18 Introduzione alla Statistica Descrittiva

18.1 Introduzione

La Statistica Descrittiva opera su dati.
Per avere i *dati*, è necessaria la *raccolta dati*, operazione preliminare – in generale complessa e fonte di innumerevoli errori e distorsioni – che tecnicamente non fa parte della Statistica Descrittiva.
Accenneremo alla raccolta dati, e cosa molto bella è fare esperienze in aula.

La Statistica Descrittiva si occupa essenzialmente di:

- **riassumere dati, tipicamente molti,**
  (se sono pochi può essere meglio elencarli)
- **presentare il riassunto,**
  il più comprensibilmente possibile: medie, diagrammi, e altro.

È cosa ben diversa, più semplice e storicamente antica, dalla Statistica Inferenziale che verrà trattata fra le Matematiche dell'Incertezza.

In una trattazione elementare della Statistica Descrittiva non si distingue fra *popolazione* e *campione*: abbiamo i dati che abbiamo, e quelli consideriamo popolazione, anche se in effetti sono un campione di una popolazione più ampia. La distinzione diventa invece essenziale nella Statistica Inferenziale.

In questo testo elementare, verrà tralasciata la questione della rilevazione materiale dei dati – con tutta la sua problematicità.

E in questa lezione **verrà trattata solo la sintesi** dei dati, ovvero la rappresentazione dei dati in una forma umanamente comprensibile e *trattabile*, cosa particolarmente utile se i dati sono più di una dozzina.

Si vuole riassumere i dati con 1 diagramma oppure 1 o pochi valori, per

comprenderli

confrontarli ← cosa fondamentale per i farmaci

divulgarli

e, a livello eticamente discutibile, presentarli manipolativamente (cosa che viene fatta amplissimamente: un fenomeno in *diminuzione* si trova sempre come presentarlo, almeno a livello di titolo riassuntivo, come *aumento*, se si vuole, usando artifici statistici formalmente legittimi).

18.2 Riassumere i dati: quanti gatti e quanti cani avete?

Dati rilevati a Farmacia a Trieste il 17/11/2023, per gatti e cani:

```
G: 0 0 0 1 0 0 0 0 0 1 0 1 1 0 1 0 0 0 1 0 0 0 0 2 3 2 0 1 0
C: 1 1 0 1 0 2 1 0 0 0 0 0 0 2 4 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
```

Gli studenti di Farmacia partecipanti alla rilevazione

hanno da 0 a 3 gatti;

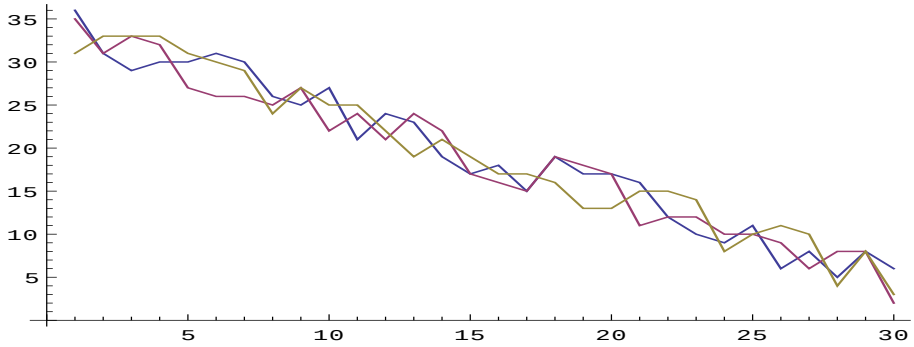
hanno da 0 a 4 cani.

