DOI: 10.13137/2039-8646/36724

L'insegnamento della Statistica in ambito multidisciplinare. Parte prima[*]

LUCIO TORELLI
Dipartimento Universitario Clinico di
Scienze Mediche, Chirurgiche e della Salute
Università di Trieste
torelli@units.it

Dedicato a Maria Chiara Magri, giovane ed acuta ricercatrice in Statistica medica, abile divulgatrice di argomenti non banali. Amica sincera con cui più volte abbiamo lavorato per presentare questi temi a platee di studiosi di discipline anche lontane dalle nostre. In aula ancora oggi presento esempi nati dalle nostre discussioni, dal desiderio di trasmettere la passione per il nostro lavoro e per le nostre ricerche.

ABSTRACT

Statistics can be a useful tool to help students grow and train. School can and must increasingly be a reality that helps and encourages the development and growth of people: not just memorizing notions soon forgotten, but a way to training oneself to think, individually and with others. This is one of the important challenges we have as teachers. We are surrounded by statistics yet we know little about it. In this paper, through my experience with students of medicine and healthcare degree courses, I want to highlight the role of Statistics in various disciplines. Here I will present a simple and general introduction to Statistics and to the so-called descriptive Statistics. A second part will be dedicated to Statistics and Probability Calculus. Finally, I will include some notes on the so-called inferential Statistics. I hope this text can be an opportunity for dialogue between University and School, dialogue where everyone has something important to say based on the experience we have in teaching and in research activity. I believe that a reciprocal relationship is essential in this era of rapid changes, changes that place us at the forefront in helping our students to grow and to orient themselves with greater confidence and passion towards the future.

PAROLE CHIAVE

STATISTICA / STATISTICS, ELABORAZIONI STATISTICHE / STATISTIC; STATISTICA DESCRITTIVA / DESCRIPTIVE STATISTICS; DIDATTICA INTERDISCIPLINARE / INTERDISCIPLINARY TEACHING.

1. INTRODUZIONE

Da alcuni anni insegno *Statistica medica* e svolgo ricerche in collaborazione con

[*] *Title: Teaching Statistics in a multidisciplinary field. Part one.*

colleghi medici e sanitari e mi colpisce soprattutto l'interesse che questa disciplina attira in ambiti molto diversi, non solo in corsi di laurea universitari.

Mi sono trovato spesso, ad esempio, a tenere delle conferenze, sia a docenti, sia a studentesse e studenti delle scuole secondarie. Queste occasioni di incontro con le scuole sono momenti importanti: da parte delle ragazze e dei ragazzi c'è l'interesse a capire meglio una disciplina che li coinvolge in molte attività quotidiane, ma che non sempre hanno studiato, dall'altra c'è il desiderio delle e dei docenti di saperne di più, di essere pronti ad affrontare in aula temi su cui talvolta ci si può sentire impreparati. Raramente in classe c'è la possibilità di affrontare argomenti specifici di Statistica, ma, al tempo stesso, siamo quotidianamente a contatto con elaborazioni di dati, e questo accade in discipline anche molto diverse, non solo necessariamente in Matematica o Fisica. Siamo bombardati da statistiche[1] ma non ci sentiamo sufficientemente attrezzati per capire quali siano quelle corrette e quali siano quelle fatte in modo affrettato o sbagliate[2].

Il cellulare a fine mese fa un rendiconto dei nostri spostamenti; i dati più frequenti raccolti sui siti che visitiamo o sulle parole che cerchiamo in un motore di ricerca indirizzano pubblicità e suggerimenti per i nostri acquisti; i dati raccolti in tempo reale ci permettono di sapere la situazione del traffico e di conseguenza possono aiutarci a decidere le strade da evitare; nello sport gli atleti vengono presentati in base alle loro statistiche; sempre di più molti indicatori ci permettono di valutare alcune situazioni o di capire la qualità del nostro operare; facciamo scelte in base a dei *ranking*, a delle classifiche. E potremmo continuare a lungo.

Viviamo ogni giorno, e in molte attività quotidiane, di elaborazioni di dati e il rischio alla fine è quello di non avere in mano la situazione, di subire le statistiche, senza avere la possibilità di capire, di conoscere a fondo quanto i dati stanno raccontando (o quanto non stanno raccontando).

[1] Da notare che uso la parola Statistica con la S maiuscola per indicare la disciplina, mentre uso la s minuscola per indicare quando si fa una statistica, un "conto", un consuntivo. In Inglese, si è soliti usare i termini *Statistics* e *statistic*.

[2] Segnalo qui il recente testo, *Molte statistiche ma poca Statistica*, che ho scritto per il volume *Cura e reciprocità - molti saperi per un contributo dialogico sulla responsabilità come nuovo paradigma di cura*, volume che nasce da un'interessante collaborazione interdisciplinare con colleghi medici (cfr. TORELLI 2024).

Non è raro sentire parlare di statistiche strane, oserei dire balorde, frutto di ragionamenti affrettati e non corretti, statistiche che fanno sorridere, in quanto senza senso[3]. Daniel Pennac nel suo *Diario di Scuola*[4] scriveva in maniera simpatica e preoccupata: «Statisticamente tutto si spiega, personalmente tutto si complica».

