

Dati (dis)aggregati e approssimazioni di numeri reali

Giancarlo Travaglini

Dipartimento di Matematica e Applicazioni

Università di Milano-Bicocca, Edificio U5

Via Cozzi 55, 20125 Milano

<https://www.unimib.it/giancarlo-travaglini>

e-mail: giancarlo.travaglini@unimib.it

I numeri non hanno modo di parlare da soli. Noi parliamo per loro. Noi li investiamo di significato... possiamo interpretarli in modi autoreferenziali che sono distaccati dalla loro realtà oggettiva.

N. Silver, *The Signal and the Noise: The Art and Science of Prediction*, Penguin Books, 2013.

Un fatto curioso

Dovete scegliere un ristorante e decidete di tenere conto dei giudizi espressi in rete. Le alternative sono il “Ristorante Chiaravita” e il “Ristorante Anjèle”. Entrambi hanno avuto 40 recensioni, e i giudizi favorevoli sono riportati nella tabella seguente.

	Rist. Chiaravita	Rist. Anjèle
giudizi favorevoli degli uomini	$18/36 = 50\%$	$5/15 = 33,333\ldots\%$
giudizi favorevoli delle donne	$4/4 = 100\%$	$20/25 = 80\%$

Quindi il Ristorante Chiaravita ottiene percentuali di giudizi più favorevoli sia tra gli uomini sia tra le donne.

Questione chiusa? Non esattamente, perché se consideriamo tutti i clienti, cioè sommiamo i giudizi favorevoli ($18 + 4$ per Chiaravita e $5 + 20$ per Anjèle), otteniamo

	Rist. Chiaravita	Rist. Anjèle
giudizi favorevoli di tutti i clienti	$22/40 = 55\%$	$25/40 = 62,5\%$

Dunque in questo caso vince Anjèle. Qualcosa non funziona?!?

È tutto corretto, ma siamo entrati nel *Paradosso di Simpson*.

Il paradosso di Simpson si verifica quando la tendenza di un gruppo di dati aggregati è diversa dalla tendenza evidenziata dai dati non aggregati.

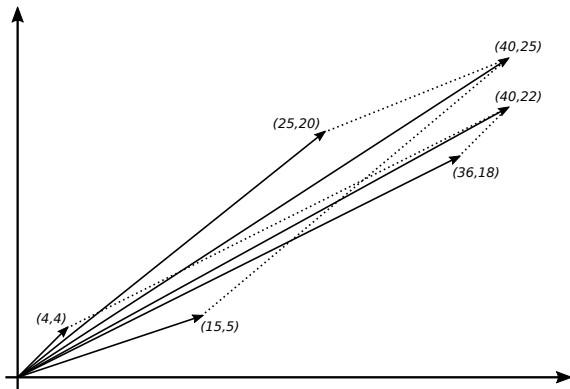
In questo caso abbiamo otto numeri interi positivi a, b, r, s, A, B, R, S che soddisfano

$$a \leq b, \quad r \leq s, \quad A \leq B, \quad R \leq S, \quad \frac{a}{b} < \frac{A}{B}, \quad \frac{r}{s} < \frac{R}{S}, \quad \text{ma} \quad \frac{a+r}{b+s} > \frac{A+R}{B+S}.$$

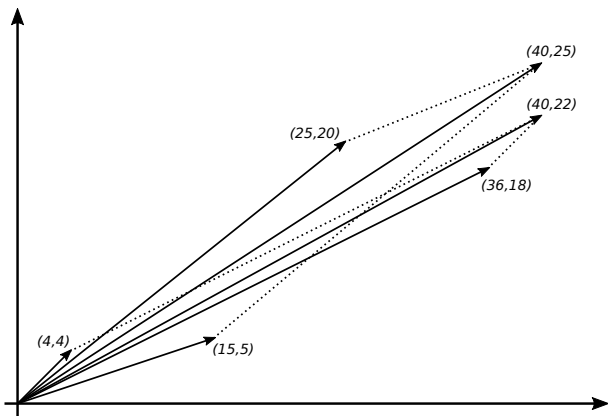
Infatti abbiamo

$$\frac{18}{36} > \frac{5}{15}, \quad \frac{4}{4} > \frac{20}{25}, \quad \text{ma} \quad \frac{18+4}{36+4} < \frac{5+20}{15+25}, \quad \text{cioè} \quad \frac{22}{40} < \frac{25}{40}.$$

Rappresentiamo questi numeri sul piano cartesiano nel seguente modo.



Possiamo vedere la frazione $18/36 = \frac{\text{ordinata}}{\text{ascissa}}$ come il coefficiente angolare della retta che collega l'origine con il punto di coordinate (36, 18), e $4/4$ come il coefficiente angolare della retta che collega l'origine con il punto di coordinate (4, 4). Analogamente $(18 + 4) / (36 + 4)$ è il coefficiente angolare della retta che collega l'origine con il punto di coordinate $(36 + 4, 18 + 4) = (40, 22)$. Cioè con la **somma dei vettori** (36, 18) e (4, 4), ottenuta attraverso la regola del parallelogrammo.



La disuguaglianza $18/36 > 5/15$ significa che la retta che collega l'origine con il punto $(36, 18)$ è più inclinata (nella figura precedente “sta sopra”) della retta che collega l'origine con il punto $(15, 5)$. In questo senso il Paradosso di Simpson significa che il vettore $(36, 18)$ è più inclinato del vettore $(15, 5)$, il vettore $(4, 4)$ è più inclinato del vettore $(25, 20)$, ma la somma $(36 + 4, 18 + 4)$ è meno inclinata della somma $(25 + 15, 20 + 5)$. Abbiamo così una rappresentazione grafica del Paradosso di Simpson.

Quella che segue può sembrare la battuta di un comico, e tuttavia rappresenta il problema che sorge di fronte al paradosso di Simpson.

Immaginiamo che un farmacista dica ad un genitore: “Per la malattia di suo figlio vendiamo due medicine: Alfa e Beta. Se suo figlio ha almeno 8 anni le consiglio Alfa, se ha meno di 8 anni le consiglio ancora Alfa; se però lei non sa l'età di suo figlio, allora le consiglio Beta”.

Situazioni reali in cui è apparso il Paradosso di Simpson

L'esempio dei ristoranti è stato "inventato a tavolino" per presentare numeri piccoli, quindi immediati da verificare, insieme ad una rappresentazione cartesiana anch'essa facile da (disegnare e) verificare personalmente. Vediamo ora tre esempi, presi dalla vita reale, in cui è apparso il Paradosso di Simpson.

Puntualità di voli

Quelli che seguono sono i numeri dei voli in orario e dei voli in ritardo per due compagnie aeree (Alaska Airlines e America West Airlines) su cinque aeroporti americani nel giugno 1991.

	Alaska Airlines			America West Airlines		
	in orario	ritardo	% ritardi	in orario	ritardo	% ritardi
Los Angeles	497	62	11,1%	694	117	14,4%
Phoenix	221	12	5,2%	4840	415	7,9%
San Diego	212	20	8,6%	383	65	14,6%
San Francisco	503	102	16,9%	320	129	28,7%
Seattle	1841	305	14,2%	201	61	23,3%
Tot. aeroporti	3274	501	13,3%	6438	787	10,9%

Come si vede, in tutti e cinque gli aeroporti considerati Alaska Airlines ha registrato percentuali di ritardo inferiori. Eppure, sommando i voli dei cinque aeroporti, la percentuale di ritardi di Alaska Airlines è superiore a quella di America West Airlines.

