R}$, con $n$ a priori indeterminato, proprio come le varie medie, viste in precedenza.
È ovvio che dataset diversi possono avere diversi gradi di "disomogeneità", che vogliamo quantificare, anche se hanno la stessa media e/o

mediana. Per esempio i 2 dataset, qua già ordinati,

$$2, 4, 6, 8, 10, 12, 14 \qquad 5, 6, 7, 8, 9, 10, 11$$

hanno uguale media, e perfino mediana, sempre 8, ma è chiaro che i valori del secondo dataset variano meno, ovvero sono più addensati intorno alla media. Si pensi a dei redditi per esempio.

**Campo di variazione, o *range*.**

$$range(x_1, ..., x_n) := \max(x_1, ..., x_n) - \min(x_1, ..., x_n) .$$

Che vale 12 e rispettivamente 6 nei 2 dataset considerati.
Il range è molto sensibile agli eventuali outlier.

Il range come sopra definito è un numero, ma altri Autori chiamano range l'intervallo $[\min, \max]$.

**Differenza interquartile.** Ben poco sensibile agli outlier:

$$iqr(x_1, ..., x_n) := q_{3/4} - q_{1/4} .$$

BOZZA DRAFT

## Varianza, deviazione standard, coefficiente di variazione.

Detta $\bar{x}$ la media aritmetica del dataset $x_1, ..., x_n$ si definiscono

$$\boxed{\textit{population variance} \quad \sigma_X^2 = Var(x_1, ..., x_n) := \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2} \tag{1}$$

*population standard deviation*

$$\boxed{\sigma_X = sd(x_1, ..., x_n) := \sqrt{Var(x_1, ..., x_n)}} \tag{2}$$

$$\textit{ovvero} \quad = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n}}$$

*coefficiente di variazione ovvero deviazione standard relativa*

$$\boxed{se\ \bar{x} \neq 0: \quad \sigma_X^* = cv(x_1, ..., x_n) := \frac{\sigma_X}{\bar{x}}}$$

e dell'ultima si può avere la forma percentuale moltiplicando per 100%, ottenendo la *deviazione standard percentuale*.
Alcuni Autori dividono per $|\bar{x}|$ invece che $\bar{x}$.

BOZZA DRAFT

**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[(6)]

> 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). The standard deviation was 7.59 cm. (...) Women were smaller on average, with a mean height of 164.7 cm, and standard deviation of 7.07 cm.

e la maggior variabilità dell'altezza maschile (7.59 vs 7.07) si rivela solo apparente perché dividendo per le altezze medie si ottiene, salvo decimali, lo stesso coefficiente di variazione del 4.3%.

(E lo scrivente si permette di ritenere che la piccola differenza sarebbe stata ancora più piccola se tutti i calcoli fossero stati fatti usando i logaritmi delle altezze – ovviamente privati delle unità di misura – arrivando a qualcosa di forse meno intelligibile ma più intrinseco della cosa).

Si trova scritto sia $\sigma$ che $sd$ che $SD$ (in inglese *standard deviation*); e sia $\sigma_X^*$ che $cv$ che $CV$ (in inglese *coefficient of variation*), detto anche $RSD$ (*relative standard deviation*), soprattutto in Chimica analitica,

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

BOZZA - DRAFT

dove spesso è espresso in percentuale, $\%RSD$.

Questi indici sono tutti sensibili agli outlier.

La varianza ha un'unità di misura diversa da quella del dataset.

La deviazione standard, di gran lunga preferita nelle Scienze Applicate alla varianza che invece si preferisce in Matematica, ha la stessa unità di misura del dataset, invece la varianza no. Per esempio litri invece che... litri quadrati.

Si noti il fattore $\frac{1}{n}$ nella varianza, e allora poi negli altri indici da essa dipendenti. Certi testi e software danno $\frac{1}{n-1}$. La questione sarà approfondita in Statistica Inferenziale, ma anticipiamo la questione, peraltro numericamente poco influente per $n$ grande:

> If the values instead were a random sample drawn from some large parent population (for example, they were 8 marks randomly and independently chosen from a class of 2 million), then one often divides by 7 (which is $n-1$) instead of 8 (which is $n$) in the denominator of the last formula. In that case the result of the original formula would be called the sample standard deviation. Dividing by $n-1$ rather than

BOZZA - DRAFT

by $n$ gives an unbiased estimate of the variance of the larger parent population. This is known as Bessel's correction.[7]

**Nota.** Grande varianza ovvero deviazione standard: valori dispersi.

Piccola varianza ovvero deviazione standard: valori addensati presso la media.

La deviazione standard può essere considerata piccola oppure grande relazionandola alla media.

Per la varianza, esistono varianze più piccole (valori più addensati) o più grandi di altre (valori più dispersi) ma non ha molto senso considerare grande o piccola una singola varianza. Similmente la deviazione standard.

BOZZA - DRAFT

[7] da Wikipedia (in inglese), l'enciclopedia libera, alla voce Standard deviation, in https://en.wikipedia.org/wiki/Standard_deviation, il 24 ottobre 2019

**Esempio 1.**

> the average height for adult men in the United States is about 70 inches (177.8 cm), with a standard deviation of around 3 inches (7.62 cm).[8]

**Esempio 2.** In un articolo scientifico[9] hanno confrontato, mediante le deviazioni standard, la variabilità dell'ora dell'andare a dormire con l'ora dell'alzarsi in popolazioni di tipo primitivo, trovando maggior variabilità nel primo orario:

> The SD of sleep onset times exceeded the SD of sleep offset times

Il confronto fra due deviazioni standard di grandezze omogenee è ben sensato.

BOZZA - DRAFT

[8] da Wikipedia (in inglese), l'enciclopedia libera, alla voce Standard deviation, in https://en.wikipedia.org/wiki/Standard_deviation, il 24 ottobre 2019

[9] "Natural sleep and its seasonal variations in three pre-industrial societies", by Yetish G, Kaplan H, Gurven M, Wood B, Pontzer H, Manger PR, Wilson C, McGregor R, Siegel JM. Curr Biol. 2015 Nov 2;25(21):2862-2868. doi:10.1016/j.cub.2015.09.046. Epub 2015 Oct 17.

## 21.2 Covarianza; correlazione; retta di regressione



Abbiamo 2 grandezze misurate sugli stessi soggetti, per esempio l'età e la funzionalità renale (Figura 3), e ci chiediamo: quando una cresce, l'altra tende a crescere o a calare?

Tracciamo nello scatterplot una croce passante per il *punto medio* $(\bar{x}, \bar{y})$: il piano resta diviso in 4 quadranti.

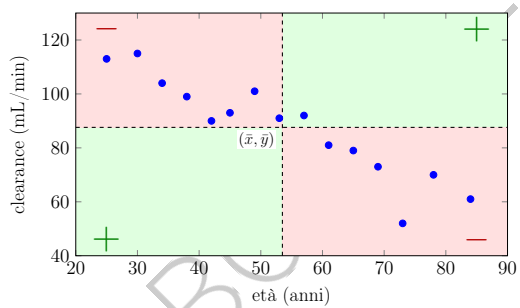



Figura 2: La croce tratteggiata passa per il punto medio $(\bar{x}, \bar{y})$. I punti nei quadranti verdi danno contributi positivi alla covarianza, quelli nei quadranti rossi negativi. Qui prevalgono i rossi: covarianza negativa.

BOZZA DRAFT

• Un soggetto in alto a destra (più anziano *e* con reni più efficienti della media) o in basso a sinistra dà un contributo *positivo*;
• un soggetto in alto a sinistra o in basso a destra dà un contributo *negativo*;
• ogni contributo è tanto più grande quanto più il punto è lontano dalla croce.

La covarianza è la media di tutti questi contributi. Se prevalgono i punti "concordi" è positiva, se prevalgono quelli "discordi" è negativa, se si compensano è circa 0. Nella nostra figura i punti stanno quasi tutti in alto a sinistra e in basso a destra: covarianza negativa.

Dividendo la covarianza di 2 dataset *appaiati* – cioè della stessa numerosità, con l'$i$-esimo valore del primo e l'$i$-esimo del secondo misurati sullo stesso soggetto, come età e clearance di una stessa persona – per il prodotto delle loro 2 deviazioni standard, si ottiene l'indice di correlazione, sempre compreso fra $-1$ e $1$: vicino a $\pm 1$ i punti stanno quasi allineati, vicino a 0 formano una nuvola senza direzione lineare.

BOZZA - DRAFT

La retta di regressione è la retta che "passa in mezzo" ai punti nel modo migliore in un sottile senso matematico (Nel senso che rende minima la somma dei quadrati delle distanze *verticali* dei punti dalla retta). La sua pendenza dice di quanto cambia *in media* la $y$ quando la $x$ aumenta di 1: nella figura, circa 1 mL/min di clearance in meno per ogni anno di età.

BOZZA - DRAFT

– La correlazione misura solo legami *lineari* (una relazione a campana può avere correlazione 0);

– pochi outlier possono spostare molto la retta;

– e soprattutto **la correlazione non dimostra causalità**. In Lombardia possono esserci molte nascite e molte cicogne, in Molise poche nascite e poche cicogne, ma questo non dimostra strane cose; più probabilmente è dovuto all'enormità della Lombardia rispetto al Molise sia come area che come popolazione.

**Formule. Non le impariamo e si calcolano con computer.**
Dati 2 dataset numerici di uguale numerosità

$$X : x_1, ..., x_n \qquad Y : y_1, ..., y_n$$

si definisce la loro *covarianza* (ovvero *covarianza di 2 $n$-uple di numeri*, ovvero *covarianza osservata*)

$$Cov(x_1, ..., x_n; y_1, ..., y_n) := \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + ... + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

dove $\bar{x}$ e $\bar{y}$ sono le medie dei 2 dataset. Si trova denotata anche $\sigma_{X,Y}$, oppure con $cov($... minuscolo. Si definisce anche l'*indice di correlazione di [Bravais-]Pearson*, o *indice di correlazione lineare* (e al posto della parola *indice* altri scrivono *coefficiente*)

$$\rho_{X,Y} := \frac{Cov(x_1, ..., x_n; y_1, ..., y_n)}{\sqrt{Var(x_1, ..., x_n) Var(y_1, ..., y_n)}} \qquad = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

e questo dà un'idea di come varino le $y_j$ relativamente alle $x_i$: un indice prossimo a 1 suggerisce che al crescere delle $x_i$ crescano linearmente le $y_i$, un indice prossimo a -1 suggerisce che al crescere delle $x_i$

BOZZA - DRAFT

decrescano linearmente le $y_i$, e un indice prossimo a 0 indica la non presenza di tali relazioni approssimate. In ogni caso

$$-1 \leq \rho_{X,Y} \leq 1\,.$$

Coi dataset $X$ : 2, 7, 4, 3, e $Y$ : 3, -1, 0, 2 si calcoli l'indice di correlazione, e si rappresentino sul piano cartesiano i punti $(x_i, y_i)$.
In un senso non banale, una retta $y = mx + q$ si avvicina meglio ai punti se ha coefficienti

$$m := \frac{\sigma_{X,Y}}{\sigma_X^2} \qquad q := \bar{y} - m\bar{x} \quad \text{equivalentemente: } m := \frac{\sigma_Y}{\sigma_X}\rho_{X,Y}$$

e in questo caso si chiama *retta di regressione* (dietro, vi è il *metodo dei minimi quadrati*, che non approfondiamo).
Per esempio per i punti $(2,1)$, $(1,3)$, $(3,-1)$, ovvero per i dataset $X:\ 2,1,3$ e $Y:\ 1,3,-1$, si trova $y = -2x + 5$, che passa *esattamente* per i punti. È $\bar{x} = 2$, $\bar{y} = 1$, $\sigma_X^2 = \frac{2}{3}$, $\sigma_{X,Y} = -\frac{4}{3}$. E con $y_2 := 3.14159265$ si trova $y \approx -2.07x + 5.19$. Si facciano i disegni.