Si racconta che Woody Allen, astuto e provocatore, avesse commentato: «Il 94,5% delle statistiche è sbagliato»[5], frase che solitamente in aula rendo ancora più provocatoria dicendo: «*È stato statisticamente dimostrato* che il 94,5% delle statistiche è sbagliato».

E ancora, troppo spesso la Statistica è vista come uno "strumento che porta alla verità", che fornisce una risposta dicotomica: un bianco o un nero, un sì o un no, un dentro o un fuori. Se non capiamo una certa situazione, ci affidiamo alle statistiche, pensando che queste ci toglieranno ogni dubbio, che ci daranno delle certezze.

La frase sopra citata «è stato statisticamente dimostrato» ci rassicura ma, al tempo stesso, non credo ne capiamo a fondo il significato. «Vorremmo avere certezze, ma abbiamo solo probabilità», così saggiamente commenta Albert Einstein nel film *Oppenheimer*[6], come risposta a una domanda dello scienziato americano. Dobbiamo pensare a una Statistica diversa, che non sia solo un freddo strumento di calcolo[7].

In aula, per presentare alcuni argomenti di Statistica e per coinvolgere la classe in discussioni critiche e costruttive, utilizzo, tra l'altro, la piattaforma *wooclap*[8], con cui formulo delle domande: è un modo che di solito coinvolge tutti, anche in situazioni di classi molto numerose.

Chiedo, ad esempio, quanto ci fidiamo delle statistiche, oppure perché le statistiche sono importanti o, ancora, cosa pensiamo di sapere già di Statistica. In classe le

[3] Suggerisco a questo riguardo il bel libro dal titolo *Mentire con le statistiche* (cfr. HUFF 2009). In rete è possibile trovare alcuni dei brevi filmati con esempi tratti da questo testo.
[4] Cfr. PENNAC 2008.
[5] Da notare la cifra decimale in 94,5%, che sembra dare maggior peso all'affermazione.
[6] Cfr. OPPENHEIMER in Filmografia.
[7] Si raccomanda ad esempio che in una sperimentazione lo statistico non venga contattato a dati raccolti, quando bisogna solo eseguire dei calcoli; sarebbe come chiamare un ingegnere, un architetto o un geometra quando una casa è già in costruzione: è essenziale lavorare insieme fin da principio, per mettere le basi del progetto e della struttura dello studio. Solo con un approccio metodologico di questo tipo sarà poi possibile raccogliere e analizzare i dati.
[8] Cfr. WOOCLAP in Siti web.

ragazze e i ragazzi rispondono in forma anonima attraverso il proprio cellulare e le risposte vengono raccolte dal software e proiettate sullo schermo in aula.

Colpisce il fatto che di solito le ragazze e i ragazzi rispondono positivamente a queste domande: dicono in generale di credere alle statistiche, cosa che di solito non avviene quando faccio le stesse domande a gruppi di persone più adulte. Mi piace constatare questa maggiore fiducia da parte dei giovani rispetto a noi adulti, che forse siamo prevenuti rispetto a quanto leggiamo sulle elaborazioni dei dati. È importante valorizzare questo pensiero e questa positività dei giovani, come pure è importante aiutare ed aiutarci a capire come va letta una statistica e come si possa verificarne la sua correttezza.

A lezione mi piace ripetere che una statistica può essere non corretta per ignoranza, quando non è stata fatta in maniera corretta la raccolta dei dati o non sono stati usati strumenti corretti di Statistica, oppure per astuzia, quando si vuole pilotare una statistica per arrivare comunque a un risultato prefissato. Sta a noi la capacità di capire la correttezza di una statistica, all'interno del contesto in cui si sta lavorando.

Sono molto interessanti le risposte delle studentesse e degli studenti alla domanda su cosa pensano possa servire la Statistica: tra i commenti che spesso vedo, ne evidenzio alcuni significativi. Scrivono, ad esempio: la Statistica serve «a capire meglio la realtà»; «capendo cosa è già successo in passato, sapere come comportarsi in futuro»; «conoscere, rendere consapevoli le persone tramite dati e informazioni utili». Sono frasi positive che fanno intravedere delle aspettative e che sottolineano ancora una volta la necessità di conoscere meglio la Statistica e di imparare a utilizzarla in maniera corretta, evitando di arrivare a conclusioni affrettate.

Dopo questa premessa, penso sia utile mettere in luce alcune idee di base, conoscere almeno alcuni concetti iniziali. Come anticipato, questo semplice manoscritto nasce dall'esperienza di insegnamento in questi anni in corsi di Laurea in Medicina, in Odontoiatria e nelle professioni sanitarie, corsi di Laurea in cui di solito la Statistica non è particolarmente gradita, ma che, alla fine, mi pare riesca a sorprendere le studentesse e gli studenti: ci si aspetta un corso pieno di formule da imparare e da

utilizzare e si ritrova invece una disciplina che aiuta a ragionare davanti ai problemi, con applicazioni utili e interessanti.