Discriminazione a Berkeley

Nel 1973 l'università di Berkeley fu uno dei primi atenei ad essere denunciato per discriminazione di genere. Per l'ammissione al semestre autunnale di quell'anno furono esaminate 12.763 domande di iscrizione (8442 ragazzi e 4321 ragazze) e le ammissioni suddivise per genere furono queste

	ammesse/i	respinte/i
ragazzi	3738	4704
ragazze	1494	2827

Quindi furono ammessi oltre il 44% dei maschi e meno del 35% delle femmine. Di fronte a questa evidenza, l'università disaggregò i dati per capire quali dipartimenti avevano contribuito a questa discrepanza.

Quelli che seguono sono i dati dei sei principali major (che indichiamo con A,B,C,D,E,F).

	ragazzi		ragazze	
	# domande	% ammessi	# domande	% ammessi
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Qui la differenza di genere è ancora più marcata: risultano globalmente ammessi circa il 45% dei maschi e il 30% delle femmine. Eppure in quattro major la percentuale di ragazze ammesse è maggiore, mentre nei rimanenti due il vantaggio dei maschi è contenuto. In altre parole, il dato disaggregato sembra indicare nessuna discriminazione o addirittura una discriminazione contro i maschi.

COVID-19 (2021)

Avvertenza. *Quello che segue non è un consiglio medico. È solo un altro esempio di paradosso di Simpson. I dati provengono dal Ministero della Salute di Israele, e non sono stati aggiornati dopo agosto 2021.*

A partire dall'inizio del 2020 la pandemia di COVID-19 ha colpito il mondo intero e una ragionevole speranza di salvezza è iniziata solo quando sono stati resi disponibili alcuni vaccini. Milioni di persone (per lo più nei paesi ricchi) hanno avuto l'opportunità di essere vaccinate, ma sfortunatamente molte di loro hanno rifiutato, spesso come conseguenza di disinformazione intenzionale o di una non intenzionale interpretazione errata di fatti reali. Di seguito vediamo due articoli tratti rispettivamente dal Corriere della Sera (18-19 agosto 2021) e dal Washington Post (21 agosto 2021).

SALUTE / MALATTIE-INFETTIVE

SPORTELLO CANCRO
NUTRIZIONE
CARDIOLOGIA
PEDIATROLOGIA
NEFROLOGIA
DIPARTIMENTO

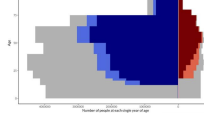


Covid, in Israele più ricoveri tra i vaccinati, ma Pfizer funziona contro la Delta: ecco come interpretare correttamente i dati

di Silvio Turin

La reale efficacia del vaccino Pfizer è ancora vicina al 90%. Quello che vediamo in Israele è il cosiddetto «paradosso di Simpson»: i dati, in realtà, confermano la bontà dell'immunizzazione. Un argomento utilizzato dai no vax confondendo l'analisi dei numeri

The US has vaccinated fewer of its older population against COVID than England
The number of people in each country who are vaccinated is lower in the US than in England, but the US has vaccinated more of its older population



Covid, Israele: «Con l'età non c'è protezione»

Si è parlato molto di dati recenti provenienti da Israele che sembrano suggerire un calo dell'efficacia del vaccino contro i casi gravi da Covid dovuti alla variante Delta.

Il paradosso di Simpson

Questa settimana i numeri mostravano che, su 65 pazienti attualmente ricoverati per Covid grave, 301 (il 46,4%) erano completamente vaccinati. Non significa che i vaccini non siano efficaci (e lo siano meno delle attese), ma solo che i numeri sono «ingannevoli» se non riferiti a un contesto specifico. Queste percentuali, infatti, sono confuse dall'età dei vaccinati e dall'alto tasso di vaccinazione di Israele e danno vita al «paradosso israeliano» che è quello conosciuto in statistica come «paradosso di Simpson». Vediamo di che cosa si tratta e come i numeri, in realtà, ci stanno dicendo che i vaccini proteggono ancora bene (in questo caso Pfizer, trattandosi di Israele, ndr).

Secondo gli ultimi dati, nel Paese l'efficacia della vaccinazione contro i casi gravi sembra essere piuttosto buona, attestandosi a circa il 89% quando sarebbe attesa a circa il 60%. La stessa efficacia misurata all'interno di due gruppi di età, però, oscilla tra l'85% e il 60% come mai? È il «paradosso israeliano», come si vede nella Tabella 1 (qui) sotto, ndr) compilata da Jeffrey S. Mehl, professore e direttore di Bionumerica presso la Penn State School of Medicine dell'Università della Pennsylvania (qui dati forniti dal governo israeliano).

CONTINUIAMO LA SERIE DI PROPOSIZIONI

Paese	Età	Vaccinati	Casi gravi	Morti
Israele	0-17	100%	0	0
Israele	18-24	100%	0	0
Israele	25-34	100%	0	0
Israele	35-44	100%	0	0
Israele	45-54	100%	0	0
Israele	55-64	100%	0	0
Israele	65-74	100%	0	0
Israele	75-84	100%	0	0
Israele	85+	100%	0	0

Covid, il bollettino di oggi: 57.400 nuovi casi e 320 morti. Calano i ricoveri in terapia intensiva

Per Everything - Perspective

Coronavirus vaccines work. But this statistical illusion makes people think they don't.

In Israel for a time, more vaccinated people were hospitalized for covid-19 than unvaccinated people. There's no reason to worry.



An Israeli man talks to a woman while waiting a first dose of the Pfizer coronavirus vaccine in Ramat Gan, Israel, on Aug. 30. Israel is grappling with a surge of infections and urging people over 12 to get a booster shot. (David H. Burton)



By Jordan Elzein

Jordan Elzein is a senior professor at the University of Wisconsin, is the author of "Shades: The Hidden Geometry of Information, Strategy, Strategy, Democracy, and Everything Else."



August 30, 2022 at 6:00 a.m. EDT

Is the vaccine wearing off? It's an exhausting thought for those of us who believed the battle against covid-19 would be won once enough people plunged into enough arms. But outbreaks of the delta variant have blossomed even in places with high levels of vaccination, including Israel, Britain and my own home of Madison, Wis. Recent reports from Israel that nearly 60 percent of people hospitalized with severe covid-19 were fully vaccinated raised particular alarms about the limits of the protection that vaccines provide.

While there is other evidence that the vaccines' protection against infection (not severe illness) declines over time — and also some evidence that booster shots work — we shouldn't be overly concerned about the Israeli hospitalization statistics. The explanation involves a famous old statistical conundrum called "Simpson's paradox" — which isn't really a paradox at all. It's just a reminder of why you have to be careful with data.

Simpson's paradox is a warning that the whole of the data often looks weirdly different than the sum of its parts. In the case of Israel — as a number of epidemiologists and other scholars have pointed out — what explains the surprising hospitalization figures is largely the relative ages of vaccinated and unvaccinated people. The University of Pennsylvania biostatistician Jeffrey Mehl wrote an especially thorough and widely shared blog post making this point — although the post was not shared nearly as widely as the worry that vaccines aren't helping you stay out of the intensive care unit. Anxiety travels halfway around the world while a spreadsheet is still getting its boots on.

