

statistica: maneggiare con cura

Giuliano Spirito

Obiettivi dell'insegnamento della statistica:

- mettere gli alunni in condizione di comprenderne tecniche, importanza e utilità
- ma anche imparare a “diffidarne” (di guardare ad essa e all'uso che se ne fa **in modo critico**)

Su entrambi questi fronti, lo studio della statistica ha a che fare con

- un “**moderno saper fare di conto**”

Un po' di storia: i censimenti

- La prima attività statistica, e tuttora una delle principali, è costituita dai **censimenti** (dal latino *censere*, valutare).
- Le prime notizie riguardo a indagini sulla popolazione risalgono addirittura al **3800 a.C.**, presso i Sumeri; tali indagini erano mirate a misurare la quantità di uomini (in vista di guerre) e di beni (in vista di carestie) di cui si poteva disporre.

Nella **Bibbia** troviamo traccia di vari censimenti nel popolo ebraico:

- Il censimento fatto da Mosè:

Il Signore parlò a Mosè e a Eleàzaro, figlio del sacerdote Aronne, e disse: - Fate il computo di tutta la comunità degli Israeliti, dai venti anni in su, suddivisi secondo il casato dei loro padri, di quanti in Israele possono andare in guerra... (in Numeri, 26)

- Il censimento condotto da re Davide narrato in due punti della Bibbia, e sempre con una connotazione negativa:

“Ma dopo che ebbe contato il popolo, il cuore di Davide gli fece sentire il rimorso ed egli disse al Signore: - Ho peccato molto per quanto ho fatto, ti prego, Signore, toglì la colpa del tuo servo, perché io ho commesso una grande stoltezza -” (in 2 Samuele 24,1)

“Il fatto dispiacque agli occhi di Dio, che perciò colpì Israele” (in 1 Cronache 21,1)

- Il censimento degli stranieri effettuato da Salomone, figlio di Davide:

“Salomone censì tutti i forestieri che erano nella terra di Israele, un nuovo censimento dopo quello effettuato da suo padre Davide. Ne furono trovati centocinquantatremilaseicento.” (in 2 Cronache 2,17)

- Nell'**antica Roma** i censimenti venivano effettuati già dalla fine del VI secolo a.C.
- Tito Livio racconta come, nel 459 a.C., si sia concluso il decimo censimento dalla fondazione di Roma, dal quale risultarono 117.319 cittadini romani. (Pochi: la grande maggioranza di coloro che abitavano a Roma non erano cittadini romani).
- I **primi censimenti moderni**, che tengono conto di tutti gli abitanti in uno stesso momento e per tutto il paese, si hanno in Svezia nel 1749 e negli Stati Uniti dal 1790.
- In **Italia** il primo censimento risale al 1861. Da allora i nostri censimenti hanno cadenza decennale e da quasi un secolo vengono curati dall'Istat.

Ma veniamo a analizzare le modalità di svolgimento delle indagini statistiche, nelle loro varie fasi e i problemi connessi a ciascuna di queste fasi.

- Che cosa fa la statistica? Per analizzare una situazione raccoglie, organizza, interpreta dati.
- Ma, nonostante la sua fama di strumento asettico e votato alla pura e efficace quantificazione, **appartiene in qualche misura alla categoria della matematica dell'incerto** (per certi versi persino dell'**arbitrario**).

Raccolta di dati.

Già in fase di raccolta dei dati, cioè

→ nella predisposizione delle domande e

→ nella registrazione delle risposte

entrano in campo **scelte** che possono **influenzare** le risultanze dell'indagine.

Ad esempio intervengono sulle risposte:

→ la scelta della griglia in cui “costringere” i dati

→ la stessa formulazione della domanda

- Sul primo punto: **predisposizione della griglia.**
- Supponiamo di svolgere un'indagine *sul colore preferito da un gruppo di persone.*

Ragionevolmente fisseremo una **griglia di risposte** tra cui far scegliere l'intervistato, a evitare ambiguità e incertezze interpretative che rendono difficilmente gestibili le risposte.

- Ci regoleremo, cioè, esattamente secondo lo stesso approccio del pittore che deve dipingere le pareti delle varie stanze nel delizioso film del 1948 “La casa dei nostri sogni”, pittore che si propone certo di esaudire le richieste della padrona di casa, ma anche di difendersi da quanto in esse vi è di eccessivo e fuorviante.