BOZZA - DRAFT

# Il coefficiente di correlazione e la retta di regressione sono fra i più usati strumenti matematici delle Scienze Biomediche.

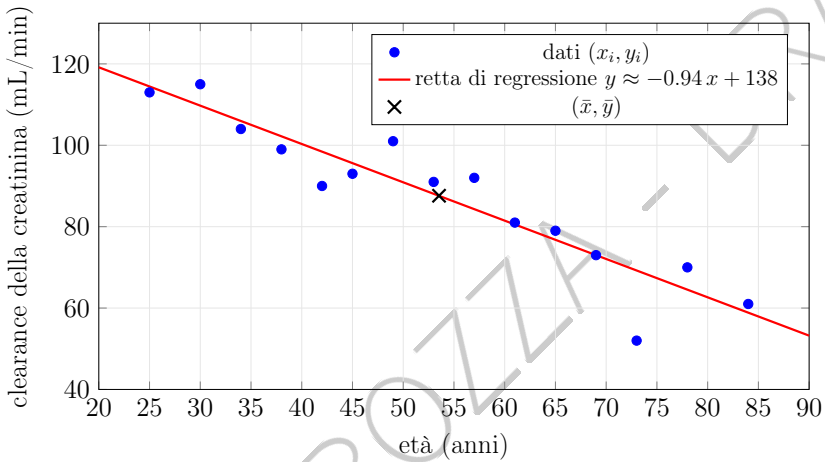



Figura 3: Scatterplot di 15 soggetti (dati *fittizi*, a scopo illustrativo) e retta di regressione. La funzionalità renale cala con l'età: indice di correlazione $\rho \approx -0.93$. La retta passa sempre per il punto delle medie $(\bar{x}, \bar{y}) \approx (53.5,\ 87.6)$, come segue da $q = \bar{y} - m\,\bar{x}$.

BOZZA - DRAFT

**È la retta che in assoluto meglio riassume uno scatterplot?** No, tant'è vero che facendo la retta di regressione $y = m\,x + q$ si trova in generale una retta diversa dalla retta di regressione $x = m'\,y + q'$, eppure lo scatterplot è lo stesso. Bisogna fissare la variabile che riteniamo indipendente, quella in questo paragrafo denotata con $x$. La questione è sottile e non la approfondiamo.

BOZZA - DRAFT

## 21.3 L'inimicizia fra progresso e variabilità 12:20

*Nominate la variabilità a un Ingegnere, che progetta quasi tutte le cose materiali che vedete intorno a voi, e gli si drizzeranno i capelli in testa: per l'Ingegneria la riduzione della varianza nei processi produttivi è obiettivo prioritario.* Anche nell'industria Farmaceutica.

Se guardate i flaconi di vetro in un museo di antichità romane, vedete che sono alquanto variabili nella forma: sono fatti a mano uno per uno, di vetro soffiato. Se una farmacia ne comprava una certa quantità per mettervi un certo farmaco, erano tutti diversi.
Se guardate in una farmacia odierna i flaconi di un qualunque prodotto, sono tutti uguali, o, per meglio dire, la variabilità – che a livello molecolare comunque esisterà – è ridottissima.

**Questo sopradetto è un fatto generalissimo del progresso.**

BOZZA - DRAFT

Tutti gli oggetti di uso tecnico e/o scientifico, dai flaconi alle automobili, sono prodotti con una variabilità piccolissima, e la sostanziale inesistenza di variabilità – comunque esistente a un livello microscopico – è uno dei principali indicatori della bontà di un processo produttivo.

Ancora all'inizio del '900 auto e aerei erano costruiti uno per uno.

Potete accettare che le clementine e i limoni, prodotti nei campi, abbiano una certa variabilità nelle dimensioni, ma non certo

le bottiglie d'acqua minerale
le siringhe
i flaconi di un certo farmaco
le compresse o capsule di un certo farmaco
i chip dei computer, eccetera eccetera.

E comunque anche per i limoni, e gli altri prodotti agricoli, col tempo i vari standard legislativi sono sempre più restrittivi riguardo la variabilità: i limoni di *size code* 7, i più piccoli, devono avere diametro fra 45 e 52 mm, e comunque anche senza quel codice non si possono vendere come limoni dei limoni di diametro minore di 45 mm. (Non stiamo qua negando che siano limoni ma non potete venderli come limoni; ve li accetteranno forse come materiale per compost). LINK

BOZZA - DRAFT

->

È in atto un secolare processo che tende a ridurre sempre più la variabilità degli oggetti prodotti in una determinata serie.

## Icasticamente:
## *il progresso odia la variabilità*

L'uniformità ovvero bassissima variabilità, misurata su qualche parametro (per esempio il diametro dei limoni) consente

- risparmio economico
- risparmio di tempo

In campo farmaceutico, il danno della variabilità è evidente nella necessità di avere dosaggi garantiti uniformi in ogni flacone, in ogni capsula, in ogni compressa, eccetera. È ovvio.
Fra 2 compresse di aspirina di una stessa marca c'è da aspettarsi una

BOZZA - DRAFT

variabilità infinitesimale.

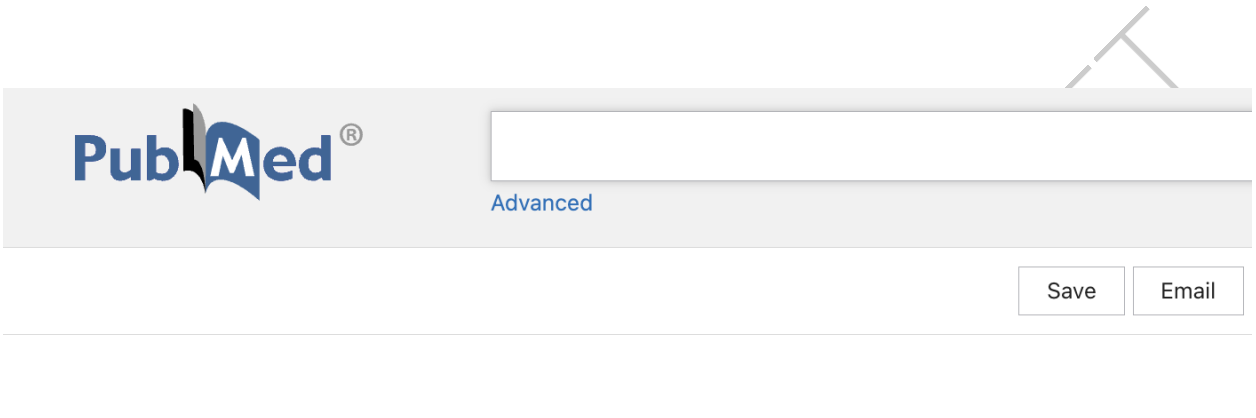



Figura 4: Anche il farmaco più miracoloso che l'Uomo avesse mai prodotto potrebbe talvolta fare male se – per ipotesi assurda – il prodotto che arriva all'utenza non è esattamente uguale a quello previsto.

Altro indice importante è in https://it.wikipedia.org/wiki/

Coefficiente_di_Gini

BOZZA - DRAFT

## 21.4 Note finali sulla Statistica Descrittiva 12:30

Ad un livello basico, nella Statistica Descrittiva **si vuole riassumere molti valori (*dataset*) con 1 valore o pochi valori, e/o con un diagramma**.

Quest'opera di sintesi non è banale.

Viene fatta per comprendere i dati, divulgarli, confrontarli. Alcuni lo fanno perfino per presentarli manipolativamente; allora noi dovremo attrezzarci conoscitivamente per non essere tratti in inganno.

Questo è centrale in Farmacia. Basta ridefinire 1 soglia, e qualche miliardata di persone da sane potrebbero trovarsi riclassificate come malate, verosimilmente bisognose di farmaci.
Leggiamo su un articolo[10] del British Medical Journal:

BOZZA - DRAFT

[10] Moynihan R, Heath I, Henry D. Selling sickness: the pharmaceutical industry and disease mongering. BMJ. 2002 Apr 13;324(7342):886-91. doi: 10.1136/bmj.324.7342.886. PMID: 11950740; PMCID:

> There's a lot of money to be made from telling healthy people they're sick. Some forms of medicalising ordinary life may now be better described as disease mongering: widening the boundaries of treatable illness in order to expand markets for those who sell and deliver treatments. Pharmaceutical companies are actively involved in sponsoring the definition of diseases and promoting them to both prescribers and consumers. The social construction of illness is being replaced by the corporate construction of disease.

BOZZA - DRAFT

## 21.5 Bugie, bugie cattive, e statistica

Da più di un secolo si dice "lies, damned lies, and statistics". Ai numeri si può far dire quel che si vuole o quasi.

# Attenzione!

## PARTE I – DATI VERI

PMC1122833.

## 21.6 La scelta dei parametri. Cosa misurare?

In questa trattazione elementare siamo sempre partiti da dati numerici, ma a monte c'è il problema di quali numeri considerare in una data situazione.

Se consideriamo un quadrato di lato 2 ovvero area 4, e 2 quadrati di lato 5 ovvero area 25, e volessimo rappresentare un "quadrato medio", facendo la media dei lati abbiamo 4 (con area 16), invece facendo la media delle aree abbiamo 18 (con lato $3\sqrt{2} \approx 4.24$).

Qual è il senso di quelle tabelle che mettono ai primi posti il peperoncino per contenuto di vitamina C? Il peperoncino si mangia, eventualmente, a grammi, mica a etti come certa frutta, che, certo, contiene meno vitamina C *all'etto*, ma ne contiene di più *per porzione* (*ragionevole*, per quanto ciò possa significare). La scelta di un diverso parametro cambia completamente l'impostazione della questione.

BOZZA - DRAFT

"Il Molise ultimo per decessi covid"

sì ne aveva solo 31, ad una certa data, meno di ogni altra regione, ma ha anche pochi abitanti. Considerando i decessi per 100mila abitanti ne aveva 10, più della Basilicata che ne aveva 8, pur avendo un numero maggiore di decessi. Non è più ultimo.

Se leggiamo

*La regione X prima per tumori*

potremmo farci l'idea che la regione X sia più inquinata, o meno assistita medicalmente, magari peggio governata da quel noto partito che ha la maggioranza nella regione X – associazioni d'idee che magari sono il motivo reale per cui è stato pubblicato quell'articolo sui media – ma il contenuto informativo di quel titolo è seminullo, e quelle argomentazioni che ci sorgono istintivamente sono praticamente insensate. Potrebbe ben essere che nella regione X i tumori siano 10 volte meno probabili *per ogni singolo residente* che nella regione Y, ma la regione X ha più del decuplo di abitanti della regione Y, per cui *il numero assoluto di casi* della regione X sopravanza quello della regione Y, molto più inquinata e meno assistita medicalmente.
Nella situazione considerata, altrettanto legittimo sarebbe stato il titolo

BOZZA - DRAFT

*La regione Y prima per tumori*

e in un caso numerico estremo potrebbe essere vero addirittura

*Nella regione X la **minima** incidenza di tumori*

(Per una disamina del termine *incidenza*: ->Link).