Washington Post LIVE
FIRST LOOK
Washington Post Live's "First Look" offers a smart, inside take on the day's politics.
Tune in every Friday
9:00am Washington, D.C.
2:00pm London
REGISTER
Streamed live on Facebook, Twitter, YouTube, LinkedIn and The Washington Post Live homepage

MOST READ OPINIONS



- 1 **Opinion**
Why Ukraine — and Russia — aggression against it — matters to Americans
- 2 **Opinion**
How Putin's missteps could lead to his overthrow
- 3 **Opinion**
This year's GPEC report signals economic doom

Abbiamo qui un esempio di “interpretazione errata non intenzionale di fatti reali”, riportato dallo statistico Jeffrey Morris nell’agosto 2021 e basato su dati israeliani del 15 agosto 2021, secondo i quali si erano verificati, a causa del COVID-19, 515 casi gravi, e 301 di essi (quindi il 58.5%) erano stati precedentemente vaccinati con il vaccino Pfizer.

Allora molte persone sollevarono dubbi sull’efficacia del vaccino.

Mostriamo che questi dati non contraddicono i risultati eccellenti (almeno quelli noti nel 2021, vedi sotto) sull’efficacia del vaccino Pfizer.

Per prima cosa, i numeri sopra riportati devono essere normalizzati. 1302912 persone non sono state vaccinate, mentre 5634634 persone sono state vaccinate. Quindi abbiamo i seguenti rapporti:

$$\frac{214}{1302912} \approx \frac{16,42}{100000} \quad \text{non vaccinati hanno sviluppato la malattia grave,}$$
$$\frac{301}{5634634} \approx \frac{5,34}{100000} \quad \text{vaccinati hanno sviluppato la malattia grave.}$$

L'efficacia di un vaccino (VE) è definita da

$$VE = \left(\frac{ARU - ARV}{ARU} \times 100 \right) \% ,$$

dove ARU (Attack rate of unvaccinated people) è il quoziente tra il numero dei nuovi casi e il numero di persone a rischio tra le persone non vaccinate, mentre ARV è il quoziente tra il numero dei nuovi casi e il numero di persone a rischio tra le persone vaccinate.

Se il vaccino non ha alcun effetto, allora $ARU = ARV$ è l'efficacia è 0%.

Se il vaccino è molto efficace, allora ARV è decisamente più piccolo di ARU e l'efficacia si avvicina al 100%.

Nel mondo reale i vaccini non sono mai completamente efficaci. Ci aspettiamo, quindi, $VE < 100\%$. Secondo vari studi disponibili nel 2021 l'efficacia del vaccino Pfizer nel prevenire la malattia grave supera il 95%. Invece, secondo i dati israeliani, l'efficacia del vaccino Pfizer sembra essere

$$VE = \left(\frac{16,42 - 5,34}{16,42} \times 100 \right) \% \approx 67,48\% ,$$

che non è molto.

Se però separiamo la popolazione secondo due livelli di età (≤ 50 e > 50) troviamo un quadro diverso (per completezza va aggiunto che non sono stati considerati i minori di età ≤ 12 anni e coloro che hanno ricevuto una sola dose di vaccino), dove nuovamente vediamo il Paradosso di Simpson:

età	non vacc. %	vacc. %	malattia grave (non vacc.)	malattia grave (vacc.)	VE
≤ 50	1116834 23,3%	3501118 73%	$\frac{43}{1116834} \approx \frac{3,9}{100000}$	$\frac{11}{3501118} \approx \frac{0,3}{100000}$	92,3%
> 50	186078 7,9%	2133516 90,4%	$\frac{171}{186078} \approx \frac{91,9}{100000}$	$\frac{290}{2133516} \approx \frac{13,6}{100000}$	85,2%
tutti	1302912 18,2%	5634634 78,7%	$\frac{214}{1302912} \approx \frac{16,4}{100000}$	$\frac{301}{5634634} \approx \frac{5,3}{100000}$	67,5%

Perché il paradosso

In termini matematici abbiamo capito cosa è successo, ma non abbiamo ancora capito perché. Se però esaminiamo i quattro esempi di Paradosso di Simpson visti in precedenza, notiamo qualche caratteristica comune.

Nel caso dei ristoranti i giudizi degli uomini appaiono più severi di quelli delle donne (50% e 33, 33... % di giudizi positivi da parte degli uomini contro 100% e 80% da parte delle donne). Nel risultato complessivo il Ristorante Chiaravita è quindi svantaggiato dal fatto di avere avuto molti giudizi di uomini (36 su 40 giudizi), mentre Anjèle ha avuto solo 15 giudizi di uomini.

Nel caso dei ritardi nei voli notiamo che i ritardi erano legati anche all'aeroporto di arrivo. A Phoenix i ritardi erano pochi (5,2% per Alaska Airlines e 7,9% per America West Airlines), mentre a San Francisco e a Seattle la percentuale di ritardi era molto alta (16,9% e 14,2% per Alaska Airlines, contro 28,7% e 23,3% per America West Airlines). Quindi nel "Totale" Alaska Airlines era penalizzata dal fatto di avere pochi voli su Phoenix e molti voli su San Francisco e Seattle. La morale è che le compagnie aeree possono decidere dove volare e come pianificare i loro voli, ma una compagnia che sceglie di operare in modo estensivo presso aeroporti difficili deve accettare le conseguenze di questa scelta.

Nel caso di Berkeley, la spiegazione è nel fatto che le ragazze tendevano ad iscriversi ai major più selettivi, nei quali la percentuale di ammissioni era inferiore, mentre i maschi si iscrivevano spesso a major in cui era più facile essere ammessi.

Rivediamo infatti i dati.

	ragazzi		ragazze	
	# domande	% ammessi	# domande	% ammessi
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Nel caso del vaccino Pfizer osserviamo che il dato aggregato (214 casi gravi tra i non vaccinati e 301 casi gravi tra i vaccinati) è stato largamente determinato dalle persone più anziane, di età > 50 . Le persone più anziane sono anche quelle più a rischio, e sono quasi tutte vaccinate.

Ricordiamo che stiamo confrontando due scelte: *vaccinarsi* e *non vaccinarsi*. Riteniamo che la vaccinazione sia la scelta migliore. La domanda è *Di quanto è migliore?* Dobbiamo cioè calcolare - e capire - il *VE*. Se stratifichiamo ulteriormente la popolazione secondo l'età otteniamo un quadro più chiaro, che conferma sostanzialmente la performance (fino al 2021) del vaccino Pfizer:

età	$\frac{\text{malattia grave}}{100000}$ (non vaccinati)	$\frac{\text{malattia grave}}{100000}$ (vaccinati)	VE
12-15	0.3	0	100%
16-19	1.6	0	100%
20-29	1.5	0	100%
30-39	6.2	0.2	96.8%
40-49	16.5	1	93.9%
50-59	40.2	2.9	92.8%
60-69	76.6	8.7	88.7%
70-79	190.1	19.8	89.6%
80-89	252.3	47.9	81.1%
90+	510.9	38.6	92.4%

In tutti gli esempi visti il paradosso nasce dalla presenza, nei dati disaggregati, di denominatori molto diversi tra loro. Ad esempio, nel caso dei ristoranti, 36 uomini contro 4 donne per Caravita, e 15 uomini contro 25 donne per Anjèle. *Vediamo quindi un problema di denominatori, non di percentuali.* Cioè un problema di frazioni, non di numeri razionali

Noi diciamo - giustamente - che

$$\frac{1}{2} = \frac{500}{1000} ,$$

cioè che 1/2 e 500/1000 rappresentano lo stesso numero razionale. Sono però frazioni diverse: difficilmente una persona di buon senso direbbe che ottenere 1 caso positivo su 2 esperimenti è “la stessa cosa” che ottenere 500 casi positivi su 1000 esperimenti.