Si tratta delle statistiche del minimo e del massimo, di calcolo abbastanza semplice, eppure già molto espressive della situazione, di non così immediata comprensione semplicemente guardando il dataset, specialmente se avesse 100 numeri, come in questo esempio ipotetico:

2, 0, 3, 0, 3, 2, 1, 1, 2, 0, 2, 2, 2, 1, 1, 2, 1, 0, 2, 0, 3, 4, 0, 1, 4, 1, 2, 3, 0, 1, 0, 2, 1, 3, 3, 2, 0, 1, 0, 2, 0, 2, 3, 2, 1, 1, 2, 3, 2, 2, 4, 2, 1, 3, 3, 3, 2, 0, 2, 3, 0, 1, 2, 2, 2, 1, 4, 0, 2, 0, 0, 1, 2, 2, 1, 1, 1, 3, 0, 4, 0, 1, 2, 2, 1, 1, 2, 4, 1, 2, 3, 4, 2, 0, 0, 3, 2, 2, 3, 2

E con più calcoli potremmo trovare il numero medio di gatti, e di cani. E vedremo molte altre forme di Statistica Descrittiva.

Esempio. Consideriamo un fenomeno in evoluzione da un giorno 1 a un giorno 30, tanto più negativo per la società quanto più sono grandi 3 parametri a, b, c.



Un *ingenuo* direbbe che la situazione sta migliorando: i parametri negativi diminuiscono. Invece i media del posto scrivono, a beneficio del *popolo bue*:

Giorno 2 Aumenta il tal parametro!
Giorno 3 Aumenta il tal parametro!
Giorno 4 Aumenta il tal parametro!
Giorno 5 Il Governo valuta la preoccupante situazione
Giorno 6 Aumenta il tal parametro!
Giorno 7 Il Governo valuta la preoccupante situazione
Giorno 8 Il Governo valuta la preoccupante situazione
Giorno 9 Aumenta il tal parametro!
Giorno 10 Aumenta il tal parametro!
Giorno 11 Aumenta il tal parametro!
Giorno 12 Aumenta il tal parametro!
Giorno 13 Aumenta il tal parametro!
Giorno 14 Aumenta il tal parametro!
Giorno 15 Il Governo valuta la preoccupante situazione
Giorno 16 Aumenta il tal parametro!
Giorno 17 Il Governo valuta la preoccupante situazione
Giorno 18 Aumenta il tal parametro!
Giorno 19 Il Governo valuta la preoccupante situazione
Giorno 20 Il Governo valuta la preoccupante situazione
Giorno 21 Aumenta il tal parametro!
Giorno 22 Aumenta il tal parametro!
Giorno 23 Il Governo valuta la preoccupante situazione
Giorno 24 Il Governo valuta la preoccupante situazione
Giorno 25 Aumenta il tal parametro!
Giorno 26 Aumenta il tal parametro!
Giorno 27 Aumenta il tal parametro!
Giorno 28 Aumenta il tal parametro!
Giorno 29 Aumenta il tal parametro!
Giorno 30 Il Governo valuta la preoccupante situazione

18.3 Esempio di possibile raccolta dati

Fare in aula una raccolta dati è interessante ma bisogna fare estrema attenzione al rispetto per la persona (non si possono chiedere il peso e l'altezza, nemmeno in questionari anonimi) e alla possibilità concreta che gli studenti conoscano il dato richiesto (non si può chiedere quanta vitamina D o rame hanno nel sangue perchè in generale non lo sanno). La raccolta dati e successive analisi statistiche sono estremamente valevoli didatticamente e piacciono molto agli Studenti. Alcune possono essere fatte a voce, altre anonimamente.

Ecco una possibile raccolta dati, fattibile addirittura pubblicamente a voce in aula: Quanti animali domestici possiedi?