Gli argomenti che tratterò sono solo alcuni, primissimi spunti per imparare a leggere una statistica con senso critico. Sono spunti che possono essere di aiuto anche nel fare alcune prime elaborazioni di dati: semplici statistiche che possono essere una sorta di palestra in cui esercitarsi.

Sono esempi e strumenti di base che non richiedono di solito conoscenze avanzate di Matematica e questo permette di presentarli anche in contesti diversi e in diversi livelli della Scuola secondaria. Mi sorprende il fatto che quando tengo dei seminari nelle Scuole, come pure in diverse Aziende, mi si richiede sempre di partire da elementi iniziali di Statistica, di rimettere a fuoco le basi della disciplina, in quanto non sono chiare nel loro significato e nel loro uso.

Anche per questo motivo, come anticipato, in questo primo lavoro mi voglio concentrare solo su alcuni elementi di *Statistica descrittiva*, lasciando a contributi successivi le considerazioni sul *Calcolo delle Probabilità* e sue applicazioni alla *Statistica* e quelle sulla *Statistica inferenziale.* Non ho ritenuto opportuno presentare qui un'introduzione classica ed esaustiva della Statistica: ci sono già testi molto validi che affrontano la Statistica a questo riguardo.

Questa presentazione nasce soprattutto dal continuo confronto con colleghi di altre discipline e credo che proprio questo rapporto interdisciplinare dia un valore aggiunto al nostro lavoro di ricerca e, di conseguenza, a quanto insegniamo ogni giorno in aula. La collaborazione tra diverse discipline è un elemento vincente nell'affrontare e nel presentare nuove tematiche in aula e aiuta le studentesse e gli studenti ad avere una visione più aperta di cosa significhi lavorare insieme, cosa che poi tornerà estremamente utile anche quando affronteranno il mondo del lavoro.

Ho scelto in questo breve testo di avere un approccio più divulgativo, partendo da esempi semplici e da applicazioni pratiche. Questa modalità, importante nel mio contesto di insegnamento in un Dipartimento di Scienze Mediche, penso possa essere

utile anche in ambiti scolastici in cui la Statistica può essere introdotta anche in maniera occasionale e non sistematica: può capitare infatti di commentare in aula una notizia, di capire cosa ci sta dicendo una statistica, di dover rispondere a delle domande degli studenti.

Mi auguro che la Statistica possa essere di aiuto a quei docenti che non hanno avuto la possibilità di studiarla approfonditamente nei corsi universitari, che si possa insegnarla sempre più in maniera coinvolgente, che possa essere motivo di maggiore comprensione e conoscenza della realtà.

Qualche anno fa, una docente di scuola secondaria osservò, alla fine di una mia conferenza, che la Statistica presentata in questo modo può essere proposta anche come una lezione di Educazione civica. Mi auguro allora che la Statistica possa essere vista come un importante strumento di calcolo, solo se prima viene contestualizzata da un punto di vista metodologico nell'ambito in cui ci si trova ad applicarla. In tal modo potrà aiutare a essere cittadini più consapevoli, preparati a prendere decisioni e a fare delle scelte per il futuro.

2. Esempi ed elementi di base di Statistica descrittiva

Spero che gli argomenti qui trattati, anche se di base, possano essere degli spunti utili a coloro che desiderano conoscere un po' di più questa disciplina e che possano essere elementi di scambio e di discussione per le attività quotidiane di insegnamento.

2.1 Anche una semplice percentuale può essere ambigua e deviante

La Statistica descrittiva, lo dice l'aggettivo, permette di descrivere i dati raccolti, dati che non sono solo numeri: possono infatti essere anche parole, ordinabili o meno, numeri ordinali e così via.

La descrizione dei dati deve tenere conto *di chi* li ha raccolti, *di quando* questi sono stati raccolti, *della modalità* della raccolta e *di quanti non sia stato possibile* raccogliere il valore. Una cosa è, ad esempio, ciò che il paziente mi riferisce circa il suo peso, altro

è avere la possibilità di misurare il peso stesso[9]. Questi elementi metodologici, uniti alla conoscenza di alcuni strumenti per descrivere i dati, sono primi passaggi importanti.

Non sempre è immediato saper scegliere un grafico o una tabella opportuni che riescano a comunicare quanto il dato ci dice, come pure non a tutti è chiaro quando usare la *media* piuttosto che la *mediana*, o la *deviazione standard* piuttosto che la *distanza interquartile*.

Inoltre, se andiamo a leggere molte statistiche, i dati vengono presentati o riassunti attraverso l'uso di *percentuali*, concetti matematici semplici che impariamo a conoscere fin dalle Scuola primaria. Perfino le percentuali possono, però, nascondere delle insidie, volute o non volute - come scrivevo sopra - per ignoranza o per astuzia: possono cioè non rappresentare fedelmente quanto si pensa di descrivere.

Dire, ad esempio, che una certa percentuale delle nostre coste è dato in concessione balneare non chiarisce se tale percentuale è stata calcolata rispetto a tutte le coste presenti nel nostro Paese, oppure solo a quelle realmente utilizzabili e accessibili per i bagnanti via terra.