Mostriamo che è possibile fare sparire il paradosso dall'esempio dei ristoranti mantenendo le percentuali (ma cambiando i denominatori). Consideriamo infatti questa nuova situazione, basata sempre su 40 giudizi.

	Rist. Chiaravita	Rist. Anjèle
giudizi positivi degli uomini	$10/20 = 50\%$	$5/15 = 33,333 \dots \%$
giudizi positivi delle donne	$20/20 = 100\%$	$20/25 = 80\%$

Allora, non essendo mutate le percentuali, il Ristorante Chiaravita continua ad ottenere giudizi più favorevoli sia tra gli uomini sia tra le donne. Ora, però, se consideriamo tutti i clienti otteniamo

	Rist. Chiaravita	Rist. Anjèle
giudizi positivi di tutti i clienti	$30/40 = 75\%$	$25/40 = 62,5\%$

e qui il Paradosso di Simpson non si presenta.

La media aritmetica di due numeri razionali s e t è $(s + t)/2$ e non dipende dal modo in cui scriviamo s e t come frazioni. All'opposto il risultato seguente mostra che possiamo ottenere ogni media di Farey tra s e t , pur di scegliere opportunamente il modo in cui scriviamo i numeri razionali s e t come frazioni.

Proposizione. Siano $r = A/B$ e $s = X/Y$ frazioni di interi positivi e consideriamo qualsiasi numero razionale v tale che $r < v < s$. Allora esistono due interi positivi p e q tali che v è uguale alla media di Farey delle frazioni $pA/pB (= r)$ e $qX/qY (= s)$. Cioè

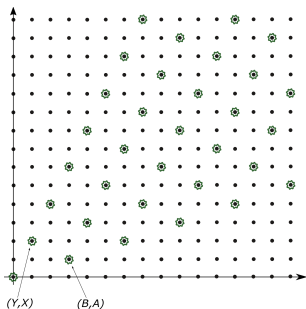
$$v = \frac{pA + qX}{pB + qY} .$$

Dimostrazione Scriviamo $v = \alpha/\beta$. Allora v è la media di Farey di pA/pB e qX/qY , con $p = X\beta - Y\alpha$ e $q = B\alpha - A\beta$.

Geometricamente questo significa che l'insieme delle pendenze delle rette che collegano l'origine con i punti dell'insieme

$$\{(pB + qY, pA + qX)\}_{p,q \in \mathbb{N}}$$

è uguale all'insieme dei numeri razionali compresi tra A/B e X/Y . Questo insieme è rappresentato dai punti evidenziati nella figura seguente.



In termini del Paradosso di Simpson questo significa che possiamo fissare le percentuali dei dati disaggregati e (giocando con p e q) ottenere qualsiasi dato aggregato desideriamo.

Definizione

La frazione

$$\frac{A + R}{B + S},$$

che abbiamo appena visto nel paradosso di Simpson, è detta *media di Farey* delle due frazioni A/B e R/S .

Rispetto alla familiare (e importante) media aritmetica

$$\frac{1}{2} \left(\frac{A}{B} + \frac{R}{S} \right) = \frac{AS + BR}{2BS},$$

la media di Farey si rivela utile nell'approssimazione diofantea, cioè nell'approssimazione di numeri reali attraverso frazioni, possibilmente con denominatori piccoli.

Sulla retta reale, sia la media aritmetica che la media di Farey sono comprese tra A/B e R/S , ma osserviamo che il denominatore della media di Farey di A/B e R/S è la somma $B + S$, mentre il denominatore della loro media aritmetica è il prodotto $2BS$, che può essere molto più grande di $B + S$.

Approssimare numeri real con frazioni aventi denominatori non troppo grandi

Sappiamo che per ogni numero reale esistono numeri razionali ad esso arbitrariamente vicini. Ad esempio, il numero

$$\pi = 3,14159\dots$$

è approssimato da $3,1 = 31/10$ a meno di $1/10$, da $3,14 = 314/100$ a meno di $1/100$, da $3,141 = 3141/1000$ a meno di $1/1000$, e così via. Possiamo però ricordare che, ad esempio

$$\frac{22}{7} = 3,14285\dots$$

è una buona approssimazione di π , ottenuta con un denominatore abbastanza piccolo. Infatti è meglio di $3,14 = 314/100 = 157/50$.

Vedremo tra poco come “trovare” $22/7$ usando le medie di Farey.

Prima però riportiamo il più famoso risultato di approssimazione diofantea, dimostrato da Dirichlet nel 1840.

Teorema. Per ogni numero irrazionale α esistono infinite frazioni p/n per ciascuna delle quali

$$\left| \alpha - \frac{p}{n} \right| < \frac{1}{n^2} .$$

Successioni di Farey

Fissato un intero positivo n consideriamo l'insieme $\mathfrak{F}_n$ (detto *successione di Farey n -esima*) costituito dalle frazioni semplificate di denominatore $\leq n$, comprese tra 0 e 1, disposte in ordine crescente:

$\mathfrak{F}_1$	$\frac{0}{1}$																		$\frac{1}{1}$
$\mathfrak{F}_2$	$\frac{0}{1}$								$\frac{1}{2}$										$\frac{1}{1}$
$\mathfrak{F}_3$	$\frac{0}{1}$						$\frac{1}{3}$		$\frac{1}{2}$		$\frac{2}{3}$							$\frac{1}{1}$	
$\mathfrak{F}_4$	$\frac{0}{1}$			$\frac{1}{4}$		$\frac{1}{3}$		$\frac{1}{2}$		$\frac{2}{3}$		$\frac{3}{4}$						$\frac{1}{1}$	
$\mathfrak{F}_5$	$\frac{0}{1}$		$\frac{1}{5}$	$\frac{1}{4}$		$\frac{1}{3}$	$\frac{2}{5}$		$\frac{1}{2}$		$\frac{3}{5}$	$\frac{2}{3}$		$\frac{3}{4}$	$\frac{4}{5}$			$\frac{1}{1}$	
$\mathfrak{F}_6$	$\frac{0}{1}$		$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{4}$		$\frac{1}{3}$	$\frac{2}{5}$		$\frac{1}{2}$		$\frac{3}{5}$	$\frac{2}{3}$		$\frac{3}{4}$	$\frac{4}{5}$	$\frac{5}{6}$	$\frac{1}{1}$	
$\mathfrak{F}_7$	$\frac{0}{1}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{2}{7}$	$\frac{1}{3}$	$\frac{2}{5}$	$\frac{3}{7}$	$\frac{1}{2}$	$\frac{4}{7}$	$\frac{3}{5}$	$\frac{2}{3}$	$\frac{5}{7}$	$\frac{3}{4}$	$\frac{4}{5}$	$\frac{5}{6}$	$\frac{6}{7}$	$\frac{1}{1}$

È un modo per mettere un po' di ordine tra le frazioni semplificate, cioè tra i numeri razionali.

Si può dimostrare il seguente

Teorema. Siano h/k e h'/k' due termini consecutivi in una successione di Farey. Allora la media di Farey

$$\frac{h + h'}{k + k'}$$

è la frazione con minor denominatore, tra tutte quelle strettamente comprese tra h/k e h'/k' .

Ad esempio, in $\mathfrak{F}_7$ troviamo le frazioni successive $4/7$ e $3/5$. Quindi la loro media di Farey $7/12$ è, tra tutte le frazioni comprese tra $4/7$ e $3/5$, quella con il denominatore più piccolo. Cioè tra $4/7$ e $3/5$ non esiste alcuna frazione avente per denominatore uno dei numeri $1, 2, 3, \dots, 10, 11$.