- Ora, signor Piterson, parliamo un po' delle pareti

- Ok

- Innanzitutto il salotto. Lo voglio di un bel verde soffice. Non un verde blu come le uova del pettirosso...

- Nooo?!

- Ma neppure un verde giallastro come i boccioli del narciso. Ora l'unico campione che ho qui è un po' troppo giallo, ma non vorrei che si andasse all'estremo opposto e si facesse troppo blu...

- Nooo?!

- Dovrebbe essere una specie di giallo, verde e grigiastro. Ora la stanza da pranzo. La vorrei gialla, ma non un giallo qualunque, un giallo molto allegro, come fosse un raggio di sole. Sentite, signor Piterson, mandate uno dei vostri operai dal salumiere a comprare un etto di burro e copiatene il colore, non potrete sbagliare.

- Sì, ahaha

- Oh! Questa è la carta che metteremo in anticamera, è a fiori, in ogni modo non voglio che il soffitto richiami il colore dei fiori...

- Nooo?!

- Sullo sfondo ci sono dei piccoli puntini, voglio che il colore del soffitto li richiami. Non questo puntino verde però vicino all'agrifoglio...

- Nooo?

- Ma questo piccolo puntino bluastro tra il bocciolo di rose e quello della delfinia. E' tutto chiaro?

- Sì, sì.

- La cucina deve essere bianca, non di un bianco freddo, asettico, da ospedale, ma un po' più caldo, però non tanto da far supporre che non sia bianco.

- Sì, sì.

- Ora per lo spogliatoio, ecco qui. Riproducete il colore di questo filo e non perdetemelo, ne ho soltanto un rocchetto e mi ci è anche voluto molto a trovarlo. Come vedete è praticamente un rosso mela, una cosa di mezzo tra la mela winesap e la mela Jonathan.

- Sì, sì.

(Lei si allontana. Il signor Piterson si rivolge al suo aiutante)

- Hai preso nota?

- Verde, giallo, blu, bianco, rosso.

- Esatto



- Un altro esempio, sempre sul primo punto: *vogliamo indagare qual è l'età dei consumatori di un certo prodotto*. Ci converrà, per gli stessi motivi sopra detti raccogliere i dati per fasce di età.
- Ma **la scelta delle fasce** (ad esempio: meno di 20, 21-40, 41-60, più di 60 o invece 11-25, 26-40, 41-55, 56-70, 70-85, più di 85) **interverrà a determinare gli esiti** (e quindi forse la “lettura” delle risultanze) dell'indagine statistica...

Sul secondo punto: **formulazione della domanda**

- E qui facciamo un esempio di scuola (Tversky e Kahneman citati da John Allen Paulos in “Gli snumerati”, Edizioni Leonardo)

Immaginate di essere un generale e di essere circondati dal nemico. Per evitare che il vostro esercito di 600 uomini sia annientato avete due sole vie di scampo. I vostri consiglieri vi spiegano che se prendete la prima salverete 200 soldati (400 dei vostri soldati moriranno); se invece prendete la seconda c'è 1 probabilità su 3 che tutti i 600 soldati ce la facciano (nessuno muoia), mentre in 2 probabilità su 3 nessuno si salverà (che tutti i soldati periscano).

Ebbene:

→ nella prima formulazione il **75%** sceglie la prima via di scampo

→ nella seconda formulazione solo il **20%** sceglie la prima di scampo

- Un altro esempio sul secondo punto:

immaginiamo di condurre un'indagine sulla posizione rispetto alla pena di morte.

Il modo in cui formuleremo il quesito può influire in modo decisivo sul risultato.

Non ho le prove ma non mi meraviglierei che si pervenga a conclusioni opposte a seconda che la domanda sia formulata in questo modo:

- *Sei favorevole a introdurre la pena di morte per gli autori di gravi reati?*

o in quest'altro modo

- *Ritieni giusta la pena di morte per gli autori di gravissimi delitti quali stragi, femminicidi, infanticidi?*

→ Specialmente in indagini che coinvolgono espressioni di opinioni, le scelte più o meno consapevoli di chi conduce l'indagine "guidano" (a volte a "pilotano") le risposte...

Scelta del campione

Le indagini statistiche in cui dati vengono raccolti sull'intera popolazione sono impegnative, in termini economici e in termini di tempo; perciò sono rare (i censimenti periodici e, a ben pensarci, anche le elezioni).