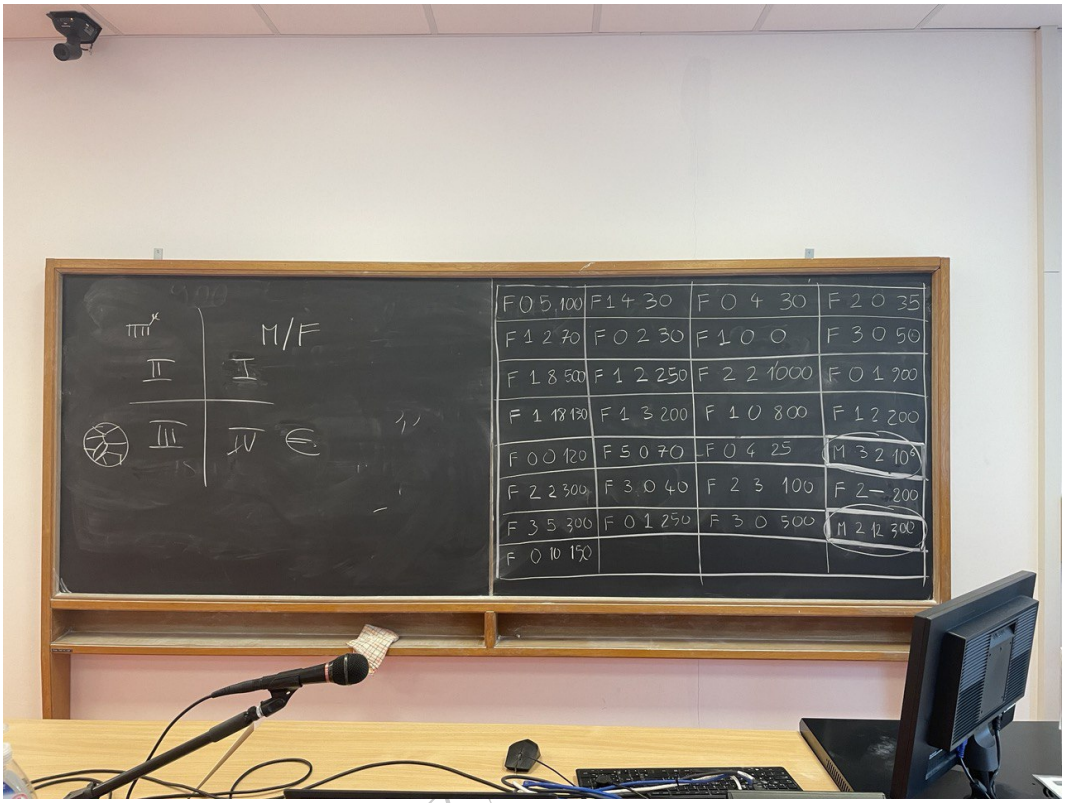


Figure 25: Raccolta dati, Matematica per Farmacia a Trieste, in un anno e con 4 domande che non specifichiamo. Si era consegnato un foglio bianco ad ogni presente in aula e le risposte andavano scritte nei 4 quadranti ideali del foglio. Raccolta dati anonima, risposte lette via via ad alta voce da una volontaria, e trascritte alla lavagna con successiva distruzione dei fogli. Il valore anomalo (*outlier*) un milione è trascritto come 10^6. Un trattino orizzontale indica un dato mancante: un partecipante ha scritto che non sa invece di rispondere con un numero, e forse intendeva 0 o proprio che non sa rispondere, in ogni caso è un dato mancante classificabile come "non sa/non risponde" oppure come dato mancante per qualsiasi causa.

18.4 Il dataset e le operazioni su esso

L'insieme dei dati $x_1, ..., x_n$ (che spesso sono numeri ma non sempre) si chiama *dataset* e non è un insieme in senso matematico perché le ripetizioni dei valori sono ammesse e non vanno elise.

Si considerano dataset, via via più trattabili matematicamente:

• *nominali*: per esempio con valori in { pari, dispari }
pari, pari, dispari, pari, dispari, dispari (è un dataset)
oppure con valori in { integro, spezzato, bifido }
integro, spezzato, spezzato, bifido, integro, integro (è un dataset)

• *ordinali*: per esempio, per valutare un farmaco antidepressivo potremmo considerare questi valori non numerici
morte
grande peggioramento
medio peggioramento
lieve peggioramento
stabile
lieve miglioramento
medio miglioramento
grande miglioramento
e dopo il *trial clinico* avere un dataset di – per esempio – 172 valori di quel tipo.

Oppure, altro caso:
sono completamente d'accordo,
sono d'accordo,
sono parzialmente d'accordo,
non sono d'accordo,
non sono assolutamente d'accordo.