Oppure, dire che al ballottaggio per l'elezione di un Sindaco il vincitore ha avuto il 60% dei voti non significa che ha scelto quel candidato il 60% degli aventi diritto, ma il 60% dei votanti. Se supponiamo che al ballottaggio si siano presentati solo il 40% dei cittadini, dobbiamo concludere che il nuovo Sindaco ha avuto, sugli aventi diritto, solo il 24% dei voti (cioè il 60% del 40%). Anche se tale esempio può sembrare banale, dobbiamo osservare che dati di questo tipo vengono spesso presentati e comunicati in maniera scorretta, magari volutamente.

Un altro semplice esempio lo riprendo da una notizia apparsa su un quotidiano alcuni anni fa, in cui i dati di due anni successivi delle matricole nelle Scuole di una certa città portavano a delle considerazioni sul lavoro di orientamento scolastico e su eventuali successi di sperimentazioni innovative. Il quotidiano presentava la Tabella 1 e commentava che le Scuole B e H erano le uniche in cui c'era stato un aumento di

[9] Accade, ad esempio, che spesso venga dichiarato un peso minore e un'altezza maggiore!

matricole. Tutte le altre Scuole, a parte la C il cui numero di matricole era rimasto invariato, avevano avuto diminuzioni più o meno importanti. L'elaborazione di questi dati portava quindi a vari commenti sulle scelte delle studentesse e degli studenti, sulla capacità di rendere i propri corsi di studio attrattivi.

	2022 - 23	2023 - 24	
A	92	59	-55,9%
B	250	280	10,7%
C	87	87	0,0%
D	178	166	-7,2%
E	250	240	-4,2%
F	120	105	-14,3%
G	96	80	-20,0%
H	117	125	6,4%
I	76	53	-43,4%
L	195	146	-33,6%

Tabella 1. Numero di matricole e variazione percentuale per righe.

La variazione percentuale della terza colonna delle tabelle è stata calcolata dividendo la differenza tra il dato relativo all'anno scolastico 2023-24 e quello del 2022-23, rispetto al dato del 2023-24. Ad esempio, per la Scuola A si ottiene : (59-92)/59.
Da una lettura più attenta, come si vede dalla tabella, quasi tutte le scuole sono andate in perdita. Ci si accorge, però, che tra i due anni scolastici c'è stato un calo complessivo di matricole: 1461 contro 1341. Le percentuali appena calcolate, pur algebricamente giuste, non tengono conto di tale variazione sui totali e pertanto bisogna in qualche modo correggerle.
Potremmo ad esempio procedere così: la Scuola A passa da 92 matricole su un totale di 1461 a 59 su 1341 e quindi passa dal 6,3% al 4,4%. A questo punto calcoliamo la

variazione relativa tra questi due nuovi valori: (4,4%-6,3%)/4,4%. Otteniamo un valore percentuale diverso, non più -55,9% ma -43,1%.

Lavorando allo stesso modo sugli altri dati, otteniamo la Tabella 2, in cui risulta, tra l'altro, che le Scuole D ed E non hanno una variazione negativa ma positiva: la Scuola D passa da un -7,2% a un +1,6%, la Scuola E, da un -4,2% a un +4,4%. La Scuola C, che ha mantenuto lo stesso numero di matricole nonostante il calo totale, ottiene un +8,2%.

	2022 - 23	2023 - 24	
A	6,3%	4,4%	-43,1%
B	17,1%	20,9%	18,0%
C	6,0%	6,5%	8,2%
D	12,2%	12,4%	1,6%
E	17,1%	17,9%	4,4%
F	8,2%	7,8%	-4,9%
G	6,6%	6,0%	-10,1%
H	8,0%	9,3%	14,1%
I	5,2%	4,0%	-31,6%
L	13,3%	10,9%	-22,6%

Tabella 2. Percentuali sul totale di colonna e corrispondenti variazioni percentuali per riga.

Due statistiche, matematicamente corrette e che lavorano con gli stessi dati, che arrivano a due conclusioni differenti! Come decidere quale sia quella corretta, cioè quella che descrive cosa è realmente successo? Non basta quindi eseguire dei calcoli, usare una formula: per fare Statistica, è necessario applicare, contestualizzare il tutto nella situazione che si sta descrivendo.

Traggo un ulteriore esempio dal bel libro dei colleghi Sgarro, Franzoi, Vicig[10],

[10] Cfr. SGARRO, FRANZOI, VICIG 2022.

esempio che prende spunto da una situazione reale:

A e B sono due Paesi di circa 10 milioni di abitanti ciascuno; sia in A sia in B durante una pandemia sono state ospedalizzate circa 10 mila persone. In A, che è un Paese ricco e con un buon sistema sanitario, sono morte 310 persone tra quelle ospedalizzate, quindi il 3,1%. In B, che è un Paese povero e con un sistema sanitario non buono, sono morte 270 persone, quindi il 2,7%. La notizia colpisce l'opinione pubblica: non ce lo aspettavamo, ma i dati sembrano dire che è andata meglio in B che non A!