Già dalla seconda metà dell'Ottocento l'indagine statistica viene condotta su un **campione**, cioè un sottoinsieme della popolazione.

È però evidente che l'attendibilità dei risultati, in un'indagine per campione, è legata:

- da una parte alla sua **consistenza numerica**: da questo punto di vista, più il campione è ampio, meglio è.
- dall'altra alla sua **rappresentatività**: più il campione è "casuale" meglio "rappresenta" le varie caratteristiche della popolazione, (che cosa significhi poi davvero il termine "casuale" è davvero un bel problema!)

Nel Novecento si è provveduto a cercare di fornire una base "scientifica" ai criteri di scelta del campione; resta però un certo tasso di arbitrarietà o almeno di **opinabilità** nelle scelte che si compiono per **selezionare un campione**.

L'amara verità è che un campione non è mai del tutto rappresentativo. L'unico campione che garantisce davvero di rispecchiare perfettamente l'intera popolazione è il campione che... coincide con l'intera popolazione!

I valori di sintesi

I valori di sintesi (**media**, **moda**, **mediana**), sono preziosi strumenti per non “perdersi” in una mole ampia di dati registrati.

Una riproduzione troppo dettagliata, infatti, finisce spesso per essere inutilizzabile, come ci dice lo scrittore argentino Jorge Luis Borges ne “L’artefice”:

“... In quell'Impero, l'Arte della Cartografia raggiunse tale Perfezione che la mappa di una sola Provincia occupava un'intera Città, e la mappa dell'Impero un'intera Provincia. Col tempo queste Mappe Smisurate non soddisfecero più e i Collegi dei Cartografi crearono una Mappa dell'Impero che aveva la grandezza stessa dell'Impero e con esso coincideva esattamente. Meno Dedite allo Studio della Cartografia, le Generazioni Successive capirono che quell'immensa Mappa era Inutile e non senza Empietà l'abbandonarono alle Inclemenze del Sole e degli Inverni...”

D'altra parte occorre tenere sempre presente che un valore di sintesi ci fornisce un'informazione molto "ridotta" e parziale rispetto all'insieme dei dati che sintetizza.

Ce lo dice con ineguagliata efficacia un celeberrimo sonetto di Trilussa:

...

*Me spiego: da li conti che se fanno
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.*

La scelta di un valore di sintesi

Oltre all'ineliminabile parzialità di un valore di sintesi (come parziale è ogni ricapitolazione, ogni riassunto), c'è un altro aspetto su cui riflettere:

la scelta di uno o di un altro valore di sintesi può risultare tutt'altro che neutrale nell'esame delle risultanze

Un esempio "scolastico": supponiamo che l'alunno X abbia preso i seguenti voti:

6 1 6 1 6 2 6

Nello scrutinio finale, un docente che vuole giungere alla promozione potrà legittimamente sostenere che il voto più frequente (**la moda**) riportato dall'allievo è 6 (voto che si presenta ben 4 volte su un totale di 7 prove) → promosso

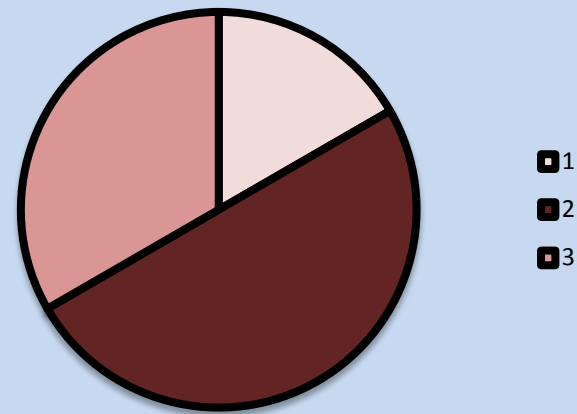
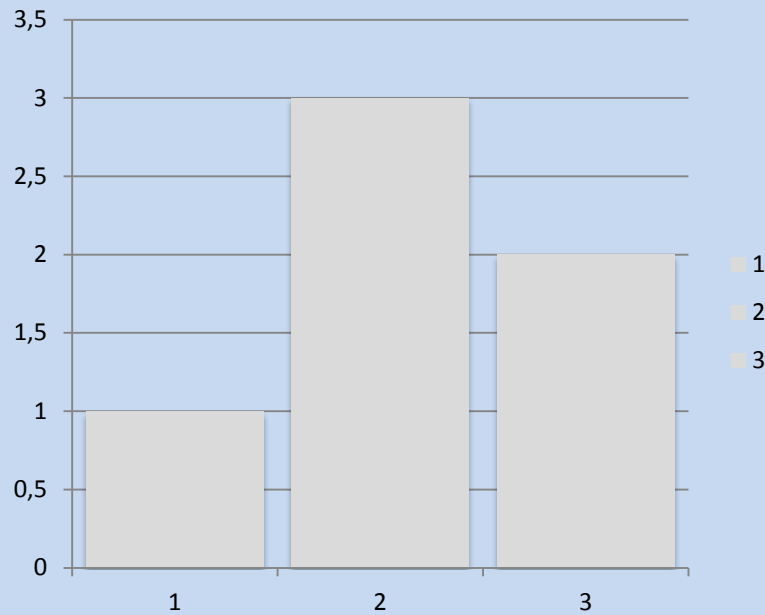
Mentre un altro docente che vuole giungere a una conclusione opposta potrà altrettanto legittimamente sottolineare che **la media** riportata dall'alunno nelle 7 prove è $(6+1+6+1+6+2+6)/7$, cioè $28/7$, cioè 4 → non promosso