Anche qua la scelta di un diverso parametro cambia completamente l'impostazione della questione.

Numero assoluto dei morti: non dice nulla su inquinamento, malasanità...

Tasso di mortalità: serio indizio, su varie possibilità.

Il numero assoluto, spesso espresso da cifre ad alto impatto emotivo ("milioni di tumori"), da un punto di vista medico e farmaceutico ci dice meno della frequenza relativa, che in generale si potrà misurare in casi per $1\,000$ residenti, o per $100\,000$, o per milione. (Per evitare scritture come $0.000077$, scarsamente leggibili: molto più chiaro, come nell'esempio del link sopra riportato, $7.7$ casi per $100\,000$).
Si noti che la frequenza relativa corrisponde alla probabilità, per 1 sin-

BOZZA - DRAFT

golo residente preso a caso. Per esempio 50 casi ogni 1000 abitanti dà una probabilità di $\frac{1}{20}$. Si tratta della concezione frequentista della probabilità.
La frequenza assoluta ha una diversa importanza: corrisponde alla *dimensione economica* complessiva della questione: è ovvio che se in un certo luogo ci sono più casi di una malattia usualmente trattata con un certo farmaco, là c'è un maggior *mercato* per quel farmaco.

**In ambito medico e farmaceutico.** Ragionando con la logica comune – sbagliata – trovare una riduzione nelle aritmie che causano morte è un buon risultato per un farmaco, e si verifica abbastanza facilmente; verificare la sopravvivenza a lungo termine è molto più oneroso (e in passato ha dato risultati opposti).

La sopravvivenza a 5 anni dalla diagnosi di cancro è un indice di altissimo valore ma richiede molto tempo e denaro; la riduzione della massa tumorale in un breve tempo si verifica prima.

BOZZA- DRAFT

## Cosa misuriamo è un serio problema generale della Medicina e della Farmacologia.

Vogliamo massa tumorale ridotta? A 2 mesi o a 6 mesi o a 1 anno dall'inizio del trattamento, o quando? O vogliamo diminuita mortalità, ma, di nuovo, quando misurata? O vogliamo aumentata sopravvivenza complessiva? Oppure aumentata *progression-free survival*? O migliore *response rate*, tasso di risposta del tumore alla terapia valutato misurando qualche *biomarker*? O diminuita spesa? (Che anche quello conta). O migliorata qualità della vita? (Sotto-problema: misurata come?) O un po' tutto, con una media pesata con coefficienti da decidere? In ognuno di quei casi si otterranno risultati in generale diversi.

Leggiamo sul Journal of Clinical Oncology[11]

BOZZA - DRAFT

[11] American Society of Clinical Oncology Perspective: Raising the Bar for Clinical Trials by Defining Clinically Meaningful Outcomes. By: Lee M. Ellis et al. DOI: 10.1200/JCO.2013.53.8009 Journal of Clinical Oncology 32, no. 12 (April 20, 2014) 1277-1280.

> produce results that are clinically meaningful to patients (ie, significantly improved survival, quality of life [QOL], or both)

(La riduzione della massa tumorale, così tanto usata, appare per certi versi la cosa che *meno* possa interessare il paziente).
*Lo stesso discorso vale praticamente per ogni patologia*.

Icasticamente:

*L'operazione è riuscita ma il paziente è morto*

Leggiamo in un articolo scientifico:[(12)]

> The last few years have seen an increase in the number of randomized controlled trials (RCTs) of new agents in metastatic solid tumors using progression-free survival (PFS) as the primary end point. Some trials showing improvement in PFS, without a corresponding increase in overall survival (OS), have led to approval of new drugs and/or changes

BOZZA - DRAFT

[12] Progression-Free Survival: Meaningful or Simply Measurable? By: Christopher M. Booth, Elizabeth A. Eisenhauer. DOI: 10.1200/JCO.2011.38.7571 Journal of Clinical Oncology 30, no. 10 (April 01, 2012) 1030-1033. https://ascopubs.org/doi/full/10.1200/JCO.2011.38.7571

> in standard of care. This suggests a growing belief in the oncology community that delaying progression in metastatic disease is a worthy goal, even if OS is not improved. But is a new treatment that improves PFS really an advance for patients? Or is it only lowering the bar to declare active some of our much-heralded new molecular targeted therapies?

## 21.7 Rispetto a quale standard fare le statistiche?

Una fonte di problematicità è quale standard scegliere. Se per esempio volessimo vedere se la carenza di una certa sostanza nel sangue è statisticamente correlata ad una malattia, avremmo – oltre al problema della diagnosi della malattia nei soggetti, problema medico che non considereremo – il problema di quali soggetti considerare carenti e quali no, in base alle analisi chimiche del sangue: infatti ci sono molti standard per i vari livelli delle sostanze nel sangue, e in generale per i parametri biomedici, come l'essere sovrappeso, e come la mortalità infantile e l'aspettativa di vita alla nascita, standard che variano

BOZZA - DRAFT

nel tempo e nelle grandi entità scientifiche, sanitarie e/o geopolitiche ragionevolmente considerabili: Ministero della Salute italiano, Unione Europea, OMS (Organizzazione Mondiale della Sanità), ONU, USA (NIH, CDC...), la rivista scientifica The Lancet e molte altre...

L'aspettativa di vita delle donne in Italia, e il ranking dell'Italia per essa, sarebbe[13]

86.49 – 3^ secondo l'ONU (luglio 2015), dati 2010-2015
85.6 secondo l'OECD (2016), che considera solo stati OECD
85.1 – 12^-13^ secondo The World Factbook (2017)
84.8 – 7^ secondo l'OMS (maggio 2016), dati 2015
83.90 – 7^ secondo (un articolo di) The Lancet (2012).

Abbastanza variabilità per fare affermazioni molto diverse, pure autorevoli, specialmente usando il dato del ranking, molto fuorviante: per esempio San Marino viene considerato solo nella lista del World Factbook.

[13]Si veda Wikipedia, l'enciclopedia libera, alla voce List of countries by life expectancy.

BOZZA - DRAFT

Gli standard variano nel tempo, e a seconda delle agenzie che li fissano, pur autorevoli.

Esempio. Attualmente (giugno 2019) leggiamo su sito governativo statunitense a proposito dell'anilina, in https://www.atsdr.cdc.gov/substances/toxsubstance.asp?toxid=79

> Cancer Classification: EPA: Probable human carcinogen. IARC: Not classifiable as to carcinogenicity to humans. NTP: Not evaluated

(Si veda per esempio il Link ->).

Altro esempio: leggiamo[14]

> several factors influence the likelihood of progression from prediabetes to Type 2 diabetes, including how we define prediabetes (there is no internationally agreed-upon standard)

## Varia nel tempo e nei luoghi cosa sia un farmaco...

...o piuttosto una droga (come l'eroina, "venduta liberamente dalla multinazionale farmaceutica Bayer dal 1899"[15] per varie patologie),

BOZZA - DRAFT

[14] Online text in https://www.levelshealth.com/blog/how-to-reverse-prediabetes read December 15, 2022.

[15] https://it.wikipedia.org/wiki/Eroina:

o un cancerogeno[16] (come la formaldeide: "Fino al momento in cui il prodotto è stato bandito, era uno degli ingredienti delle pastiglie Formitrol per il mal di gola"[17]).

---

L'intento era quello di ottenere una molecola più efficace della codeina nel sedare la tosse, la tubercolosi e le patologie respiratorie. (...) In breve tempo l'impiego terapeutico si ampliò alle più disparate patologie pneumologiche, ma anche neurologiche, ginecologiche, o a semplici dolori; si diffusero pertanto svariate preparazioni farmaceutiche acquistabili liberamente, e questo fece sì che l'eroina divenisse velocemente uno dei farmaci più venduti in assoluto.

[16] https://it.wikipedia.org/wiki/Formaldeide:

L'Associazione Internazionale per la Ricerca sul Cancro (AIRC) sin dal 2004 ha inserito la formaldeide nell'elenco delle sostanze considerate con certezza cancerogene per la specie umana. (...) Dal 1° gennaio 2016, la formaldeide è passata dalla classificazione di "sospettato di provocare il cancro" a "può provocare il cancro" (Regolamento (UE) N. 605/2014).

[17] https://it.wikipedia.org/wiki/Formaldeide

BOZZA - DRAFT







Le conoscenze scientifiche variano nel tempo, e con esse i vari standard, seppure in parte contraddittori, e in uno stesso tempo e luogo esistono

opinioni mediche opposte.






# Come cambiano nel tempo le opinioni mediche e farmaceutiche!

Un orbitoclasto è uno strumento chirurgico utilizzato per eseguire lobotomie transorbitali. Esso fu inventato dal dottor Walter Freeman nel 1948 come sostituto all'unica forma di leucotomo allora utilizzato per gli interventi di tal genere. Lo strumento in questione è a forma di punteruolo rompighiaccio, con alcuni segni incisi sul manico ed una struttura a cuneo sull'estremità esterna. L'operazione consiste nell'inserire lo strumento dietro la palpebra del paziente, perforare il sottile strato osseo che è posto inferiormente ad essa, e attraverso colpi di martello far penetrare lo strumento fino al lobo frontale. L'orbitoclasto viene poi mosso in varie direzioni al fine di distaccare i lobi frontali dal talamo. Nel 1948, Freeman perfezionò la procedura introducendo un profondo taglio frontale ed un movimento ulteriore della punta nel lobo, al fine di evitare - come succedeva occasionalmente in precedenza - che l'orbitoclasto si rompesse nella testa del paziente, necessitando così una rimozione chirurgica. In più, Freeman riuscì a commissionare la produzione di orbitoclasti di materiale più resistente ed affidabile.

Poiché durante i primi interventi venivano usati veri e propri punteruoli da ghiaccio, la procedura fu soprannominata "lobotomia con punteruolo da ghiaccio" ("ice pick lobotomy"). Nel 1949 fu assegnato il Premio Nobel per la Medicina a Egas Moniz, "per la sua scoperta del valore terapeutico della lobotomia in alcune psicosi". (Da Wikipedia, l'enciclopedia libera).

BOZZA - DRAFT




# NOW... Scientific Evidence on Effects of Smoking!

A MEDICAL SPECIALIST is making regular bi-monthly examinations of a group of people from various walks of life. 45 percent of this group have smoked Chesterfield for an average of over ten years.

After ten months, the medical specialist reports that he observed...

***no adverse effects on the nose, throat and sinuses of the group from smoking Chesterfield.***

***MUCH MILDER***

## CHESTERFIELD IS BEST FOR YOU

*First and Only Premium Quality Cigarette in Both Regular and King-Size*




CONTAINS TOBACCOS OF BETTER QUALITY AND HIGHER PRICE THAN ANY OTHER KING-SIZE CIGARETTE

Copyright 1952, Liggett & Myers Tobacco Co.