Approssimazioni razionali esplicite

Supponiamo di volere approssimare π . Basta approssimare $\pi - 3$, e, ovviamente, $\frac{0}{1} < \pi - 3 < \frac{1}{1}$. Confrontiamo $\pi - 3 = 0,14159\dots$ con la media di Farey di $0/1$ e $1/1$, cioè con $1/2$. Otteniamo $0/1 < \pi - 3 < 1/2 < 1/1$ e scartiamo $1/1$. Poi confrontiamo $\pi - 3$ con la media di Farey di $0/1$ e $1/2$, cioè con $1/3$. Otteniamo $0 < \pi - 3 < 1/3 < 1/2$ e scartiamo $1/2$. E così via. Scriviamo così la tabella

media di Farey			abbiamo	scartiamo
0/1	1/1	1/2	$0 < \pi - 3 < 1/2 < 1/1$	1/1
0/1	1/2	1/3	$0 < \pi - 3 < 1/3 < 1/2$	1/2
0/1	1/3	1/4	$0 < \pi - 3 < 1/4 < 1/3$	1/3
0/1	1/4	1/5	$0 < \pi - 3 < 1/5 < 1/4$	1/4
0/1	1/5	1/6	$0 < \pi - 3 < 1/6 < 1/5$	1/5
0/1	1/6	1/7	$0 < \pi - 3 < 1/7 < 1/6$	1/6
0/1	1/7	1/8	$0 < 1/8 < \pi - 3 < 1/7$	0/1
1/8	1/7	2/15	$1/8 < 2/15 < \pi - 3 < 1/7$	1/8

Tra $2/15 = 0,13333\dots$ e $1/7 = 0,14286\dots$, osserviamo che $1/7$ è più vicina a $\pi - 3 = 0,14159\dots$. Scegliamo quindi $1/7$ come approssimazione di $\pi - 3$, cioè $22/7$ come approssimazione di π .

Come ulteriore esercizio cerchiamo una approssimazione p/n di $\sqrt{\pi}$, con n (e quindi anche p) non troppo grande. Di nuovo, consideriamo $\sqrt{\pi} - 1$. Con qualsiasi calcolatrice possiamo ottenere

$$\sqrt{\pi} - 1 = 0,77245 \dots$$

ma non è quello che vogliamo, poiché

$$0,77245 = \frac{77245}{100000} = \frac{15449}{20000}$$

soddisfa

$$\left| \sqrt{\pi} - 1 - 0.77245 \right| = \left| \sqrt{\pi} - 1 - \frac{15449}{20000} \right| = 0.0000038 \dots < \frac{1}{200000} \quad (1)$$

al costo di cinque cifre a denominatore. Cerchiamo una approssimazione migliore.

Se guardiamo le prime successioni di Farey vediamo che

$$\frac{3}{4} < \sqrt{\pi} - 1 < \frac{4}{5}, \quad (2)$$

dove $3/4$ e $4/5$ sono consecutive nelle successioni di Farey $\mathfrak{F}_N$ con $5 \leq N < 9$. Confrontiamo $\sqrt{\pi} - 1$ con la media di Farey $7/9$ di $3/4$ e $4/5$. Otteniamo

$$3/4 < \sqrt{\pi} - 1 < 7/9 < 4/5,$$

e allora scartiamo $4/5$. Poi confrontiamo $\sqrt{\pi} - 1$ con la media di Farey $10/13$ di $3/4$ e $7/9$. Procedendo così otteniamo la tabella seguente:

		media di Farey	abbiamo	scartiamo
3/4	4/5	7/9	$3/4 < \sqrt{\pi} - 1 < 7/9 < 4/5$	4/5
3/4	7/9	10/13	$3/4 < 10/13 < \sqrt{\pi} - 1 < 7/9$	3/4
10/13	7/9	17/22	$10/13 < \sqrt{\pi} - 1 < 17/22 < 7/9$	7/9
10/13	17/22	27/35	$10/13 < 27/35 < \sqrt{\pi} - 1 < 17/22$	10/13
27/35	17/22	44/57	$27/35 < 44/57 < \sqrt{\pi} - 1 < 17/22$	27/35
44/57	17/22	61/79	$44/57 < 61/79 < \sqrt{\pi} - 1 < 17/22$	44/57
61/79	17/22	78/101	$61/79 < 78/101 < \sqrt{\pi} - 1 < 17/22$	61/79
78/101	17/22	95/123	$78/101 < 95/123 < \sqrt{\pi} - 1 < 17/22$	78/101
95/123	17/22	112/145	$95/123 < 112/145 < \sqrt{\pi} - 1 < 17/22$	95/123
112/145	17/22	129/167	$112/145 < \sqrt{\pi} - 1 < 129/167 < 17/22$	17/22
112/145	129/167	241/312	$112/145 < 241/312 < \sqrt{\pi} - 1 < 129/167$	112/145
241/312	129/167	370/479	$241/312 < 370/479 < \sqrt{\pi} - 1 < 129/167$	241/312
370/479	129/167	499/646	$370/479 < 499/646 < \sqrt{\pi} - 1 < 129/167$	370/499
499/646	129/167	628/813	$499/646 < 628/813 < \sqrt{\pi} - 1 < 129/167$	499/646
628/813	129/167	757/980	$628/813 < 757/980 < \sqrt{\pi} - 1 < 129/167$	628/813
757/980	129/167	886/1147	$757/980 < 886/1147 < \sqrt{\pi} - 1 < 129/167$	757/980

Quindi scegliamo l'approssimazione razionale del numero $\sqrt{\pi} - 1$ tra $886/1147 = 0,7724498692\dots$ e $129/167 = 0,7724550898\dots$. La migliore è $129/167$ e abbiamo, al costo di un denominatore di tre sole cifre,

$$\left| \sqrt{\pi} - 1 - \frac{129}{167} \right| = 0,00000123\dots < \frac{1}{800000}.$$

Numeri di Fibonacci e successioni di Farey

Il problema originale studiato da Fibonacci all'inizio del 1200 riguardava la riproduzione dei conigli ed aveva la seguente forma.

Un certo uomo mette una coppia di conigli in un posto circondato su tutti i lati da un muro. Quante coppie di conigli possono essere prodotte da quella coppia in un anno, se si suppone che ogni mese ogni coppia genera una nuova coppia, che dal secondo mese in avanti diventa produttiva?

Assumiamo dunque le seguenti ipotesi.

- Un coniglio si riproduce per la prima volta dopo un mese dalla nascita e successivamente ogni mese. Chiamiamo mature le coppie di conigli di età non inferiore ad un mese.
- La gestazione dura un mese, al termine del quale una femmina partorisce un maschio e una femmina.
- I conigli non muoiono.

Vediamo la situazione ad ogni mese (dove “mese” significa “inizio mese”).

Mese 1. C'è una coppia di conigli appena nati.

Mese 2. C'è una coppia matura.

Mese 3. Ci sono una coppia matura e una coppia appena nata.

Mese 4. Ci sono due coppie mature e una coppia appena nata.

Mese 5. Ci sono tre coppie mature e due coppie appena nate.

Mese 6. Ci sono cinque coppie mature e tre coppie appena nate.

A questo punto la legge è chiara. Nel mese $n + 1$ ci sono tutte le coppie del mese n più tutte le coppie nate nel mese $n + 1$, e queste ultime sono tante quante le coppie mature al mese n , che a loro volta sono tante quante le coppie (senza aggettivi) al mese $n - 1$. Se indichiamo con F_n è il numero delle coppie di conigli presenti al mese n otteniamo la legge

$$\begin{cases} F_1 = F_2 = 1 \\ F_{n+1} = F_n + F_{n-1} \quad \text{se } n > 2, \end{cases}$$

da cui

$$\{F_n\} = \{F_1, F_2, F_3, F_4, \dots\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, \dots\}$$

In particolare, dopo un anno abbiamo $F_{12} = 144$ coppie di conigli.