Ho detto volutamente "sembra". Infatti, da un'analisi più attenta, si scopre che A ha bassa natalità e alta aspettativa di vita: sono morti 10 giovani su 1000, l'1%, 60 adulti su 3000, il 2%, 240 anziani su 6000, il 4%.

B ha invece alta natalità e bassa aspettativa di vita: sono morti 100 giovani su 5000, il 2%, 120 adulti su 4000, il 3%, 50 anziani su 1000, il 5%. Questa seconda analisi ci dice, al contrario della precedente, che è andata meglio in A![11]

Come leggiamo, cosa capiamo e cosa comunichiamo di queste statistiche, che altro non sono che semplici percentuali, calcolate entrambe in maniera corretta? Qual è la statistica che meglio racconta i dati raccolti?

Nel prossimo manoscritto parleremo anche di altre percentuali, che tutti abbiamo sentito nominare e che abbiamo usato per diagnosticare il COVID, percentuali che ci raccontavano che il tampone molecolare funzionava meglio di quello salivare: sono la sensibilità, la specificità e i valori predittivi di test diagnostici.

Vedremo che la *Sensibilità* è la probabilità che una persona malata risulti positiva a un test e che può essere calcolata come rapporto tra veri positivi e malati. Il *valore predittivo positivo* è invece la probabilità che una persona positiva sia malata e la calcoleremo come rapporto tra veri positivi e test positivi. In maniera analoga verranno definiti la *Specificità* e il *valore predittivo negativo* e vedremo come tali percentuali sono un'applicazione importante di quella che definiamo *probabilità condizionata*.

[11] Come fanno notare gli autori, una situazione di tale tipo è un esempio del cosiddetto *paradosso di Simpson*, un fenomeno statistico che, come vedremo, si verifica quando la relazione tra due fenomeni appare modificata o invertita, paradosso che è alla base di errori nel considerare con poca attenzione le analisi statistiche.

2.2 La Statistica non è solo fare il calcolo della media

Capita in aula di parlare di Statistica ricordando la bella poesia *La statistica* di Trilussa, poesia che per molti anni apostrofava come "polli di Trilussa" le statistiche fatte male[12]. Troppo spesso molte statistiche si fermano a calcolare solo la media dei dati raccolti[13]. Come convincere le studentesse e gli studenti che la *media*, pur calcolata in maniera corretta, non sempre è informativa e pertanto non sempre descrive i dati in maniera opportuna?

In aula propongo questo semplice esercizio: la classe è stata invitata a un aperitivo. Il gruppo che invita la classe e che offre l'aperitivo ha un'età media pari a 20 anni. Ovvio accettare! Ma che informazioni abbiamo sapendo solo il valore dell'età media? In prima battuta potremmo pensare che il gruppo è composto in gran parte da ventenni, ma in classe c'è spesso qualcuno incerto che afferma che potrebbe esserci qualche adulto, qualche anziano o qualche bambino.

La discussione di solito fa nascere altri dubbi fino ad arrivare a dire che nel gruppo che ci invita potrebbe addirittura non esserci alcun giovane, ma solo neonati e nonni! La cosa non è grave per l'esito dell'aperitivo, ma ci si chiede a questo punto quali siano le informazioni avute dal sapere che l'età media è pari a 20 anni. La media in questo caso non basta a descrivere i dati e potrebbe addirittura portarci a delle considerazioni sbagliate[14].

Dobbiamo pertanto utilizzare anche altri strumenti[15]: in questo modo potremo rispondere a Trilussa, al suo dubbio sulla statistica "curiosa".

---

[12] Così recita la poesia: «Sai ched'è la statistica? È na' cosa / che serve pe fà un conto in generale / de la gente che nasce, / che sta male, / che more, che va in carcere e che spósa. / Ma pè me la statistica curiosa / è dove c'entra la percentuale, / pè via che, lì, la media è sempre eguale / puro co' la persona bisognosa. / 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 / perch'è c'è un antro che ne magna due. [...]» (cfr. Trilussa 1954, p. 189).

[13] Sempre che sia corretto calcolare la media, cosa ad esempio non possibile se stiamo lavorando con numeri ordinali – in una gara di sci se un atleta arriva primo nella prima *manche* e terzo nella seconda *manche*, non risulterà secondo, ma il risultato dipenderà dai tempi che ha fatto.

[14] Pensiamo ad esempio al significato che potremmo dare al tempo medio d'attesa per un esame diagnostico: dire che il tempo medio di attesa è pari a 20 giorni potrebbe dare informazioni molto diverse e quindi non farci capire quale è in realtà la situazione.

[15] Tempo fa mi colpì un episodio raccontato da un collega medico: per la valutazione di alcuni laboratori di analisi veniva calcolata la media di singole prestazioni e, a seguire, si calcolava la media di quanto ottenuto nelle diverse prestazioni (si faceva quindi una media di medie). L'obiettivo era quello di arrivare a una classifica che permettesse di stabilire un ordine di valore dei laboratori. Come si può intuire, questo procedimento non funziona e non valuta in maniera corretta singoli laboratori: ci possono essere ad esempio prestazioni da considerare con *pesi* diversi.