DRAFT

## 21.8 Col ranking facciamo ben figurare il Paraguay

Supponiamo che abbiamo questi dati, relativi a una cosa qualunque che i lettori di un giornale percepiscono in generale come *buona*:

USA 3723
Canada 3518
Paraguay 18
India 16
eccetera

Oggi vogliamo dire bene del Paraguay, a beneficio di qualche potentato, ovvero interesse geopolitico – il lettore acuto può trovarne uno – a cui il nostro giornale, per il quale scriviamo, risponde. Che fare, con quel misero valore 18? Potremmo dire che

il Paraguay è terzo dopo USA e Canada, davanti all'India.

BOZZA - DRAFT

## 21.9 Illusioni percettive nella presentazione dei dati

In questo paragrafo usiamo sia lo standard del punto decimale che della virgola decimale.

La dose giornaliera raccomandata di vitamina C è stata portata da 60 milligrammi a 90 milligrammi:

possiamo dire che adesso è del 50% maggiore

possiamo dire che prima era del 33% minore.

Ovviamente col primo modo diamo l'impressione di una variazione maggiore, almeno al lettore poco esperto di numeri.

L'epidemia di influenza del 2016/17 ha fatto circa $25\,000$ morti, circa il 28% in meno della prima ondata del covid-19 nel 2020 che ha fatto circa $35\,000$ morti, il 40% in più...

In linea generale, i titoli faranno più conseguenze dei testi sottostanti.

BOZZA - DRAFT

Se il partito dei Vispi Volpini un anno riceve lo 0,6% e l'anno dopo lo 0,9% (è un partito piccolo...) la televisione – se onesta – dovrebbe dire che è cresciuto *di 0,3 punti percentuali*, mentre il Corriere dei Vispi Volpini titolerà che il loro elettorato è *aumentato del 50%*. Hanno ragione entrambi:

$$0{,}9\% - 0{,}6\% = 0{,}3 \text{ punti percentuali} \qquad \text{(differenza assoluta)}$$

$$\frac{0{,}9 - 0{,}6}{0{,}6} = 0{,}5 = +50\% \text{ variazione percentuale} \qquad \text{(differenza relativa)}$$

Fuorviante invece, benché frequentissimo, dire "è cresciuto dello 0,3%": crescere dello 0,3% vorrebbe dire passare da 0,6% a $0{,}6 \cdot 1{,}003 \approx 0{,}602\%$.
(Magari anche un telegiornale dirà "+50%", chissà chi è il direttore...)

In Farmacia è la stessa cosa: se un effetto collaterale passa dall'1%

BOZZA - DRAFT

dei pazienti al 2%, il rischio aumenta di *1 punto percentuale* (rischio assoluto), ovvero *raddoppia*, $+100\%$ (rischio relativo). Chi vuole allarmare sceglierà la seconda formulazione, chi vuole rassicurare la prima.

N.d.A. Ricordo, decenni fa, un esempio di presentazione fuorviante dei risultati elettorali del tipo "il partito A scende dal ...% al ...%, il partito B prende il ...% contro il ...% della volta precedente" dove nel profluvio di cifre e decimali l'ascoltatore medio ricorda solo che A *scende* e B *prende*, ma in effetti B scendeva più di A...

Fare raffronti con dati collaterali può permettere di titolare "aumento" un articolo che poi parla di un fenomeno in diminuzione, ma moltissimi si limiteranno a leggere il titolo.

In ambito medico, affermare che con diagnosi precoci si allunga la speranza di vita dalla diagnosi può essere molto fuorviante: "Una volta il paziente viveva mediamente due anni dalla diagnosi, adesso dodici

BOZZA - DRAFT

anni"... ma facendo la diagnosi mediamente quanti anni prima?

Attenzione alle fallacie della memoria: vi pare di ricordare che secondo l'Organizzazione Mondiale della Sanità ogni anno 12.6 milioni di persone muoiono a causa dell'inquinamento, e adesso leggete che ogni anno 4.2 milioni di persone lasciano questo mondo a causa dell'inquinamento dell'aria? Basta rileggere con attenzione.

Attenzione soprattutto alle rappresentazioni grafiche.
Iniziare un diagramma a colonne o un grafico da un certo tempo invece che da un tempo precedente può permettere di fare affermazioni completamente diverse, magari entrambe formalmente vere, ma che genereranno emozioni e poi reazioni completamente diverse.
Grafici che sull'asse delle ordinate non iniziano da 0 possono dare l'impressione di variazioni enormi, anche raddoppi o dimezzamenti, a variazioni minuscole.
Diagrammi a torte tridimensionali in rappresentazione prospettica pos-

BOZZA - DRAFT

sono far apparire più grandi certi dati rispetto agli altri.

## 21.10 Di cosa parliamo?

Diamo solo un cenno di un argomento ciclopico. Ogni persona vive immersa in un flusso di informazioni che gli arrivano, in base al quale costruisce la sua percezione della realtà, che lui crede vera.
Anche se tutte quelle informazioni fossero vere (per ipotesi assurda, ovvio) l'immagine complessiva che l'individuo si forma è in generale (molto!) distorta: una *diversa scelta* di informazioni *altrettanto vere*, da far arrivare coi media – dal livello più basso fino agli articoli scientifici – creerebbe in quella stessa persona un'immagine del mondo completamente diversa, per molti versi quasi opposta.

BOZZA - DRAFT

Supponiamo qua in via ipotetica che sul pianeta Kepler-22 b, simile alla Terra, il danno, in qualche modo misurato, delle sigarette si disponga a campana gaussiana, con un'ampia zona mediana di un

danno medio per una grande generalità di fumatori. E che il danno delle sigarette elettroniche si disponga analogamente, ma più a sinistra sull'asse delle ascisse, cioè con un danno mediamente minore. Per quanto il danno sia mediamente minore, in moltissime persone è grave, per molte addirittura gravissimo, per alcune perfino maggiore di quello medio delle sigarette. Ecco, dai, invece di fare alcuni milioni di articoli su quelli mediamente meno danneggiati (impossibile a farsi), e invece di dire che in media il danno è minore (chi mai ci obbliga?) facciamo un bell'articolo titolando "ragazzo fuma sigarette elettroniche e ha i polmoni devastati".

La gente dirà: ecco, vedi, non serve, soldi buttati, meglio continuare con le sigarette normali!

**È impressionante quante persone ragionino per casi singoli, appresi dai media, senza garanzia che siano rappresentativi di situazioni medie invece che eccezionali.**

BOZZA - DRAFT

## 21.11 Cherry picking – bias di selezione

Affine alla questione precedente. Scegliere le prove favorevoli a una tesi e ignorare quelle contrarie. Se un proprietario di media volesse sostenere la tesi che bisogna chiudere le scuole per evitare il propagarsi del covid, potrebbe cercare di citare affermazioni favorevoli alla sua tesi, al livello autoritativo maggiore possibile, trascurando le prove contrarie; e similmente se vuole far passare la tesi opposta. Troverà articoli scientifici per qualunque dei 2 desiderata. L'articolo di review sistematica in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8375447/ considera decine di articoli scientifici di alto livello (dopo averne scartati migliaia di altri) e verifica che metà trovano che la chiusura delle scuole ha ridotto la diffusione del contagio, e metà hanno trovato che non l'ha ridotto.

BOZZA - DRAFT

## 21.12 Approfittare delle fluttuazioni statistiche




BOZZA - DRAFT

Mentre le funzioni elementari di base della Matematica hanno andamenti molto regolari, come $x^2$ o $x^3$, e similmente le successioni propriamente dette come quella di Fibonacci e quella dei fattoriali, le successioni di valori numerici rappresentativi di fenomeni reali normal-

mente presentano nel tempo fluttuazioni, alternando crescite a diminuzioni.

Esempio, morti in Italia:

Esempio, morti in Italia (dati ISTAT, arrotondati alle migliaia):

2013 $601\,000$

2014 $598\,000$ $-3\,000$ rispetto all'anno precedente

2015 $648\,000$ $+50\,000$

2016 $615\,000$ $-33\,000$

2017 $649\,000$ $+34\,000$

2018 $633\,000$ $-16\,000$

2019 $634\,000$ $+1\,000$ ("Aumentano i morti in Italia!" Sì, dello 0.2% rispetto al 2018).

(Si noti che anche dividere per il numero di abitanti non basterebbe: la popolazione invecchia, e i morti possono aumentare anche se, *a parità di età*, si muore sempre meno. Per questo in Demografia si usano i *tassi standardizzati per età*.)

BOZZA - DRAFT

Consideriamo come ulteriore esempio il grafico *Sperimentazioni autorizzate dall'Autorità competente* a pagina 10 del documento La Sperimentazione Clinica dei Medicinali in Italia – 17° Rapporto Nazionale – Anno 2018 dell'AIFA, Agenzia Italiana del Farmaco. L'ultimo dato è del 2017 e relativamente ad esso potremmo dire *veridicamente*

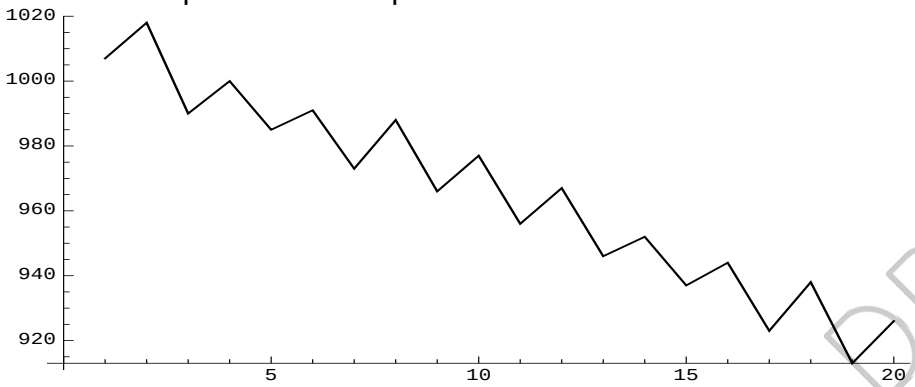
aumentate rispetto al 2000
aumentate rispetto a 15 anni prima
diminuite rispetto all'anno precedente
al livello minimo da 10 anni

ottenendo nel lettore emozioni diverse, che potranno supportare interventi decisionali diversi.

Si noti che anche in situazioni di palese diminuzione sostanziale sul lungo periodo, tali fluttuazioni possono permettere in vari anni di parlare di aumento. E negli anni in cui proprio non si può dirlo... si può parlare della situazione dell'anno precedente, che magari vedeva un aumento rispetto all'anno precedente – mica il lettore sa qual è l'ultimo dato rilasciato, in generale. Agendo così, si potrà *ogni anno* parlare di

BOZZA - DRAFT

*aumento* in questo esempio di evidente *sostanziale diminuzione*:




## 21.13 Capziosità verbali

Un'intera enciclopedia si potrebbe scrivere, di infiniti volumi, sulle capziosità verbali, e, tranne che per la Medicina e la Farmacia, in cui c'è da piangere, sarebbe molto divertente.

Leggiamo per esempio in un articolo[18] online:

[18] https://www.today.it/cronaca/trivelle-terremoto-oggi-9-novembre-2022-marche.html letto il 24 novembre 2022

BOZZA - DRAFT

(Titolo)

perché le trivelle non c'entrano niente con il terremoto nelle Marche

(Sottotitolo)

Le insinuazioni dei contestatori dell'estrazione di gas sul territorio italiano e la spiegazione dell'esperto

(Corpo dell'articolo)

(...) ha chiarito che "la trivellazione di per sé non comporta alcuno scuotimento sismico, sono semmai le attività di estrazione e stoccaggio che possono farlo. (...)