F_n ammette una scrittura esplicita. Dimostriamo infatti il seguente

Teorema.

$$F_n = \frac{1}{\sqrt{5}} \left\{ \left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right\} .$$

Dimostrazione. Qui è conveniente premettere $F_0 = 0$ alla successione di Fibonacci. Dobbiamo quindi risolvere

$$\begin{cases} F_0 = 0, F_1 = 1 \\ F_{n+1} = F_n + F_{n-1} \quad \text{se } n \geq 2, \end{cases}$$

L'idea è di cercare una soluzione della forma $F_n = k^n$ (per il momento dimentichiamo che F_n deve essere un intero positivo). Sostituendo otteniamo

$$k^{n+1} = k^n + k^{n-1}.$$

Il caso $k = 0$ non ci interessa perché non soddisfa $F_1 = 1$; allora, dividendo per k^{n-1} , studiamo

$$k^2 = k + 1,$$

che ha soluzioni $k = (1 \pm \sqrt{5})/2$. Dunque $(1 + \sqrt{5}/2)^n$ e $((1 - \sqrt{5})/2)^n$ soddisfano l'equazione $F_{n+1} = F_n + F_{n-1}$:

$$\begin{aligned} \left(\frac{1 + \sqrt{5}}{2}\right)^{n+1} &= \left(\frac{1 + \sqrt{5}}{2}\right)^n + \left(\frac{1 + \sqrt{5}}{2}\right)^{n-1} \\ \left(\frac{1 - \sqrt{5}}{2}\right)^{n+1} &= \left(\frac{1 - \sqrt{5}}{2}\right)^n + \left(\frac{1 - \sqrt{5}}{2}\right)^{n-1}. \end{aligned}$$

Osserviamo però che $(1 + \sqrt{5}/2)^n$ e $((1 - \sqrt{5})/2)^n$ non soddisfano $F_1 = 1$.

Possiamo moltiplicare per due costanti A, B arbitrarie e sommare, ottenendo

$$F_n = A \left(\frac{1 + \sqrt{5}}{2} \right)^n + B \left(\frac{1 - \sqrt{5}}{2} \right)^n ,$$

che ancora soddisfa l'identità $F_{n+1} = F_n + F_{n-1}$, comunque si scelgano A e B . Ora ricordiamo che vogliamo $F_0 = 0$, $F_1 = 1$. Il sistema

$$\begin{cases} F_0 = A + B = 0 \\ F_1 = A \frac{1+\sqrt{5}}{2} + B \frac{1-\sqrt{5}}{2} = 1 \end{cases}$$

ha soluzione $A = 1/\sqrt{5}$, $B = -1/\sqrt{5}$, e da qui segue

$$F_n = \frac{1}{\sqrt{5}} \left\{ \left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right\} .$$

N.B. Questa scrittura di F_n non sembra un numero intero, ma deve esserlo (e in effetti lo è), poiché la condizione

$$\begin{cases} F_0 = 0 , F_1 = 1 \\ F_{n+1} = F_n + F_{n-1} \quad \text{se } n \geq 2, \end{cases}$$

è soddisfatta solo da numeri interi.

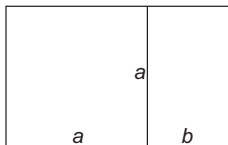
Il numero $(1 + \sqrt{5})/2$ è chiamato *sezione aurea* ed è uguale al rapporto a/b fra due numeri positivi soggetti alla condizione

$$\frac{a}{b} = \frac{a+b}{a} .$$

Infatti, scrivendo $a/b = x$, la scrittura precedente diventa $x = 1 + x^{-1}$, cioè

$$x^2 - x - 1 = 0 ,$$

che ha $(1 + \sqrt{5})/2$ come unica radice positiva.



Nella figura precedente il rettangolo grande, i cui lati misurano $a + b$ e a , è simile al rettangolo a destra, ottenuto togliendo il quadrato di lato a .

Un rettangolo il cui rapporto tra i lati sia uguale alla sezione aurea è detto *Rettangolo Aureo*, e non è difficile incontrarlo in varie manifestazioni della natura e dell'arte.

Vediamo ora una connessione tra i numeri di Fibonacci e le medie di Farey. Per questo scriviamo le frazioni di numeri di Fibonacci successivi:

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{3}, \frac{3}{5}, \frac{5}{8}, \frac{8}{13}, \frac{13}{21}, \frac{21}{34}, \frac{34}{55}, \frac{55}{89}, \frac{89}{144}, \frac{144}{233}, \dots$$

Osserviamo che

$$\frac{F_n}{F_{n+1}} = \frac{F_{n-1} + F_{n-2}}{F_n + F_{n-1}},$$

cioè F_n/F_{n+1} è la media di Farey delle due frazioni F_{n-2}/F_{n-1} e F_{n-1}/F_n che la precedono.

Gerrymandering

Nella seconda parte di questo seminario mostriamo un altro famoso fenomeno dove dati aggregati e dati disaggregati possono non andare d'accordo.

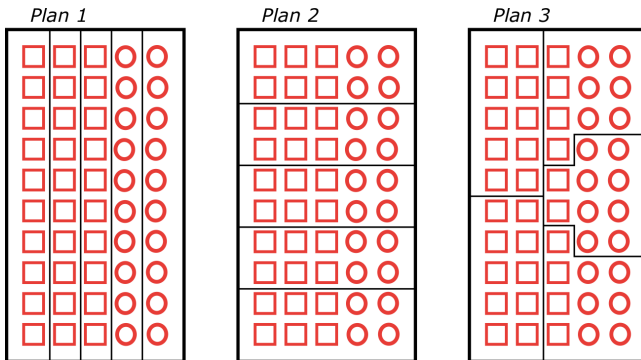
Il termine (*partisan*) *gerrymandering* si riferisce alla manipolazione di distretti elettorali allo scopo di avvantaggiare un partito politico.

Il gerrymandering è un problema particolarmente serio riguardo all'elezione della Camera dei Rappresentanti degli Stati Uniti (cioè la camera bassa del Congresso), per la quale devono essere eletti 435 membri.

Stiamo parlando di un sistema elettorale maggioritario, nel quale il paese è diviso in distretti elettorali, ciascuno dei quali elegge a maggioranza un rappresentante, che dovrebbe agire a nome e nell'interesse del distretto che lo ha eletto.

Come esempio, assumiamo che il paese abbia 50 votanti, di cui 30 fedeli al *Partito Quadrato* e 20 fedeli al *Partito Rotondo*. Supponiamo che tutti votino e che il paese debba essere diviso in 5 distretti elettorali, con 10 elettori ciascuno.

In un sistema elettorale proporzionale il Partito Quadrato avrebbe il 60% dei voti (dunque 3 rappresentanti) e il Partito Rotondo avrebbe il 40%. Nella figura seguente vediamo a confronto i risultati di tre diverse scelte della suddivisione del paese in 5 distretti.



Nel Piano 1 ogni partito ha la percentuale di rappresentanti (3 per il Partito Quadrato e 2 per il Partito Rotondo) uguale alla sua percentuale di voti.

Il Piano 2 è il “sogno del Partito Quadrato”, che ottiene tutti e 5 i rappresentanti.