Vedete bene come la scelta di un diverso valore di sintesi può addirittura decidere se il povero X avrà diritto o no all'agognato motorino, con conseguenze a catena sulle prospettive dell'industria italiana dei veicoli a due ruote!

Rappresentazione dei dati

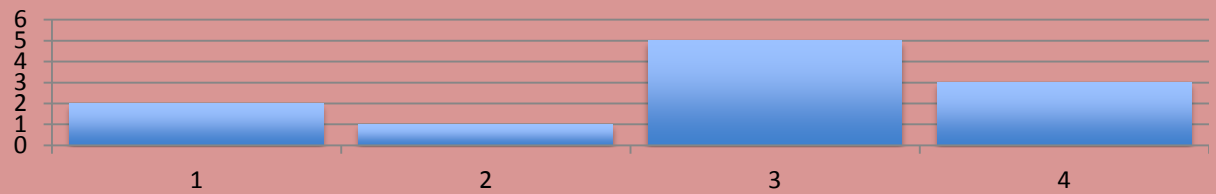
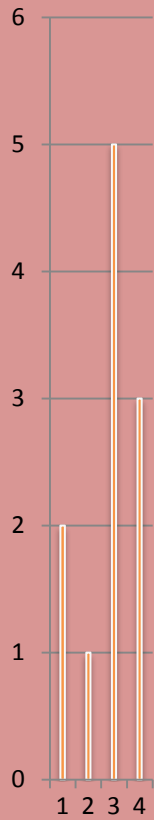
Anche la fase di **rappresentazione** delle risultanze statistiche contiene degli elementi di arbitrarietà:

→ i dati raccolti possono essere rappresentati in più modi, con elementi di suggestione molto diversi...



Le rappresentazioni raccontano la stessa situazione?

E queste due rappresentazioni?



Correlazione e causalità

Un'altra insidia della statistica si nasconde nella **correlazione** che si stabilisce quando due grandezze “variano insieme” (co-variano).

Ad esempio: *prezzo e domanda di una merce, statura e peso di un individuo, ma anche: lettura abituale di un certo quotidiano e opzione politica, nervosismo dei genitori e capricciosità dei figli.*

La scoperta di una correlazione non permette però di per sé (e comunque questo non è compito della statistica) di decidere ***cosa determina cosa***.

Di fatto, invece, la correlazione viene frequentemente usata, in modo azzardato se non scorretto, per indurre da essa un presunto rapporto lineare di **causa-effetto**.

Molto spesso, in realtà, la faccenda è più complessa (***genitori nervosi → figli capricciosi, ma anche figli capricciosi → genitori nervosi!***) e la dipendenza tra il variare di una grandezza e dell'altra è assunta arbitrariamente da colui che evidenzia la correlazione.

→ Qui il terreno è davvero scivoloso. Però anche in questo caso, la responsabilità non è della statistica, ma del suo uso improprio e sprovveduto se non capzioso...

Attenzione, attenzione...