Insomma i contestatori delle trivelle sono fuori strada: che si trivelli pure, come *divertissement*, casomai si contestino le successive "attività di estrazione e stoccaggio".

BOZZA - DRAFT

## 21.14 Illustrazione di un triplice esempio reale

Aumento dell'aspettativa di vita negli USA: `link<-`

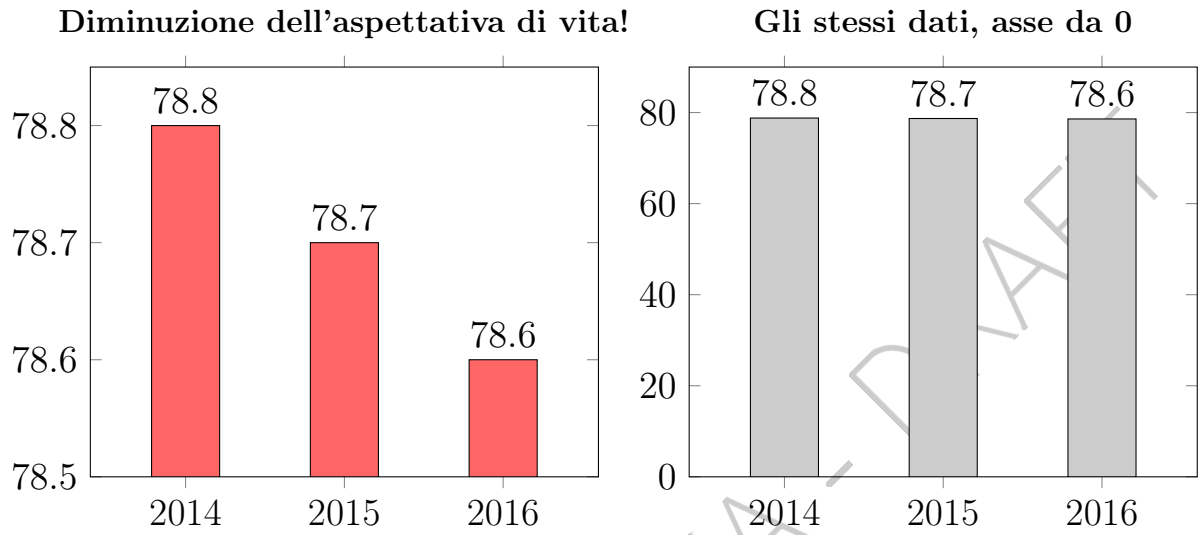



Figura 5: Aspettativa di vita alla nascita negli USA, in anni (dati CDC/NCHS). A sinistra l'asse verticale parte da 78.5: un calo di 2 decimi di anno in 2 anni (circa lo 0.25%) appare come un crollo, e l'ultima barra sembra alta un terzo della prima. A destra gli stessi 3 numeri con l'asse che parte da 0: le barre sono praticamente uguali. Entrambi i grafici sono *veri*.

BOZZA - DRAFT

Aumento del tasso di mortalità negli USA:

2009 ..... 7.9

```
2010 ..... 8.0
2011 ..... 8.1
2012 ..... 8.1
2013 ..... 8.2
2014 ..... 8.2
2015 ..... 8.4
2016 ..... 8.5
2017 ..... 8.6
```

Capire questi diagrammi richiede:

1) conoscere la Statistica Descrittiva, notando in particolare che
   il diagramma considera solo gli ultimi 3 anni;
   il diagramma ha una linea base posta a circa 78.5 anni.

1 bis) conoscere la Demografia, che possiamo considerare semplificatamente una branca specialistica della Statistica Descrittiva. Questo ci farebbe capire che non è per nulla strano che si alzino da tempo i tassi di mortalità, anche prima della riduzione dell'aspettativa di vita. Il tasso di mortalità può ben alzarsi mentre l'aspettativa di vita si alza, proprio perché la popolazione mediamente invecchia, in compresenza di un fattore demografico completamente diverso: la scarsità di nuovi nati.

2) Conoscere la <u>situazione reale</u> il che è completamente diverso dal

BOZZA - DRAFT

guardare questi numeri o grafici.
Per una parziale spiegazione della situazione reale dietro alla diminuzione dell'aspettativa di vita negli USA si segua questo link<-.

## 21.15 Conclusioni sulla fallace presentazione di dati veri

Pensare di aver capito una situazione perché si è vista una qualche statistica o lista di numeri è in generale *ingenuo*, e i grafici sono, per le persone non istruite, perfino più fuorvianti. Le parole, poi, possono essere ancora più ingannevoli anche se sembrano esprimere dati oggettivi, numerici: aumento, diminuzione...

È necessario acquisire una buona competenza statistica per non essere tratti in inganno. Sia da errori fatti in buona fede, che da manipolazioni volontarie, fatte da soggetti interessati.

**Superato il problema della comprensione formale di sta-**

BOZZA - DRAFT

**tistiche e grafici, il problema**
**delle statistiche**
**e dei grafici,**
**e perfino dei dati grezzi/completi,**
**è che ce ne possono sempre essere altri, che non stiamo vedendo ovvero non ci stanno mostrando, che cambierebbero completamente la nostra prospettiva sui fenomeni considerati.**

In linea generale, le statistiche economiche dei media non valgono la carta su cui sono scritte. Per le statistiche scientifiche, non è vietato sperare una maggiore oggettività.

Molti testi divulgativi affermano che *la maggior parte* dei contagi da toxoplasmosi non avvengono dai gatti ma dalla sporcizia. Anche se questa statistica – praticamente impossibile da verificare – fosse vera, quegli Autori magari trascurano di dire, nei loro articoli divulgativi, che sebbene il toxoplasma si trovi in tanti animaletti,

BOZZA - DRAFT

si riproduce solo nei felini, per cui se tutti i gatti venissero inceneriti (orrore!) la toxoplasmosi semplicemente sparirebbe dall'Italia, dove praticamente l'unico felino che frequenta l'uomo è il gatto.

I gattofili non costituiscono una lobby molto pericolosa e allora qua possiamo prendercela scherzosamente con loro senza rischi, per spiegare cosa vuol dire comunicare una statistica fuorviante perché omette fatti che cambierebbero radicalmente la prospettiva.
Ma la triste verità per uno statistico è che le notizie di statistica medica che arrivano al grande pubblico dalla televisione e dai media *lasciano a desiderare*, diciamo così.

Ulteriori problematiche si presentano nella Statistica Inferenziale.

*If you torture the data long enough, it will confess to anything*. (In questo `link<-` è in un articolo su sito governativo statunitense).

**Esempio finale.** Da La Peste, aprile 2022, su Facebook:

BOZZA - DRAFT

> Il primo gennaio 2015 sono decedute 126 persone di 95-99 anni, mentre il primo gennaio 2021 ne sono decedute 198; 72 in più? Per fare il confronto corretto dobbiamo considerare come è variata la numerosità della popolazione di persone di 95-99 anni. Queste erano 88.799 al primo gennaio 2015, ma sono diventate 158.581 al primo gennaio 2021. Con una variazione così imponente non si può fare un confronto tra le cifre assolute. Il problema si può risolvere considerando la mortalità. Il primo gennaio 2015 sono morte lo 0,142% delle persone di 90-95 [*sic*: 95-99] anni mentre il primo gennaio 2021 ne sono morte lo 0,125%. La mortalità è un po' inferiore.

A seconda del risultato che si vuole ottenere, si potrà dire che ne sono morti di più, o di meno, dicendo sostanzialmente il vero.

BOZZA - DRAFT

# PARTE II – DATI QUASI VERI (QUASI FALSI)

Già abbiamo visto nella Lezione 18 le problematiche del trasformare dati ordinali in dati numerici. Nella pratica, se abbiamo già il dataset ordinale ma siamo ancora liberi nel fissare le corrispondenze numeriche, possiamo largamente manipolare il risultato.

## 21.16 Il valore anomalo: outlier

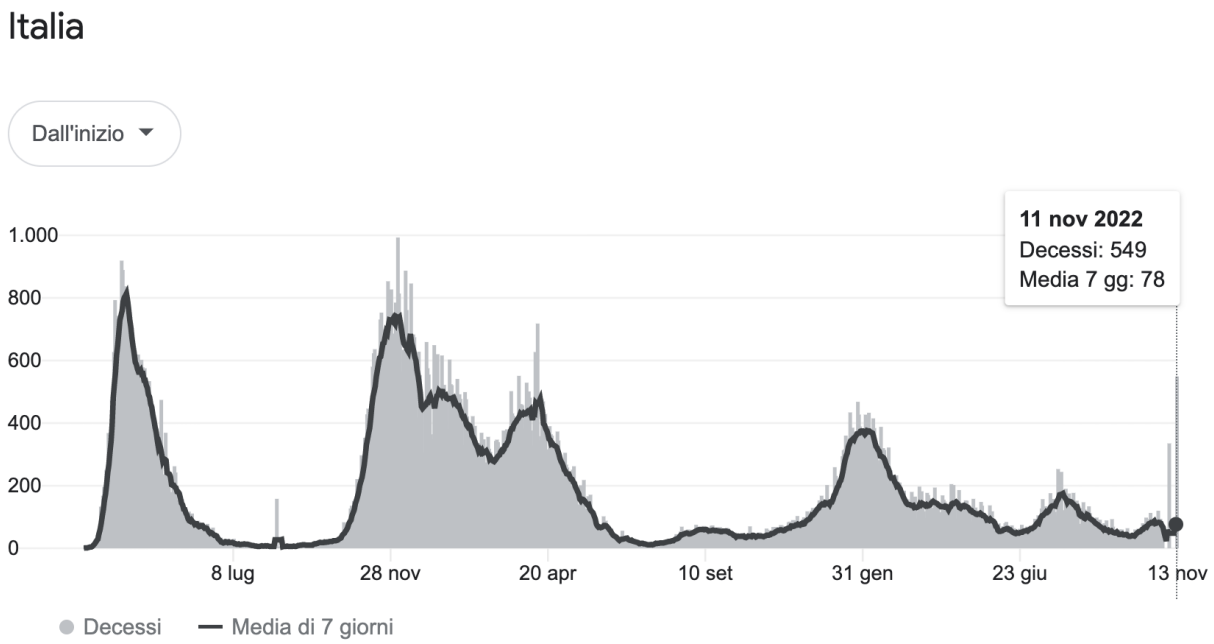



Figura 6: In questa peraltro autorevole statistica – screenshot da Google – l'esorbitante numero di 549 morti dell'11 novembre 2022 è dovuto al fatto che il Ministero ha cambiato modo di comunicare i dati: quei morti non sono di un solo giorno bensì 7.

Per un'infinità di motivi, in un dataset ci può essere qualche valore che il ricercatore ritiene *assurdo*, e tenderà ad eliminarlo prima di procedere con qualunque analisi dei dati stessi, per non comprometterla.

Questi *outlier*, valori anomali ovvero aberranti, provengono da risposte scherzose a questionari, truffe ("reddito zero"...), malfunzionamento di apparati, confusione fra punto decimale e virgola decimale, confusione fra milligrammi e microgrammi, eccetera. Ma talvolta sono dati *veri* benché *eccezionali*.