• *numerici*: 3, 1, 2, 4, 6, 1000, 6. Quelli che più ci interessano.

Anche nei dataset ordinali si potrà fissare una corrispondenza con numeri, ma non avrà alcun senso fare la media fra 3 soggetti con: lieve miglioramento, medio miglioramento, morte. Se invece la "perdita infinita" viene esclusa, una media in qualche modo si potrà anche fare, con opportuna trasformazione in valori numerici. (E alla fin fine – di necessità in virtù – qualcosa riusciranno a fare anche con la perdita infinita).

Fra le variabili nominali si distinguono quelle *dicotomiche* ovvero *binarie*, con 2 soli valori, come pari/dispari, vivo/morto, successo/fallimento, 0/1: Link->

Considereremo 2 capitoli della Statistica Descrittiva:

◇ le *rappresentazioni grafiche*: Lezioni 19, 20;

◇ le *statistiche di sintesi*, funzioni $f(x_1, ..., x_n)$ con un numero n a priori indeterminato di argomenti. Inizieremo da queste.

Ecco alcune statistiche di sintesi per un dataset:

• il minimo: $\min(x_1, ..., x_n)$ ← per un dataset almeno ordinale;

• il massimo: $\max(x_1, ..., x_n)$ ← per un dataset almeno ordinale;

• la somma: $x_1 + ... + x_n$ ← per un dataset numerico.

Il significato delle soprastanti statistiche di sintesi è ovvio. Si calcolino quei 3 valori per questo dataset: $\frac{1}{\pi}$, $\frac{2}{7}$, 0.3, $\frac{1}{\sqrt{10}}$, e, e^{-1}, $\frac{1}{3}$.

Per esempio (ipotetico, di farmacologia veterinaria) è immensamente più significativo dire che fra 100 000 ostriche sono state trovate perle con un minimo di 0 e un massimo di 4 in ogni esemplare per un totale di 41 320 perle, piuttosto che elencare il numero di perle 0, 0, 1, 0, 2, 0, 0, 0, 0, 1... di ciascuno degli esemplari. Abbiamo riassunto 100 000 numeri con 3, minimo massimo e somma, e capiamo perfino meglio la situazione che con pagine di numeri! E soprattutto possiamo meglio *confrontarla* con la situazione di un'altra vasca di 100 000 ostriche, magari nutrite/trattate diversamente, che analogamente riassumeremo con quei 3 indici.

Più in dettaglio considereremo queste altre statistiche di sintesi:

• gli indici di posizione:

  ◦ quelli che vorrebbero riassumere in 1 solo valore "medio" il complesso dei dati, e saranno l'argomento di questa Lezione 18.10;

  ◦ quelli che vorrebbero riassumere in 5 valori o corrispondentemente 1 diagramma il complesso dei dati, Lezione 20;

• gli indici di dispersione ovvero variabilità, Lezione 21, che vorrebbero quantificare con 1 numero la non omogeneità di dati numerici.

◇ Esiste anche 1 indice numerico, la *skewness*, che con una complicata formula misura quantitativamente l'asimmetria dei dati, che però noi tratteremo solo qualitativamente.

Ed esistono altri indici che non tratteremo (come la *curtosi*).

18.5 Medie

Media (aritmetica). Da ora consideriamo un dataset $\{x_1, ..., x_n\}$.

$$M(x_1, ..., x_n) := \frac{x_1 + x_2 + ... + x_n}{n}$$

Esempio.
Sul covid-19 in Italia, leggiamo in
https://www.epicentro.iss.it/coronavirus/bollettino/Report-COVID-2019_4_ottobre.pdf

> Dati al 7 settembre 2020 (...) da 4190 deceduti per i quali è stato possibile analizzare le cartelle cliniche. Il numero medio di patologie osservate in questa popolazione è di 3,4

Facile, ottimo anche per i voti, ma il reddito medio di questi

$$3, 1, 2, 4, 6, 1000, 6$$

è 146, non poi così espressivo della situazione globale, a causa dell'*outlier* 1000, valore anomalo ovvero aberrante.