Brano tratto da un sussidiario, a proposito di educazione ambientale:

“Anche se la superficie terrestre è coperta per il 71% di acqua, questa è costituita per il 97,5% da acqua salata. L’acqua dolce è per il 68,9% contenuta in ghiacciai e nevi perenni, per il 29,9% nel sottosuolo e solo lo 0,3% è localizzata in fiumi e laghi, e quindi potenzialmente disponibile.”

$$\rightarrow 68,9+29,9+0,3 = 99,2$$

Da “la Repubblica”:

“Web o librerie. Senti chi vince.

Le librerie restano il principale canale per vendere i libri: vi acquistiamo il 76% dei libri che leggiamo contro il 29% comprati su Amazon o altri store online. E’ quanto emerge da una ricerca dell’Associazione Italiana Editori.”

$$\rightarrow 76+29 = 105$$

Una raccomandazione: non buttare il bambino con l'acqua sporca...

- Abbiamo visto una serie di “precauzioni per l'uso” da adottare rispetto alla statistica.
- Ma allora sarà forse meglio evitare di introdurre nella scuola la statistica?

Certo che **no!**

→ No, perché la statistica è occasione per svolgere **attività con alta valenza formativa, non banali e coinvolgenti.**

→ E di nuovo no, anche perché l'acquisizione dei principi di prudenza di cui abbiamo ampiamente parlato è parte integrante di un progetto di **matematica per la cittadinanza.**

Per concludere

Per concludere due “indagini” statistiche.

- La prima, sgangherata e scanzonata, surreale e esilarante, ovviamente del tutto inattendibile, è quella “narrata” da **Nino Frassica** nella trasmissione televisiva “Indietro tutta 30 e l’ode”.
- La seconda, poetica e suggestiva (e con dentro molta verità), è opera della poetessa polacca **Wislava Szymborska**, premio Nobel per la letteratura nel 1996.

(Alla Szymborska dobbiamo anche, sempre in ambito “matematico”, la poesia “Pi greco”, insolito omaggio a questo numero: *“E’ degno di ammirazione il Pi greco...”*)



Son passati 30 anni da Indietro tutta, 30 anni, a me mi sembrano 29. Quanto successo; si parlò di grandi ascolti, grossissimi ascolti [...] si parla di 37 milioni di telespettatori, 37, quasi 100 milioni ...

30 anni fa l'auditel ci premiò; l'auditel 30 anni fa era fatta a mano: c'era un incaricato per ogni città d'Italia, la mattina passava sera per sera per 7 sere passava per e col suo block-note segnava "lei signora cosa ha visto ieri" segnava, e lei cosa ha visto, segnava tutto. Di solito l'incaricato lo si sceglie una persona affidabile possibilmente seria, sposato, padre di famiglia con una sola testa sopra una sola spalla. Si parla di ascolti orribilmente alti. L'odiens, odiens ma perché odiens se ci fecilitavano, e Indietro tutta raggiunge la bella cifra dello share del 106%: su 100 italiani 106 italiani vedevano Indietro tutta, la trasmissione vista in tutte le televisione italiane e estere. Allora ci fu una statista della Oxa. Fece la statista e risultò: tv accese 49 milioni, tv spente 155, pochissime, tv portate a riparare 1008, uomini 70%, donne 71%, esseri umani 96%, esseri disumani 11%, bambini accompagnati 45%, 700 sindaci, 5000 carabinieri, 888 critici, lettere in arrivo, nebbia in val Padana e complimenti per la trasmissione.

Contributo alla statistica

di Wislawa Szymborska (Nobel per la letteratura 1996)

Su cento persone:

che ne sanno sempre più degli altri – cinquantadue;

insicuri a ogni passo – quasi tutti gli altri;

pronti ad aiutare, / purché la cosa non duri molto – ben quarantanove;

buoni sempre, / perché non sanno fare altrimenti – quattro, be', forse cinque;

propensi ad ammirare senza invidia – diciotto;

viventi con la continua paura / di qualcuno o qualcosa – settantasette;

dotati per la felicità – al massimo poco più di venti;

innocui singolarmente, / che imbarbariscono nella folla – di sicuro più della metà;

crudeli / se costretti dalle circostanze – è meglio non saperlo / neppure approssimativamente;

quelli col senno di poi – non molti di più / di quelli col senno di prima;

che dalla vita prendono solo cose – quaranta, / anche se vorrei sbagliarmi;

ripiegati, dolenti / e senza torcia nel buio – ottantatre, / prima o poi;

degni di compassione – novantanove;

mortali – cento su cento. / Numero al momento invariato.