La decisione di quali valori siano da considerare outlier ed eventualmente da eliminare, può essere

soggettiva

oppure dettata da formule, che comunque non considereremo,

e in ogni caso è problematica, anche perché ce ne sono varie.
In una ricerca seria *et* onesta, si *dovrebbe* avvertire se sono stati eliminati outlier.

BOZZA - DRAFT

Leggiamo su Wikipedia, l'enciclopedia libera, alla voce "Outlier", e non è chi non veda le pericolose conseguenze di ciò:

> There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. There are various methods of outlier detection. (...) Deletion of outlier data is a controversial practice

## 21.17 Omissione di dati ritenuti poco significativi

Dati ritenuti poco significativi vengono talvolta omessi, per esempio San Marino da molte statistiche; facendoli rientrare si possono ottenere affermazioni molto diverse, ma per certi versi effettivamente meno significative.

BOZZA - DRAFT

## 21.18 "Fatta la legge, trovato l'inganno"

Fissato uno standard per valutare statisticamente un qualche aspetto della realtà, qualcuno trova conveniente, diciamo così, fare azioni che migliorano di molto la valutazione statistica con scarso o nullo miglioramento della realtà. Se per esempio una legge attribuisce ai *"comuni montani"* particolari finanziamenti, fissando ad una certa altitudine del municipio la denominazione di "comune montano", state sicuri che qualche comune sposterà la sede del municipio, più a monte... *et voilà*, il comune – esattamente lo stesso di prima – diventa "montano" (e ottiene i finanziamenti).

In ambito scientifico e biomedico in particolare, a rischio, diciamo così, sono le statistiche che valutano la *credibilità percepita* ("impact factor") delle riviste scientifiche; si veda per esempio *A user's guide to inflated and manipulated impact factors*, di John P. A. Ioannidis e Brett D. Thombs, European Journal of Clinical Investigation (17

BOZZA - DRAFT

giugno 2019) Link->

## 21.19 Traslazioni e confusioni linguistiche

Spesso le statistiche vengono presentate con quelle che potremmo considerare "traslazioni di linguaggio".
Per necessità di sintesi, soprattutto nei titoli – a non voler pensare male.

I cani non sono la terza causa di morte nel mondo, no. Sono, più precisamente, la terza causa di morte *fra quelle causate da animali*, dopo zanzare e serpenti.

Titolo (vero, al momento ancora reperibile in rete, ottobre 2020):

Piene al 150% le rianimazioni in Abruzzo, prima Regione in Italia ad esaurire i posti letto

BOZZA - DRAFT

Più discretamente la prestigiosa agenzia di stampa ANSA, nel testo di un articolo, non nel titolo, e con qualche parola in più:

> Secondo il report settimanale dell'Alta Scuola di Economia e Management dei Sistemi Sanitari dell'Università Cattolica, campus di Roma (Altems) l'Abruzzo ha saturato il 150% dei posti letto aggiuntivi implementati

Dunque, vediamo che hanno saturato dei posti letto *aggiuntivi*, non quelli ordinari che avevano già prima della crisi.
Un'attenta analisi delle fonti che nessun normale lettore farà rivela che si tratta di 10 posti in terapia intensiva – non sappiamo quanti eventualmente per covid, di cui parlano gli articoli considerati, o incidenti stradali o altro – su 7 di certi posti definiti "aggiuntivi", aggiunti per la crisi covid, mentre *molti più di 100 restavano non occupati*.

È stato tremendo, ho fatto 3 ore di straordinario
(Nelle 8 ore lavorative ordinarie riposavo)

BOZZA - DRAFT

Naturalmente questi titoli e queste notizie, oltre all'incalcolabile danno di ansia e paura che possono procurare alle persone fragili, manipolano il consenso popolare verso una o l'altra azione politica, che decreterà questo o quell'altro *spostamento di risorse* o, addirittura, determinate limitazioni alla vita delle persone.

**Un esempio documentato del danno della paura.**
Uno studio della Società Italiana di Cardiologia[19] ha confrontato una settimana di marzo 2020 (12-19 marzo, in pieno allarme covid) con la stessa settimana del 2019, in molti ospedali italiani:

| | 2019 | | 2020 |
|---|---|---|---|
| ricoveri per infarto | 618 | 319 | $(-48\%)$ |
| mortalità degli infarti STEMI ricoverati | $4.1\%$ | $13.7\%$ | $(\times 3.3)$ |

Gli infarti non erano spariti: i pazienti arrivavano in ospedale più tardi (il tempo fra i primi sintomi e la coronarografia è aumentato del $39\%$), o non arrivavano affatto, verosimilmente anche per paura di contagiarsi in ospedale. E un infarto trattato tardi uccide molto di più.

Si noti anche la lezione statistica: un titolo come "Dimezzati gli infarti durante il lockdown" sarebbe stato formalmente vero per i *ricoveri*, e completamente fuorviante.

BOZZA - DRAFT

[19] De Rosa S, Spaccarotella C, Basso C, et al. *Reduction of hospitalizations for myocardial infarction in Italy in the COVID-19 era.* European Heart Journal 2020;41(22):2083–2088. doi:10.1093/eurheartj/ehaa409.

## 21.20 Il problema della discrezionalità

Nella rilevazione dei dati può essere presente un ampio margine di discrezionalità: classificare al microscopio cellule in "regolari/irregolari", "simmetriche/asimmetriche"...

Esempio lampante di discrezionalità:[20]

> "Moribund larvae were considered dead and included in the analyses."

## 21.21 Esempio importante: le cause di morte

Il referto sulle cause di morte presenta una certa discrezionalità.

Le persone in generale giungono alla morte con diverse patologie con-

BOZZA - DRAFT

[20] Nartey, Rita et al. "Use of Bacillus thuringiensis var israelensis as a viable option in an Integrated Malaria Vector Control Programme in the Kumasi Metropolis, Ghana." Parasites & vectors vol. 6 116. 22 Apr. 2013, doi:10.1186/1756-3305-6-116

comitanti (si veda *comorbidità*).
E alcune muoiono pure dissanguate in incidenti automobilistici, in realtà causati dalle loro patologie (non è il dissanguamento la "vera" causa, a monte). Si pensi per esempio all'incidente automobilistico da crisi ipoglicemica.

Freddie Mercury è morto di polmonite (causa formale) o di AIDS (causa sostanziale)?
Anche alcol e droga forse non gli hanno fatto molto bene. Ma la causa di morte non viene ripartita in percentuali.

Il problema diventa critico per concomitanti patologie entrambe registrabili come causa di morte. Se sul piano di verità entrambe hanno favorito la morte, nelle statistiche ufficiali una sola l'ha causata, e quelle statistiche hanno conseguenze sulle politiche sanitarie e farmaceutiche – magari con la mediazione dei media.

BOZZA - DRAFT

Si pensi all'enorme difficoltà che può esserci dietro questa scelta:

morto per leucemia ma con morbillo
morto con leucemia ma per morbillo.

La decisione di cosa scrivere spetta a persone che sono generalmente molto competenti, tuttavia la situazione talvolta può essere complessa.

A proposito del Coronavirus un virologo avrebbe (2020) affermato:

> Tutte e 7 le vittime avevano anche altre patologie, quindi sarebbe più opportuno parlare di Coronavirus come concausa non come causa diretta.[21]

Molto problematico da un punto di vista statistico questo passaggio di Repubblica:

> Si tratta in gran parte inoltre - viene precisato - di pazienti contagiati dal virus nei mesi scorsi, nel frattempo negativizzatisi, ma che su indicazione del Ministero della Salute vanno registrati comunque come soggetti con infezione da Covid.

Per capire meglio la complessità della questione, su un sito internet del Ministero della Salute leggiamo[22]

BOZZA - DRAFT

[21] Matteo Bassetti, direttore della clinica di malattie infettive dell'ospedale policlinico San Martino di Genova, citato in https://www.primocanale.it/notizie/vittime-coronavirus-l-esperto-con-l-influenza-probabilmente-sarebbero-morti-lo-stesso--216491.html, letto il 28 febbraio 2020

[22] In https://www.epicentro.iss.it/influenza/sorveglianza-mortalita-influenza, accesso il 9 marzo 2020.

> il virus influenzale non viene identificato o perché **non ricercato** o perché il decesso viene attribuito a polmoniti generiche.

(Enfasi aggiunta).

L'attribuzione di una causa di morte piuttosto che un'altra può avere grandi ripercussioni nelle decisioni di politica sanitaria, sostenuta dalle istanze dei cittadini, mediate o indirizzate dai media.

*

Nel 2026 la procura di Siena ha riaperto il caso di David Rossi, il capo comunicazione del Monte dei Paschi precipitato dalla finestra del suo ufficio nel 2013 e la cui morte era stata archiviata due volte come suicidio. Nuove perizie, anche sui video di sorveglianza noti da tempo, che mostrano persone nel vicolo, fanno ora pensare a un omicidio. Il filmato della telecamera che riprese la caduta è agli atti dal 2013. All'epoca fu acquisito il video di una sola telecamera, pur essendocene una dozzina nella sede, e da anni si discute delle figure che compaiono all'imbocco del vicolo

*

Nel 2022 fu ritrovata a Trieste a pochi metri dall'Università (precisamente presso il Museo dell'Antartide) una donna[23] morta dentro due sacchi neri della spazzatura, con un sacchetto sulla testa chiuso da un laccio. La procura chiese di archiviare il caso come suicidio, ma il giudice si oppose: una nuova autopsia ha poi indicato

[23] Lliana Resinovich.

BOZZA - DRAFT

un omicidio, e dal 2025 si indaga per omicidio, caso (nel 2026) ancora aperto. D'altra parte, nel 2010 un agente[24] dei servizi britannici, fu trovato morto in un borsone chiuso con un lucchetto, nella vasca del suo appartamento londinese. Per la polizia fu un incidente. E nel 2004 un giornalista[25] dell'inchiesta sulla CIA e il crack fu dichiarato suicida nonostante avesse due colpi di pistola alla testa. Ci sono moltissimi altri casi strani.

## 21.22 Cose vere in articolo titolato oppostamente

In un articolo divulgativo, per esempio su un quotidiano, si possono riportare correttamente i risultati di uno studio scientifico, così nessuno può obiettare, mettendo però un titolo che dice l'opposto, così nessuno può obiettare comunque, perché il titolo l'ha messo il titolista, mica l'autore che firma l'articolo – e senz'altro gliela si perdona. Ovviamente il titolo farà immensamente più effetto del corpo dell'articolo.

[24] Gareth Williams.

[25] Gary Webb (2004).

BOZZA - DRAFT



**Covid, effetto boomerang; a ogni reinfezione conseguenze più gravi**

di Fabio Di Todaro



*Il lavoro di un gruppo di ricercatori dell'Università di Washington pubblicato s possibili conseguenze di infezioni plurime da coronavirus: ad ogni nuovo cont incontro a complicanze*

11 NOVEMBRE 2022 AGGIORNATO ALLE 09:44

In assoluto determinano quadri meno gravi rispetto a quelli relativi al primo contagio, ma le reinfezioni da Sars-CoV-2 non è detto che evolvano sempre senza particolari conseguenze. Ragion per cui continuare ad adottare le misure di protezione non

Figura 7: Fotografia di la Repubblica online, 11 novembre 2022. Si legga la prima riga dell'articolo.

BOZZA - DRAFT

# PARTE III – DATI FALSI

## 21.23 La falsificazione dei dati

Finora abbiamo considerato solo dati *veri*, magari presentati con omissioni o in modo fuorviante ma comunque veri: nella realtà pratica, esiste anche la falsificazione dei dati.