Supponiamo in un altro esempio di valutare i tempi di spedizione di una ditta di trasporti. Un responsabile dichiara che il *tempo mediano* di ricezione di un pacco con determinate caratteristiche è pari a 5 giorni. Il responsabile sta dicendo che metà delle spedizioni in oggetto ha raggiunto la destinazione entro 5 giorni, l'altra metà dopo 5 giorni, e giustifica l'uso della *mediana*[16] affermando che ci sono dei pacchi che sono stati recapitati dopo molto tempo.

La *media*, pur corretta, sarebbe pertanto stata sporcata da questi valori estremi e non avrebbe descritto in maniera informativa la situazione. La mediana, dopo aver messo in ordine i dati, prende un valore che lasci alla sua sinistra il 50% dei dati stessi e non è condizionata da eventuali valori estremi: per la mediana conta l'ordine dei dati e non i valori degli stessi. Un tempo medio di spedizione più alto rispetto al tempo mediano fa pensare alla presenza di pacchi recapitati dopo molti giorni. Da notare che la mediana, per quanto detto sopra, può essere usata anche quando i dati sono solamente ordinabili, cosa non possibile per la media: potremo parlare di cognome mediano in una lista di persone, di figlio mediano, di giocatore mediano, e così via.

2.3 La deviazione standard: quando già il nome incute timore

Anni fa, a un convegno, ero stato avvicinato da un collega filosofo. Cominciammo a parlare delle nostre discipline e lui mi disse tutta la sua preoccupazione nel capire poco o niente di Matematica e di Statistica.

È buffo osservare come il mondo si divida in chi ama la Matematica e in chi non la sopporta e dice o si vanta di non capirci niente! Abbiamo una grande responsabilità come docenti nel cercare di trasmettere, per quanto possibile, la bellezza e la passione per questa disciplina. Il discorso è un po' diverso per la Statistica: si dà spesso per scontato di conoscerla e quindi si pensa di poterla usare, arrivando però frequentemente a risultati non corretti.

[16] Data una successione finita di valori disposti in ordine crescente (o decrescente), la *mediana* è quel valore che occupa il posto centrale, se il numero dei termini è dispari, o qualunque valore compreso nell'intervallo dei due termini centrali, se il numero dei termini è pari.

Tornando al collega filosofo, si diceva angosciato dalla cosiddetta *deviazione standard*. Mi confidò che non l'aveva mai capita del tutto, pur essendosi trovato diverse volte a doverla usare. Mi resi subito disponibile a fornirgli una semplice spiegazione, ma replicò che preferiva di no, in quanto era sicuro che comunque non avrebbe capito e che mi avrebbe fatto solo perdere del tempo.

Feci finta di accettare ma, al tempo steso, gli dissi semplicemente che questo strumento statistico altro non è che un indice di dispersione fatto rispetto alla media. Sorpreso, replicò che forse cominciava a seguirmi, avendo capito l'*idea* e non la *formula*. La sola formula infatti era per lui illeggibile ma ora riusciva a intravedere in quella simbologia matematica una sorta di traduzione. Fu un momento simpatico e divertente che spero lo abbia aiutato a capire di più.

In breve la introdurrei così, prima di arrivare a una formulazione matematica: possiamo avere due gruppi di età media 20, con valori più o meno dispersi: ecco allora che insieme alla *media*, che è un *indice di centralità*, occorre descrivere i dati attraverso un *indice di dispersione*, ad esempio la *deviazione standard*, che va a calcolare quanto i dati sono dispersi rispetto alla media[17].

La deviazione standard presenta però criticità simili a quelle della media: non sarà ad esempio informativa nelle situazioni in cui sono presenti valori estremi (valori che vengono chiamati *outlier*), valori che modificano la media e, di conseguenza, anche la deviazione standard. Queste criticità spiegano come mai in alcuni articoli non è presente questo indice di dispersione ma si parla di distanza interquartile. La *distanza interquartile* è la differenza tra il terzo e il primo quartile, dove il primo quartile è un valore che lascia alla sua sinistra il 25% dei dati, il terzo quartile ne lascia invece il 75% (la *mediana* è quindi il secondo quartile).

Supponiamo per esempio che un gruppo di persone abbia età minima 18 anni, età

[17] Se abbiamo n dati x_1 x_2 x_n con media μ, si calcolano e si sommano le distanze al quadrato dei valori x_i rispetto alla media (si eleva al quadrato in quanto alcuni contributi sono negativi e altri positivi) e si divide per la numerosità n. La deviazione standard è definita come la radice quadrata di quanto appena calcolato: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$.

massima 45, mediana 28, primo quartile 21, terzo quartile 34. In questo caso il 25% dei dati è compreso tra 18 e 21, un altro 25% tra 21 e 28, un altro 25% tra 28 e 34, l'ultimo quarto tra 34 e 45. Questi cinque elementi: massimo e minimo, primo e terzo quartile e la mediana, descrivono in maniera molto informativa la distribuzione delle età del gruppo considerato.