Tratteremo in seguito la questione dei valori anomali ma anticipiamo che talvolta vengono semplicemente *eliminati*.

Media geometrica. Per n numeri positivi (invece la media aritmetica non lo richiede), è la radice n-esima del loro prodotto:

$$GM(x_1, ..., x_n) := (x_1 \cdot ... \cdot x_n)^{\frac{1}{n}} .$$

Nell'esempio numerico soprastante ≈ 7.05; con WolpframAlpha
`geometric mean 3,1,2,4,6,1000,6`

Nota 1. Per quanto possa interessare il lettore, l'opinione dello scrivente è che più della metà delle volte in cui negli articoli scientifici si fa una media aritmetica, la media geometrica sarebbe stata più pertinente; e plausibilmente, in progresso di tempo la media geometrica sostituirà largamente la media aritmetica. Certo, la media

geometrica è meno semplice calcolarla, almeno a mano.

Nota 2. Alcuni Autori ritengono che la media geometrica sia meno sensibile agli outlier della media aritmetica, ma questo non può essere affermato con certezza matematica, dipende da ogni singolo dataset.
Riguardo alle applicazioni non lontane dalle Scienze Biomediche, leggiamo in Wikipedia, l'enciclopedia libera, alla voce "Geometric mean":

> starting from 2010 the United Nations Human Development Index did switch to this mode of calculation, on the grounds that it better reflected the non-substitutable nature of the statistics being compiled and compared:
>
> > The geometric mean decreases the level of substitutability between dimensions [being compared] and at the same time ensures that a 1 percent decline in say life expectancy at birth has the same impact on the HDI as a 1 percent decline in education or income. Thus, as a basis for comparisons of achievements, this method is also more respectful of the intrinsic differences across the dimensions than a simple average.

In pratica, se un Paese porta l'aspettativa di vita da 40 a 60 anni, questo verrà ben rilevato facendo la media geometrica con il reddito pro-capite; se invece si facesse la media aritmetica, l'incremento di reddito da 4 000 a 4 040 dollari peserebbe di più: l'incremento è di 40 invece che 20, ma è solo l'1%, mentre da 40 a 60 è del 50%. La media geometrica appare migliore quando dobbiamo riassumere dati che variano su scale molto diverse. E anche in altri casi.

Media (aritmetica) ponderata (o pesata). Dati dei *pesi* $a_1, ..., a_n$, di somma 1, la media pesata del dataset $\{x_1, ..., x_n\}$ è $a_1x_1 + ... + a_nx_n$.
(Detto semplificatamente) pesi a_k maggiori di $\frac{1}{n}$ sono associati a dati che si vuol far "pesare" di più nella considerazione finale.

Per esempio nella media pesata

$$0.4\,voto\_Matematica + 0.4\,voto\_Chimica + 0.2\,voto\_Marketing$$

il voto nell'esame di Marketing pesa/conta metà di ciascuno degli altri voti.

Per il lettore interessato: esistono varie altre "medie".

Mediana. Il numero centrale dei dati riordinati, adesso 4: $1, 2, 3, \underline{4}, 6, 6, 1000$. E se i dati sono in numero pari, si considera la media dei 2 centrali. La mediana è definita anche per valori *ordinali* almeno se il numero di dati è dispari.

Leggiamo (11 novembre 2022) in https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia

> al 10 gennaio 2022 (...) Complessivamente, le donne decedute dopo aver contratto infezione da SARS-CoV-2 hanno un'età più alta rispetto agli uomini (età mediane: donne 85 anni – uomini 80 anni).

Moda. Il valore più frequente. Nel nostro esempio è 6 ma in generale non è definita perché i numeri sono tutti diversi o perché 2 o più valori sono ripetuti ugualmente. Ecco per esempio un dataset *bimodale*: $6, 6, 6, 6, 7, 7.5, 7.5, 8, 8, 8, 8, 8.5, 8.5, 9, 9.5, 10, 10$.
La moda è definita anche per dati *nominali*, neppure ordinali, per esempio in ogni nazione c'è un cognome più frequente.