### Falsificazioni negli articoli divulgativi

Per dare solo una vaga idea del livello stratosferico di menzogna continua – costantemente monitorata dallo scrivente – della generalità dei media italiani (l'Italia è internazionalmente classificata nel 2022 al 58esimo posto nel mondo per libertà di stampa) durante la pandemia del covid-19, si consideri il caso della Svezia, che ha tenuto una politica anticovid diversissima dall'Italia, e in particolare non ha mai (2020 - OTT 2022) chiuso le scuole dell'obbligo: la Svezia veniva presentata come un luogo in cui il covid faceva strage, molto più che in Italia; il lettore interessato troverà in rete che la politica anticovid della Svezia veniva addirittura assimilata a un sacrificio umano. La verità fattuale è opposta:

– la Svezia ha costantemente avuto una mortalità covid molto più bassa dell'Italia;

– questo fatto non può essere liquidato dicendo che là non hanno contato bene i morti per covid, perché anzi, la Svezia ha avuto un eccesso di mortalità *per tutte le cause* rispetto agli anni precedenti minore dell'Italia.

BOZZA - DRAFT

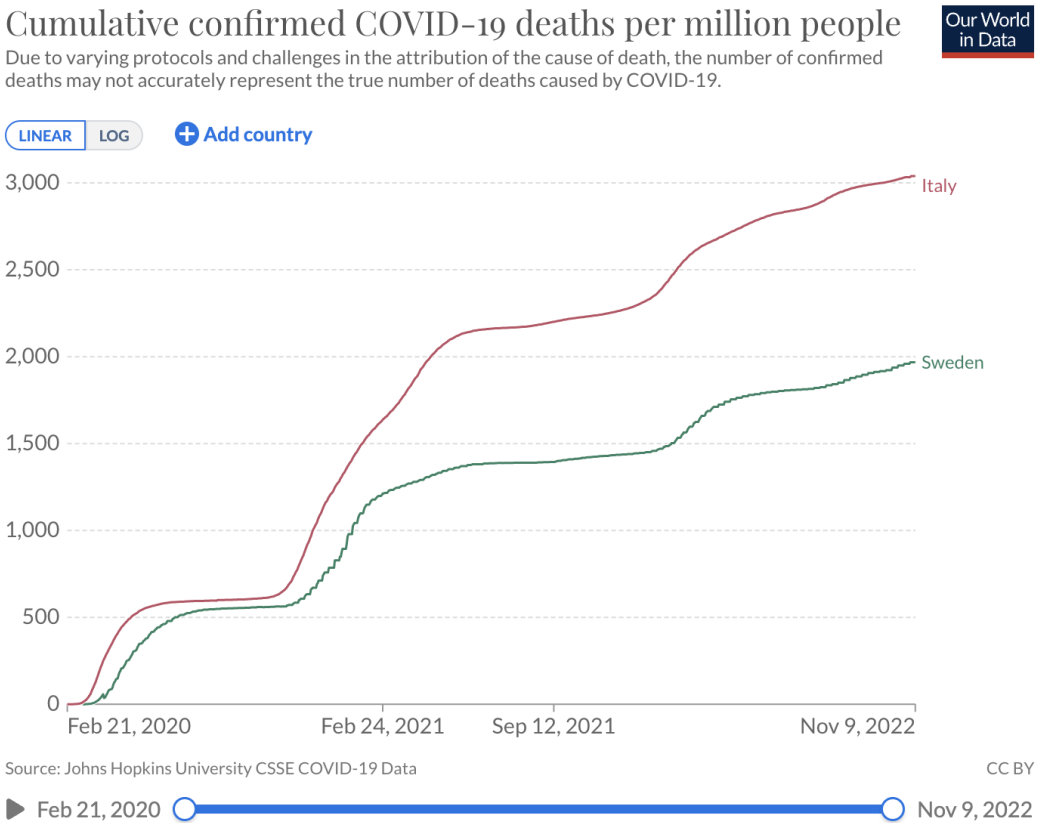



Figura 8: https://ourworldindata.org/covid-deaths?country=ITA~SWE, screenshot.

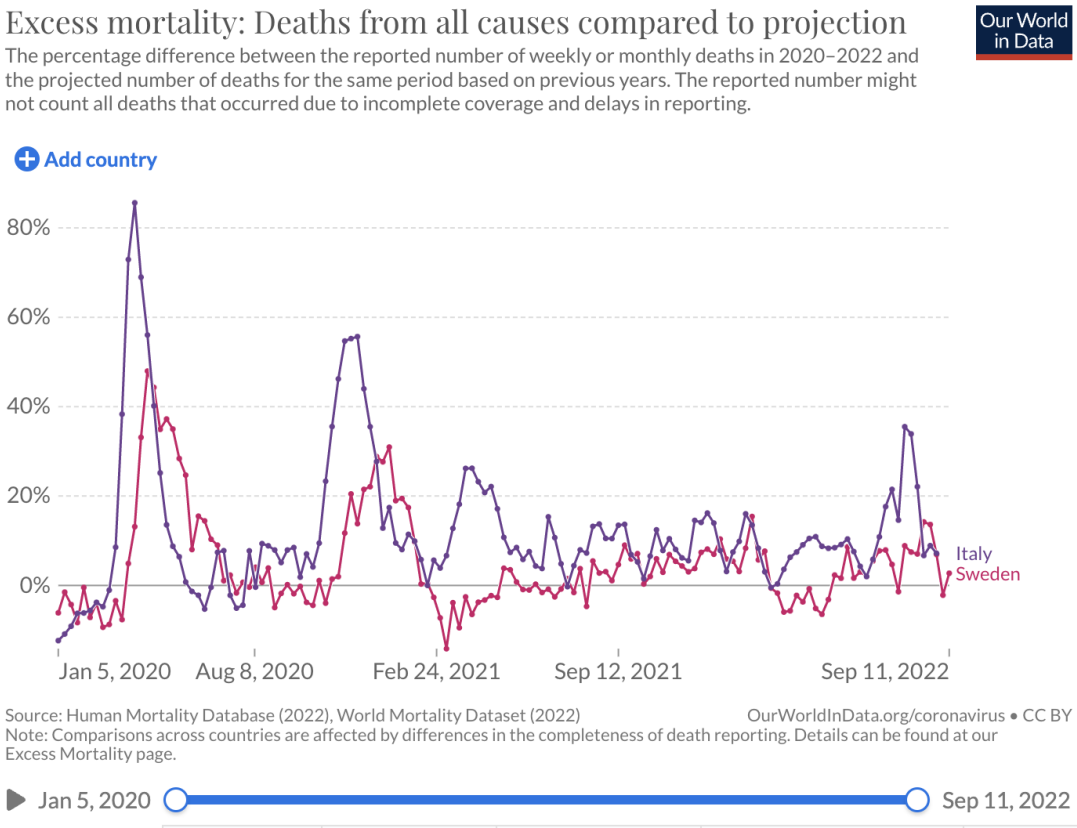



Figura 9: https://ourworldindata.org/excess-mortality-covid, screenshot.

## E negli articoli scientifici è legale mentire?

Leggiamo[26]

> Giugno 2019 [...] Scienziati impegnati nella ricerca sul cancro hanno manipolato le immagini dei loro studi, riuscendo così a ottenere successo, carriera, nuovi fondi per le loro ricerche. La Procura (...) ha appena concluso un'indagine lunga e complessa, che fornisce un quadro devastante

e volendo troviamo[27] molti dettagli fra cui

> la procura deve archiviare (...) non è reato

Dal punto di vista giuridico, le cose stanno nei termini seguenti.

Le *bugie* possono essere eventualmente immorali, ma non sono illegali, tranne in alcune fattispecie elencate dalla legge: il falso in atto pubblico, il falso in bilancio, il falso ideologico, eccetera, ma assolutamente non è compreso il falso in articolo

BOZZA - DRAFT

[26] In https://www.ilfattoquotidiano.it/in-edicola/articoli/2019/06/30/ricerca-sul-cancro-risultati-ritoccati-per-ottenere-milioni/5291204/

[27] https://codacons.it/ricerca-sul-cancro-risultati-ritoccati-per-ottenere-milioni/

scientifico, che quindi è assimilabile a una bugia fra privati cittadini – anche se l'articolo contiene descrizioni di esperimenti mai avvenuti, cioè non semplici *errori* ma falsità scritte con dolo.

Naturalmente, un cittadino che ritenesse che dalla pubblicazione di affermazioni false in un articolo scientifico gliene sia conseguito un danno economico, o un danno biologico, potrebbe ipotizzare una querela, o una denuncia – ma si capisce bene che in generale sarà impresa ardua, specialmente per articoli scientifici che non dicono bene di farmaci specifici, ma facciano *Scienza di Base*.

Tuttavia, l'autore truffaldino rischia pur sempre qualcosa.

Se egli è, come avviene quasi sempre, dipendente di un'istituzione scientifica (per esempio un'università), o anche solo affiliato scientificamente ad essa, è ben probabile che la pubblicazione di dati deliberatamente falsificati sia contraria al codice deontologico dell'istituzione. Allora al falsario l'istituzione può comminare una sanzione disciplinare in base al proprio regolamento, per violazione del codice deontologico (comunque solo in base a regolamenti e non a leggi, e quindi ad un livello molto più basso: le pene detentive sono escluse per la violazione di regolamenti).

Inoltre se la falsificazione diviene di dominio pubblico, l'istituzione può rivalersi

BOZZA-DRAFT

contro il truffaldino per avere il risarcimento per danno di immagine - con una denuncia o querela (e quindi al livello di legge).

Si noti comunque che mentre la Giustizia tende a perseguire le violazioni delle leggi (cioè i reati), l'istituzione scientifica avrebbe in generale tutto l'interesse a occultare o almeno a minimizzare la violazione del regolamento (avvenuta con la falsificazione dei dati) per evitare un danno d'immagine, e ciò fa intuire come verosimilmente andrebbero le cose.

BOZZA - DRAFT

## 21.24 Legami finanziari e risultati della ricerca farmacologica

Quando un ricercatore valuta un farmaco e ha rapporti economici con il suo produttore, il suo interesse finanziario può entrare in conflitto con quello di valutare il farmaco imparzialmente. Quanto spesso, in queste circostanze, una sperimentazione riporta un risultato favorevole?

Ahn e colleghi hanno esaminato 195 sperimentazioni cliniche randomizzate sull'efficacia di farmaci, pubblicate nel 2013. La tabella 1 distingue i trial in cui almeno un ricercatore principale aveva legami finanziari con il produttore del farmaco da quelli in cui gli autori dell'indagine non hanno rilevato tali legami. Fra i legami considerati figuravano consulenze retribuite, compensi per conferenze e partecipazioni azionarie. Ogni unità contata è un trial.[28]

BOZZA - DRAFT

[28] R. Ahn et al., *Financial ties of principal investigators and randomized controlled trial outcomes: cross sectional study*, BMJ 356 (2017), i6770. Articolo e dati originali.