Il Piano 3 è il “colpaccio del Partito Rotondo”, che ottiene 3 rappresentanti e ha quindi, contro tutte le aspettative, la maggioranza nel paese.

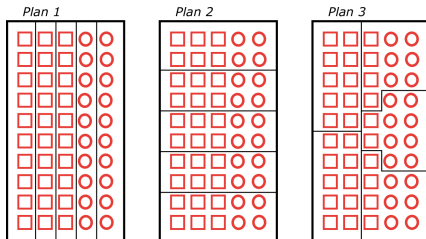
Il Piano 2 e il Piano 3 sono esempi (con esiti opposti) di gerrymandering.

La tentazione del gerrymandering appare nel momento in cui le autorità devono ridefinire i confini dei distretti elettorali (negli Stati Uniti questo avviene ogni 10 anni, dopo il censimento, con regole che variano da stato a stato).

Supponiamo che per le prossime elezioni sia prevista una distribuzione di voti quadrati e rotondi localizzata come nella figura precedente. Se al governo si trova il Partito Quadrato, sarà tentato di suddividere i distretti elettorali come nel Piano 2, mentre se al governo si trova il Partito Rotondo, la tentazione del Piano 3 sarà forte.

L'obiettivo del gerrymandering è *sprecare i voti del partito avversario*. Questo può essere fatto in due modi.

- **convogliando i voti del partito avversario in inutili “supermaggioranze”**, vedi i due distretti a sinistra nel Piano 3.
- **sprecando i voti del partito avversario su candidati perdenti**, vedi i cinque distretti nel Piano 2 e i tre distretti a destra nel Piano 3.



La nascita del termine “gerrymandering” risale al 1812, quando il Boston Gazette pubblicò una vignetta che mostrava con una rappresentazione suggestiva i contorni contorti di un distretto elettorale del Massachusetts, riprodotti nella figura seguente.



In questi contorni si vedeva, con un po' di fantasia, la forma di un mostro o di una salamandra (*salamander* in inglese), e poiché erano stati approvati dal governatore Elbridge Gerry, nacque il termine *gerry-mander*.

Nel 2019 la Corte Suprema degli Stati Uniti ha stabilito che la Corte non ha il diritto di intervenire sul problema del gerrymandering, pur dichiarando che la pratica del gerrymandering è incompatibile con i principi democratici. Il problema deve quindi essere affrontato all'interno dei singoli stati.

La Corte era chiamata a pronunciarsi su due opposti ricorsi per gerrymandering: nel primo i querelanti erano cittadini del North Carolina e denunciavano discriminazioni contro il Partito Democratico, nel secondo venivano denunciate discriminazioni contro il Partito Repubblicano nel Maryland.

È interessante leggere la presentazione dei due casi nella sentenza della Corte Suprema.

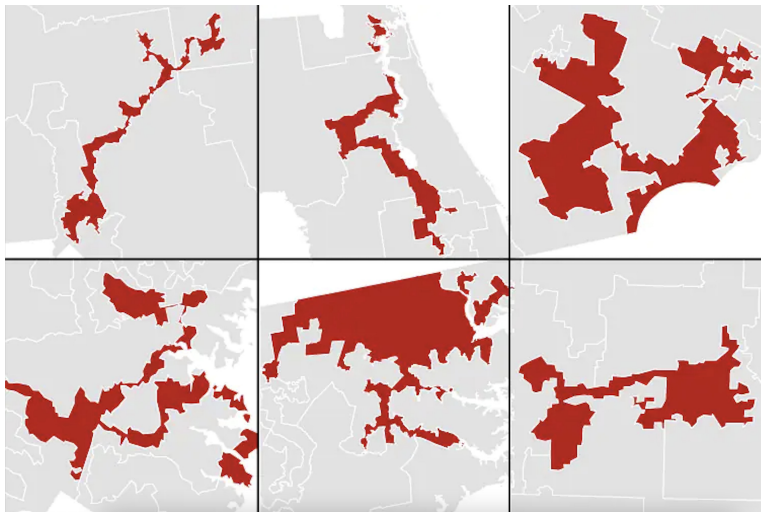
Nel caso del North Carolina si legge:

I legislatori Repubblicani hanno istruito i loro cartografi nell'usare i dati politici per ottenere una mappa che producesse una delegazione congressuale di 10 Repubblicani e 3 Democratici. ... Come dichiarato da uno dei due Repubblicani che controllavano la commissione: "Io penso che eleggere i Repubblicani sia meglio che eleggere i Democratici. Quindi ho disegnato la mappa per contribuire a promuovere ciò che credo meglio per il paese".

Nel caso del Maryland si legge:

L'allora Governatore Martin O'Malley ... ha testimoniato che il suo scopo era "usare la ridisegnazione dei distretti per cambiare la composizione complessiva della delegazione congressuale del Maryland in 7 Democratici e 1 Repubblicano, modificando un distretto. Si decise di agire sul Sesto [Distretto], che era stato tenuto dai Repubblicani per quasi due decenni. ... Il piano del 2011 si realizzò spostando circa 360.000 votanti fuori dal Sesto Distretto e inserendone 350.000 nuovi."

La forma del distretto elettorale del Massachusetts che ha segnato negativamente la memoria di Elbridge Gerry sembra meno grave di alcune suddivisioni di distretti viste in tempi recenti, ad opera di entrambi i partiti. La figura seguente mostra le mappe di alcuni distretti elettorali, negli Stati Uniti, che sembrano casi moderni di gerrymandering, evidenziati in un articolo del Washington Post del 2014.



Nell'articolo del Washington Post leggiamo:

La compattezza di un distretto – una misura di quanto irregolare è la sua forma, determinata dal quoziente tra l'area del distretto e l'area di un disco con lo stesso perimetro (N.B. Questa non è la definizione di compattezza in Matematica) – può essere un'utile indicazione di quanto [la forma di] quel distretto è frutto di gerrymandering. I distretti che hanno una forma generalmente regolare tendono ad essere compatti, mentre quelli che hanno molti “ghirigori”, diramazioni e protuberanze tentacolari hanno un risultato scadente, secondo questa misurazione.

Questo commento del Washington Post sulla compattezza, forse un po' ottimistico, pone la domanda su eventuali condizioni geometriche che permettano di definire a priori il gerrymandering.

Individuare con criteri geometrici il gerrymandering “intenzionale” non sembra però facile.

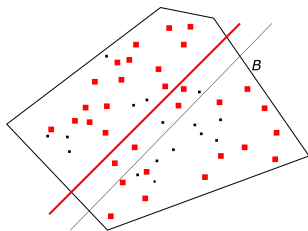
Infatti il gerrymandering non è necessariamente associato a “forme strane”.

Si possono sostanzialmente costruire situazioni di gerrymandering anche separando i distretti attraverso poche rette, creando cioè dei distretti che siano sostanzialmente dei poligoni convessi (quindi niente salamandre). Dimostriamo infatti la proposizione seguente.

Proposizione. Siano $n < r$ interi positivi e poniamo $8n$ quadrati piccoli e $8r$ quadrati grandi in un poligono convesso B , in modo che nessuna retta tocchi tre quadrati. Allora è possibile suddividere B in otto poligoni convessi in ciascuno dei quali il numero dei quadrati grandi è maggiore del numero dei quadrati piccoli (e quindi si crea una situazione di gerrymandering simile a quella del Piano 2).