Confronto fra media, mediana e moda.
Su Wikipedia, l'enciclopedia libera, alla voce "Mode (statistics)" troviamo questo significativo esempio:
1, 2, 2, 3, 4, 7, 9; media=4, mediana=3, moda=2.

Leggiamo (11 novembre 2022) in https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia

> al 10 gennaio 2022 (...) L'età media dei pazienti deceduti e positivi a SARS-CoV-2 è 80 anni (mediana 82 (...)).

Non ordinamento fisso di media, moda e mediana
Media, mediana e moda sono 3 valori che possono presentarsi in qualunque ordine fra loro: dipende dal dataset.

Media interquartile. Leggiamo su Wikipedia, l'enciclopedia libera, alla voce Central tendency:

> The method is best explained with an example. Consider the following dataset:
> 5, 8, 4, 38, 8, 6, 9, 7, 7, 3, 1, 6
> First sort the list from lowest-to-highest:
> 1, 3, 4, 5, 6, 6, 7, 7, 8, 8, 9, 38
> There are 12 observations (datapoints) in the dataset, thus we have 4 quartiles of 3 numbers. Discard the lowest and the highest 3 values:
> ~~1~~, ~~3~~, ~~4~~, 5, 6, 6, 7, 7, 8, ~~8~~, ~~9~~, ~~38~~
> We now have 6 of the 12 observations remaining; next, we calculate the arithmetic mean of these numbers:
> xIQM = (5 + 6 + 6 + 7 + 7 + 8) / 6 = 6.5
> This is the interquartile mean.
> For comparison, the arithmetic mean of the original dataset is
> (5 + 8 + 4 + 38 + 8 + 6 + 9 + 7 + 7 + 3 + 1 + 6) / 12 = 8.5
> due to the strong influence of the outlier, 38.

(Che 38 vada considerato outlier è opinabile).

Con artifici si definisce per un numero di dati non quadruplo.

Detto grezzamente, consideriamo il reddito medio delle persone medie, al netto di mendicanti e ricconi, che quelli ci sono comunque dappertutto.

ES. 2$_{\mu 2019}$
≈ Il carbonio risulta avere – almeno secondo alcuni Autori: non possiamo qua garantirlo in forma assoluta e fare Chimica o Fisica; si veda Wikipedia in inglese alla voce *Isotopes of carbon* – 12 isotopi non reperibili in natura (oltre a 3 reperibili in natura) con queste emivite approssimative:

$3,5 \times 10^{-21}$ s, 126,5 ms, 19,3 s, 20,364 min, 2,45 s, 0,747 s,
193 ms, 92 ms, 46,2 ms, 16 ms, <30 ns, 6,2 ms.

Dopo aver convertito minuti, millisecondi e nanosecondi in secondi con le note

formule

1 min=60 s, 1 ms=0,001 s, 1 ns=$0,000\,000\,001$ s,

determinare la media interquartile delle emivite.

SVOLGIMENTO

Con la conversione in secondi le emivite sono, nell'ordine iniziale dei dati,

$3,5 \times 10^{-21}$ s, 0,126 5 s, 19,3 s, 1 221,84 s, 2,45 s, 0,747 s,

0,193 s, 0,092 s, $0,046\,2$ s, 0,016 s, $< 0,000\,000\,030$ s, $0,006\,2$ s.

Ovvero, in ordine crescente, omettendo l'unità di misura,

$3,5 \times 10^{-21}$ → questo e il seguente potrebbero doversi scambiare

$< 0,000\,000\,030$ → vedi nota alla linea precedente

$0,006\,2$

$0,016$

$0,046\,2$

$0,092$

$0,126\,5$

$0,193$

$0,747$

$2,45$

$19,3$

$1\,221,84$.