Se in un gruppo di persone dai 20 ai 25 anni facciamo uscire il giovane con età più alta e, al suo posto, facciamo entrare un adulto sessantenne, la media, cambierà e così pure sarà per la deviazione standard, cosa che invece non avverrà per la mediana e per i quartili.

Un altro strumento statistico, credo abbastanza conosciuto, che estende i concetti appena presentati di mediana e di quartili è quello di *percentile*[18]: la mediana possiamo pensarla come secondo quartile ma anche come 50° percentile; il primo quartile come 25° percentile, e così via. Dire pertanto che il 97° percentile del tempo d'attesa per un esame diagnostico è pari a 20 giorni, significa che nel 97% dei casi l'esame è stato effettuato entro 20 giorni.

Un bambino che nasce come peso nel 6° percentile, è un bambino piuttosto piccolo in quanto solo il 6% dei bambini nasce con un peso inferiore. Per seguire la crescita fisiologica i pediatri ci insegnano che è importante considerare le curve dei percentili: un bimbo che nasce come peso al 60° percentile e che dopo qualche settimana si ritrova al 25° percentile indica una qualche sofferenza, ad esempio una alimentazione non corretta.

2.4 Un grafico brutto ma utile: il boxplot o grafico a scatola e baffi

Il *boxplot*, o grafico a scatola e baffi, non è ancora molto conosciuto, ma risulta essere molto utile in quanto fornisce una descrizione molto valida della distribuzione dei dati. Questo grafico nasce da un'idea del matematico e statistico John Wilder Tukey[19]. Immagino che Tukey stesse cercando qualche strumento statistico che potesse descrivere

[18] Non è raro sentire ad esempio che un bimbo nasce in un percentile come peso e in un percentile come altezza, dando così alcune prime caratteristiche importanti del neonato.

[19] In realtà, come riportato nel testo di Sgarro, Franzoi, Vicig (cfr. Sgarro, Franzoi, Vicig 2022), una prima versione del boxplot, chiamata *range-bar*, fu proposta nel 1952 da Mary Eleanor Spear.

i dati in maniera più informativa e, come abbiamo visto, la media e la deviazione standard non sempre sono informative, mentre la mediana e i quartili, indici più robusti in quanto più stabili rispetto a misure estreme, raccontano la distribuzione dei dati in maniera più dettagliata.

L'idea di Tukey è molto semplice: utilizza i valori del minimo, del massimo, del primo e terzo quartile q_1 e q_3, della mediana e, come si può vedere dal grafico, disegna una scatola (*box*) tra q_1 e q_3 e aggiunge dei baffi (*whiskers*) alle due estremità. La mediana viene quindi riportata nella scatola. Il boxplot di Figura 1 ci dice allora che un quarto dei valori si trova tra 20 e 21, un quarto tra 21 e 24, un altro quarto tra 24 e 34, e infine un ultimo quarto tra 34 e 38. Il fatto che la mediana non si trovi esattamente a metà della scatola, così come le due lunghezze diverse dei baffi, dicono la non simmetria della distribuzione.

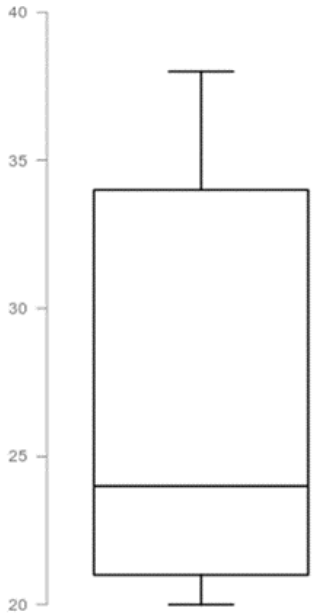

Figura 1. Boxplot.

Il secondo boxplot (cfr. Figura 2) non ha il baffo inferiore in quanto, in questo caso, il minimo coincide con il primo quartile: questo significa che il valore 20 è presente almeno il 25% delle volte (e meno del 50% in quanto la mediana vale 24).

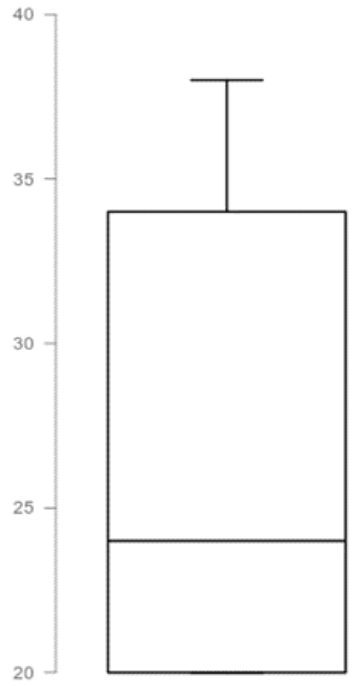

Figura 2. Boxplot con min=q_1.