In altre parole, supponiamo che i quadrati piccoli siano gli elettori del Partito QP , e i quadrati grandi siano gli elettori del Partito QG . Allora è possibile suddividere il poligono B , che rappresenta il paese, in otto poligoni convessi in modo che i quadrati piccoli e grandi si suddividano proporzionalmente, dando così al partito QG la vittoria in tutti i distretti.

Dimostrazione. Tracciamo due rette parallele: la retta più spessa separa l'insieme dei quadrati grandi a metà, la retta più sottile separa l'insieme dei quadrati piccoli a metà.



Se le due rette non coincidono, iniziamo a ruotarle (ad esempio in senso orario), mantenendole parallele e chiedendo che ciascuna di esse separi sempre a metà il suo insieme di quadrati. Dopo una rotazione di 180° ciascuna retta torna sostanzialmente nella sua posizione originaria. Poiché hanno fatto solo mezzo giro, ad un certo punto le due rette parallele devono essersi sovrapposte. Quindi esiste una retta che separa a metà sia i quadrati grandi, sia i quadrati piccoli. Chiamiamo B_1 and B_2 i due poligoni in cui questa retta ha suddiviso B . Ciascuno di essi contiene $4n$ quadrati piccoli e $4r$ quadrati grandi. Ripetiamo la costruzione precedente e suddividiamo B_1 and B_2 in poligoni più piccoli, contenenti ciascuno $2n$ quadrati piccoli e $2r$ quadrati grandi. Ancora un passo (analogo ai precedenti) e la dimostrazione è terminata.

Il gerrymandering non sembra quindi ridursi ad un problema geometrico, e richiede la collaborazione di matematici, statistici, avvocati, scienziati della politica, e informatici.

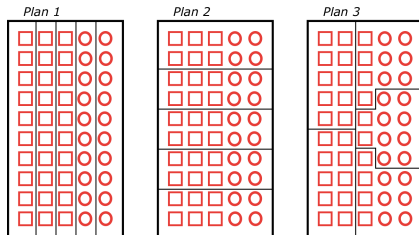
Accenniamo ora a due dei tentativi di definire (e quindi riconoscere) a priori il gerrymandering.

Divario di efficienza

Contiamo per ciascun partito il numero totale dei voti "inutili", cioè dei voti dati in distretti in cui ha vinto l'altro candidato, oppure, quando ha vinto il proprio candidato, dei voti oltre il 50% dei voti +1.

L'idea è che in una situazione equa i due partiti ottengano sostanzialmente lo stesso numero di voti inutili. Calcoliamo quindi la differenza di voti inutili tra i due partiti (o meglio, la sua percentuale).

Nel caso della figura vista all'inizio



il Partito Quadrato ha divario di efficienza +8% nel Piano 1 (poiché i Quadrati sprecano 12 voti, i Rotondi ne sprecano 8, quindi la differenza è 4 su 50, cioè 8%), -40% nel Piano 2, e +32% nel Piano 3 (un divario di efficienza fortemente negativo, come nel Piano 2, significa che è stato il Partito Rotondo a sprecare molti voti, ed è quindi un vantaggio per il Partito Quadrato).

Campionamento di insieme

Un approccio diverso, molto più complesso, ma più “ritagliato” sul singolo stato, è dato dal *campionamento di insieme*.

L'idea è di considerare tutte le possibili suddivisioni in distretti aventi un numero sostanzialmente uguale di elettori e soddisfacenti le condizioni specifiche dello stato. **Il numero di queste suddivisioni è enorme, di fatto intrattabile.** Poi si estrae in modo casuale un numero elevato, ma trattabile, di configurazioni dei distretti, ad esempio qualche decina di migliaia (l'estrazione casuale delle mappe è la parte più difficile). Successivamente si calcola il risultato delle elezioni conseguente a ciascuno di questi campionamenti, si fa una media e si confronta il risultato elettorale che sarebbe stato ottenuto sui campionamenti con il risultato reale. In questo modo l'eventuale gerrymandering dovrebbe essere evidenziato dal confronto con designazioni casuali dei distretti elettorali, non motivate da interessi di parte.

Qui, come spesso accade in matematica, il problema è dunque trovare una **definizione** accettabile di una famiglia di oggetti che già *vediamo* (i distretti oggetto di gerrymandering).

Bibliografia

-  Q. Berger, F. Caravenna, *Le paradoxe de Simpson illustré par des données de vaccination contre le Covid-19*, The Conversation, November 3, 2021.
<https://theconversation.com/le-paradoxe-de-simpson-illustre-par-des-donnees-de-vaccination-contre-le-covid-19-170159>
-  P. J. Bickel, E. A. Hammel, J. W. O'Connell (1975). *Sex bias in graduate admissions: Data from Berkeley*. Science, 187(4175), 398–404.
-  J. P. Caulkins, A. Barnett, P. D. Larkey, Y. Yuan, J. Goranson (1993). *The On-Time Machines: Some Analyses of Airline Punctuality*. Operations Research **41**, 710–720.
<http://www.jstor.org/stable/pdf/171966.pdf>
-  J. Ellemberg. Geometry, inference, complexity, and democracy, Bull. Amer. Math. Soc. **58**, 57–77, 2020.
<https://www.ams.org/journals/bull/2021-58-01/S0273-0979-2020-01708-8/S0273-0979-2020-01708-8.pdf>

-  L. Ford, *Fractions*, Amer. Math. Monthly, **45**, 1938, 586–601.
-  T. Grigg, *Simpson's Paradox and Interpreting Data*, Towards Data Science, Dec 9, 2018,
<https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765>
-  S. B. Guthery, *A motif of Mathematics. History and application of the mediant and the Farey sequence*, Docent Press, Boston, 2011.
-  A. Hatcher, *Topology of numbers*, preprint,
<https://pi.math.cornell.edu/~hatcher/TN/TNpage.html>
-  C. Ingraham, *America's most gerrymandered congressional districts*, Washington Post, May 15, 2014.
-  Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, **339**, 1973.
-  K. Katella, *Comparing the COVID-19 Vaccines: How Are They Different?*.
<https://www.yalemedicine.org/news/covid-19-vaccine-comparison>

-  W. Koehrsen, *Simpson's Paradox: How to Prove Opposite Arguments with the same data*, Towards Data Science, Oct. 12, 2018.
<https://towardsdatascience.com/simpsons-paradox-how-to-prove-two-opposite-arguments-using-one-dataset-1c9c917f5ff9>
-  M. Livio, *The golden ratio. The story of phi, the world's most astonishing number*. Broadway Books, 2002.
-  J. Morris, Israeli data: How can efficacy vs. severe disease be strong when 60% of hospitalized are vaccinated?. Covid-19 Data Science. August 2021.
<https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated>
-  D. Redmond, *Number Theory*. Marcell Dekker, 1996.
-  Supreme Court of The United States. 18-422,
https://www.supremecourt.gov/opinions/18pdf/18-422_9o11.pdf
-  Supreme Court of The United States. 21-1271. https://www.supremecourt.gov/opinions/22pdf/21-1271_3f14.pdf

-  P. Soberon. Gerrymandering, Sandwiches, and Topology, Notices Amer. Math. Soc. **64** (2017), 1010-1013. <https://www-ams-org.unimib.idm.oclc.org/journals/notices/201709/rnoti-p1010.pdf>
-  G. Travaglini, Numbers and Figures: Six Math Conversations Starting from Scratch, American Mathematical Society, 2023.
-  G. Travaglini, *Gerrymandering*, Emmequattro **86**, 2024.
<https://www.ilsussidiario.net/news/scienza-in-atto-matematica-e-sistemi-elettorali-il-caso-del-gerrymandering/2639083/>