Tabella 1: Legami finanziari dei ricercatori principali ed esito dei trial. Conteggi ricostruiti dai dati di Ahn et al.

| Legami con il produttore | Esito positivo | Esito negativo | Totale |
|---|---|---|---|
| Presenti | 103 | 29 | 132 |
| Nessuno rilevato | 33 | 30 | 63 |
| Totale | 136 | 59 | 195 |

BOZZA - DRAFT

## Quale denominatore?

Scegliamo a caso un trial fra i 195. Indichiamo con $L$ la presenza di legami finanziari e con $P$ l'esito positivo. Se sappiamo che vi sono legami, restringiamo l'attenzione ai 132 trial della prima riga. Se sappiamo che l'esito è positivo, consideriamo invece i 136 trial della prima colonna. Il numeratore resta lo stesso:

$$\Pr(P \mid L) = \frac{103}{132} \simeq 78{,}0\%, \qquad \Pr(L \mid P) = \frac{103}{136} \simeq 75{,}7\%.$$

Nell'abstract si confrontano la percentuale di trial con legami fra i positivi (103/136, circa il 76%) e la percentuale corrispondente fra i negativi (29/59, circa il 49%). La figura 10 risponde alla domanda inversa, raggruppando prima i trial secondo la presenza di legami.

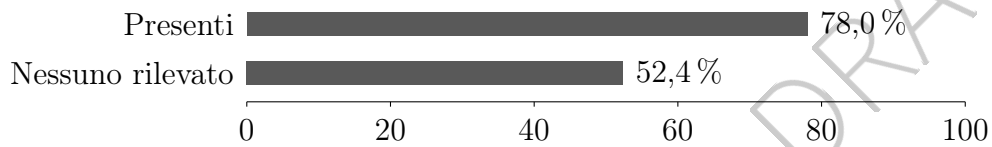



Figura 10: Percentuale di trial con esito positivo in ciascun gruppo. La barra superiore rappresenta 103 trial positivi su 132, quella inferiore 33 su 63. La scala comune parte da zero e arriva al 100%.

BOZZA - DRAFT

## Tre modi di misurare la differenza

Poniamo $p_1 = 103/132$ e $p_0 = 33/63$. La differenza $p_1 - p_0$ vale circa $0{,}256$, ossia **25,6 punti percentuali**. Il rapporto $p_1/p_0 \simeq 1{,}49$ indica che la frequenza relativa di trial positivi nel primo gruppo supera di circa il $49\%$ quella del secondo. Possiamo anche confrontare le *odds*,

ossia i rapporti fra casi positivi e negativi all'interno di ciascun gruppo:

$$\mathrm{OR} = \frac{103/29}{33/30} \simeq 3{,}23.$$

Dire che le odds sono più che triple e dire che la frequenza dei positivi è circa una volta e mezza quella dell'altro gruppo sono dunque affermazioni compatibili. La randomizzazione assegnava i pazienti ai trattamenti. Il confronto fra trial con diversi legami finanziari è osservazionale e da solo non determina la causa della differenza.

## 21.25 Chi finanzia lo studio?

Il legame personale di un ricercatore con un'azienda e il finanziamento di un intero studio sono due caratteristiche distinte. La revisione Cochrane di Lundh e colleghi permette di studiare la seconda.[29]

BOZZA - DRAFT

[29] A. Lundh et al., *Industry sponsorship and research outcome*, Cochrane Database of Systematic Reviews (2017), MR000033. Revisione e diagrammi originali, identificati dalle diciture *Analysis 1.1, 1.2, 2.1* e *3.1.*

## Finanziamento industriale e non industriale

Per ciascun esito si confronta la frequenza di studi favorevoli al prodotto nel gruppo a finanziamento industriale con la frequenza corrispondente nell'altro gruppo. Il rapporto fra le frequenze è indicato con $RR$ (*risk ratio*). Per esempio, l'analisi sull'efficacia riunisce 25 lavori relativi a 2923 studi e stima $RR = 1{,}27$. Ciò corrisponde a una frequenza di risultati favorevoli superiore del $27\%$ in termini relativi. La figura 11 distingue i risultati sull'efficacia, quelli sugli effetti dannosi e il giudizio complessivo espresso dagli autori nelle conclusioni.

BOZZA - DRAFT

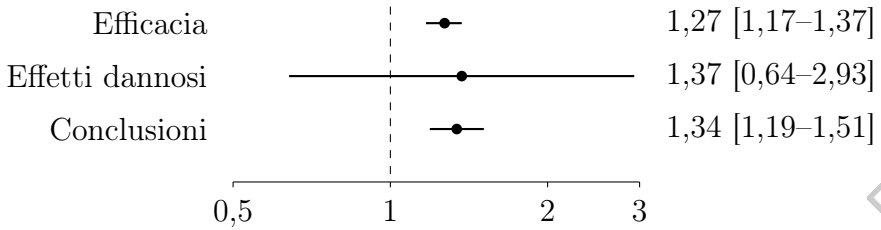



Figura 11: Tre stime di sintesi dalle analisi Cochrane 1.1, 1.2 e 3.1. Il punto indica il rapporto stimato e il segmento il suo intervallo di confidenza al 95%. L'asse del $RR$ è logaritmico. La linea tratteggiata in 1 indica uguale frequenza nei due gruppi.

La stima per gli effetti dannosi (1,37) supera quella per l'efficacia (1,27), ma il suo intervallo, basato su quattro lavori relativi a 826 studi, si estende da 0,64 a 2,93. Comprende quindi sia rapporti inferiori sia rapporti superiori a uno. Per l'efficacia l'intero intervallo, da 1,17 a 1,37, si trova sopra uno.

BOZZA - DRAFT

## Due farmaci a confronto: quale produttore finanzia il trial?

Chiamiamo $A$ il farmaco in esame e $B$ il farmaco di confronto. L'evento contato è sempre un risultato di efficacia favorevole ad $A$. Al numeratore del $RR$ si considera la proporzione di tali risultati quando finanzia il produttore di $A$, al denominatore quando finanzia il produttore di $B$. L'analisi Cochrane 2.1 comprende 131 trial relativi a statine e tiazolidinedioni, distribuiti fra due lavori. Nel lavoro di Bero il rapporto si può ricostruire direttamente:

$$\frac{43/65}{3/30} \simeq 6{,}62.$$

Nel lavoro di Rattinger lo stesso calcolo dà $(22/30)/(2/6) = 2{,}20$. La figura 12 mostra quanto siano incerte entrambe le stime, soprattutto perché i trial favorevoli ad $A$ finanziati dal produttore di $B$ sono rispettivamente soltanto tre e due.

BOZZA - DRAFT

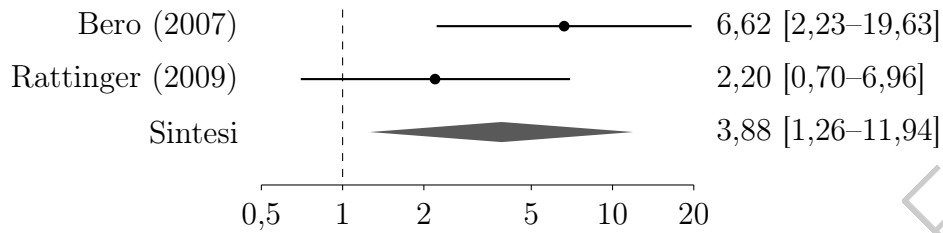



Figura 12: Rielaborazione dell'analisi Cochrane 2.1. Sull'asse logaritmico i valori a destra di 1 indicano che i risultati favorevoli ad $A$ sono più frequenti quando finanzia il suo produttore. Gli estremi del rombo delimitano l'intervallo di confidenza al 95% della stima combinata.

BOZZA - DRAFT

La metanalisi a effetti casuali stima che un risultato favorevole ad $A$ sia $3{,}88$ volte più frequente quando il trial è finanziato dal produttore di $A$. Il limite inferiore dell'intervallo ($1{,}26$) e quello superiore ($11{,}94$) suggeriscono grandezze molto diverse per la stessa associazione.

**Esercizio.** Un lettore somma i conteggi dei due lavori e calcola $(65/95)/(5/36)$. Quale valore ottiene? Perché differisce da $3{,}88$? *Risposta.* Il rapporto grezzo vale circa $4{,}93$. Sommando i conteggi si trattano tutti i trial co-

me un'unica raccolta. La metanalisi combina invece i rapporti stimati nei singoli lavori, tenendo conto della loro precisione e della variabilità fra lavori.

## 21.26 ESERCIZI SULLA LEZIONE 21

**Esempio svolto.** Calcoliamo i 5 indici per i 2 dataset visti all'inizio del paragrafo:

$$A:\ 2,4,6,8,10,12,14 \qquad\qquad B:\ 5,6,7,8,9,10,11$$

Entrambi hanno $n = 7$ e media $\bar{x} = 8$.

**Range.** $A$: $14 - 2 = 12$; $\quad B$: $11 - 5 = 6$.

**Differenza interquartile.** Escluso il valore centrale 8, le 2 metà sono
$A$: $2,4,6$ e $10,12,14$, quindi $q_{1/4} = 4$, $q_{3/4} = 12$, $iqr = 8$;
$B$: $5,6,7$ e $9,10,11$, quindi $q_{1/4} = 6$, $q_{3/4} = 10$, $iqr = 4$.

**Varianza.** Scarti dalla media e loro quadrati:
$A$: $-6,-4,-2,0,2,4,6 \ \rightarrow\ 36+16+4+0+4+16+36 = 112, \quad Var = \frac{112}{7} = 16$;

BOZZA - DRAFT

$B$: $-3, -2, -1, 0, 1, 2, 3 \;\to\; 9 + 4 + 1 + 0 + 1 + 4 + 9 = 28, \quad Var = \frac{28}{7} = 4.$

**Deviazione standard.** $A$: $\sqrt{16} = 4$; $\quad B$: $\sqrt{4} = 2$.

**Coefficiente di variazione.** $A$: $\frac{4}{8} = 0.5 = 50\%$; $\quad B$: $\frac{2}{8} = 0.25 = 25\%$.

| | $A$ | $B$ | $A/B$ |
|---|---|---|---|
| range | 12 | 6 | 2 |
| iqr | 8 | 4 | 2 |
| varianza | 16 | 4 | 4 |
| deviazione standard | 4 | 2 | 2 |
| coefficiente di variazione | 0.5 | 0.25 | 2 |

Gli scarti di $A$ dalla media sono esattamente il doppio di quelli di $B$: tutti gli indici raddoppiano, *tranne la varianza, che quadruplica*, perché è costruita coi quadrati degli scarti. È lo stesso motivo per cui la varianza ha come unità di misura il quadrato di quella dei dati, e per cui nelle Scienze Applicate si preferisce la deviazione standard.

BOZZA - DRAFT

**Esercizio**$_{\mu}$ Si afferma che un intervento del 1845 ha causato una rapida diminuzione del parametro X, e si presenta questo diagramma a colonne:

```
1845 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
1846 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
1847 XXXXXXXXXXXXXXXXXXXXXXXXX
1848 XXXXXXXXXXXXXXXXXXXX
1849 XXXXXXXXXXXXXXXX
1850 XXXXXXXXXXXXXX
```

È plausibile quell'affermazione? Ecco un dataset più completo:

```
1842 7.68
1843 6.14
1844 4.92
1845 3.93
1846 3.19
1847 2.58
1848 2.09
```

BOZZA - DRAFT

1849 1.69

1850 1.37

(Risposta: no, non è plausibile. Infatti fino al 1845 il parametro diminuiva circa del $20\%$ annuo e poi circa del $19\%$).

BOZZA - DRAFT