A Tukey rimaneva però un problema da risolvere: la possibilità di individuare eventuali valori estremi, i valori *outlier*. Supponiamo che nei dati del primo boxplot il valore massimo sia 90 anziché 38, e supponiamo ancora che tutti i valori, a parte il 90, siano inferiori a 40: quindi min=20, q_1=21, Mediana=24, q_3=34, max=90. Il baffo superiore andrebbe in tal caso da 34 a 90 e ci porterebbe a dire che il 25% dei dati sta tra questi due estremi, quando in realtà i dati di quel baffo sono compresi tra 34 e 40, a parte il valore 90, che risulta pertanto un valore isolato, un *outlier*[20].

Come descrivere meglio questa situazione, come evidenziare i valori isolati, valori che conviene considerare separatamente? Venne così proposta da Tukey questa soluzione, inventando una nuova regola: il baffo può essere lungo al massimo una volta e mezza la distanza interquartile. In formule, $lungh(baffo) \leq 1{,}5(q_3-q_1)$.

In questo modo il boxplot precedente risulta avere il baffo superiore che parte da 34 e arriva a 40, mentre il valore 90 viene evidenziato in maniera differente (cfr. Figura 3).

[20] Gli *outlier* potrebbero essere errori di battitura, valori patologici o altro, valori che è importante poter evidenziare direttamente in un grafico.

I baffi – che potrebbero essere lunghi al massimo 19,5 – sono tali che quello inferiore, che va da 20 a 21, ha lunghezza pari a 1, quello superiore, che va da 34 a 40 (ricordo che non ci sono persone con valori sopra il 40) ha lunghezza 6.

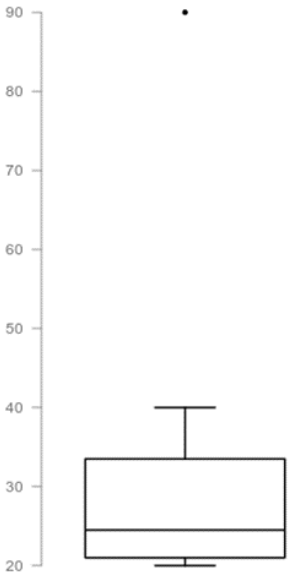

Figura 3. Boxplot con outlier.

Una soluzione semplice, che permette una buona descrizione dei dati e della loro distribuzione, come pure di eventuali valori estremi.

2.5 Proviamo ad andare verso una prima conclusione

Mi piace concludere questo primo testo di supporto all'insegnamento della Statistica, facendomi aiutare da alcune considerazioni che avevo trovato tempo fa su un sito[21] dell'Università di Padova. La Statistica, così recitava il testo, riguarda una delle grandi sfide che la Filosofia pone alla Scienza e cioè come tradurre l'informazione in conoscenza. La Statistica suggerisce come valutare ciò che osserviamo e come prendere di conseguenza delle decisioni. Come già osservato, la Statistica tratta una componente essenziale del mondo reale, la *casualità*, l'*aleatorietà*, *l'incertezza,* elementi presenti nella raccolta del dato, nella scelta dei cosiddetti materiali e metodi, e così via. Il testo

[21] Cfr. Università di Padova – Dipartimento di Scienze Statistiche in Siti web.

concludeva osservando che la capacità di far fronte all'incertezza è pertanto un'importante caratteristica della Statistica stessa.

Sono affermazioni molto importanti, apparentemente lontane dal modo in cui siamo soliti pensare alla Statistica, idee che offrono dei punti di vista nuovi e interessanti anche nelle nostre attività di insegnamento. In un tempo in cui i dati sono sempre più presenti nelle nostre attività e nelle nostre scelte, diventa sempre più rilevante la capacità di saperli capire e di attribuire loro un significato corretto.

BIBLIOGRAFIA

HUFF DARRELL
2009, *Mentire con le statistiche / How to lie with Statistics*, M&A editore.

PENNAC D.
2008, *Diario di scuola*, Milano, Feltrinelli.

SGARRO A., FRANZOI L., VICIG P.
2022, *Statistica di base. Idee e tecniche*, Bologna, Zanichelli.

TORELLI L.
2024, *Molte statistiche ma poca Statistica,* in V. GIANTIN, G. GUANDALINI (a cura di), «Cura e reciprocità – molti saperi per un contributo dialogico sulla responsabilità come nuovo paradigma di cura», Volume II, Bologna, Il Mulino.

TRILUSSA[22]
1954, *Tutte le poesie*, a cura di Pietro Pancrazi, note di Luigi Huetter, Milano, Mondadori.

SITI WEB

UNIVERSITÀ DEGLI STUDI DI PADOVA – DIPARTIMENTO DI SCIENZE STATISTICHE
Statistica e Statistici. Cos'è la statistica?,
<https://www.stat.unipd.it/futuri-studenti/statistica-e-statistici>, sito consultato il 19.10.2024.

WOOCLAP
<https://www.wooclap.com/it/>, sito consultato il 19.10.2024.

FILMOGRAFIA

OPPENHEIMER
2023, film diretto e co-prodotto da Christopher Nolan.

[22] Pseudonimo anagrammatico di Carlo Alberto Camillo Salustri (1871-1950).