I 12 valori ordinati sono 4 terne di valori, ed eliminate la prima terna (coi valori più bassi) e l'ultima (coi valori più alti), i 6 valori centrali sono

$0,016$

$0,046\,2$

$0,092$

$0,126\,5$

$0,193$

$0,747$

e la loro media è il valore cercato, la media interquartile, in secondi:

$$\text{IQM} = \frac{0,016 + 0,046\,2 + 0,092 + 0,126\,5 + 0,193 + 0,747}{6} = \frac{1,220\,7}{6} =$$

$$\boxed{\approx 0,203 \text{ s}}$$

(IQM (acronimo di InterQuartile Mean) è un simbolo classicamente usato per la media interquartile; anche iqm e x_{IQM}).

18.6 Inesistenza di una media migliore di tutte

Non esiste una media che sia la migliore di tutte in ogni caso.

Eppure avremo sempre bisogno di una qualche media, mica possiamo fermarci a dare liste di migliaia o milioni o miliardi di numeri: è necessario quel passo avanti che riassume i dati in 1 numero medio.

Se veramente siamo interessati a produrre
(1) un valore in qualche modo "medio" dei dati
(2) che prescinda dai casi/valori *eccezionali*
allora
- mediana e
- media interquartile

– sono decisamente valide
+ ma solo la mediana è *sempre* facilmente calcolabile
∗ invece la media aritmetica richiede di escludere gli outlier:
  ◇ se i dati sono pochi si può fare a mano ma lasciandoci dubbi
  ◇ se i dati sono moltissimi non si può proprio fare a mano
  però
  > si può fare informaticamente con varie regole che definiscono gli outlier, ma in diversi modi a seconda degli Autori.

18.7 Disilludersi sulle medie anche senza outlier

Non ci si illuda: non esiste alcun tipo di statistica di sintesi che riassuma bene con 1 solo numero ogni dataset, neppure in assenza di outlier. Spesso la mediana rappresenta bene il soggetto "tipico", ma non sempre: in un cortile con 6 cani e 6 galline, la media e la media interquartile e la mediana del numero di zampe per animale è 3. Che non rappresenta alcun caso tipico. (E la media geometrica è ≈ 2.83, di male in peggio). Nemmeno l'*unimodalità* rappresenta una garanzia per la mediana, pur in generale così valida: con 1 grillo, 5 cani e 6 galline

$$6,\ 4,\ 4,\ 4,\ 4,\ 4,\ 2,\ 2,\ 2,\ 2,\ 2,\ 2 \quad (moda = 2)$$

la mediana è ancora 3.

Così, il numero medio di tumori alla prostata e all'utero, di I, II e III grado, in un determinato gruppo di soggetti non ha nessun valore statistico: bisognerà fare statistiche separate per genere.

Nota 3. WolframAlpha calcola online
la media con `mean` seguito dalla lista di numeri: link->
la mediana con `median` seguito dalla lista di numeri: link->
la media geometrica con `geometricmean` seguito dalla lista di numeri: link->

18.8 Stratificazione per età in Medicina e Farmacia

Nelle statistiche di Medicina e Farmacia, i dati relativi all'essere umano talvolta vengono variamente stratificati per età e tipicamente sono presentati in:
0-4
5-9
10-14
...
oppure a decenni
10-19
20-29
...
o anche, a ventenni,
20-39
40-59
60-79
...
Si considerano gli anni compiuti. e allora chi avesse 19 anni e 11 mesi sta nella classe 10-19.

18.9 Sul trasformare dati ordinali in dati numerici

La trasformazione di dati ordinali in numerici introduce elementi di arbitrarietà. Vediamo un esempio.

Supponiamo che un'organizzazione raccolga dati sulla soddisfazione degli utenti sulle prestazioni di un suo organo: per esempio un'università, sull'insegnamento di un suo docente – caso reale in Italia. L'utente può scegliere, per valutare il suo grado di soddisfazione, fra

no
più no che sì
più sì che no
sì

Un'università trasforma i valori ordinali in numerici con lo schema

2
5
7
10

e un'altra con quest'altro, altrettanto legittimo:

1
2
3
4

Consideriamo i giudizi sui famosi Professori Pinco e Pallino, dati dai loro 2 studenti per ciascuno:

Pinco: più no che sì, sì: 5+10=15 oppure 2+4=6

Pallino: più sì che no, più sì che no: 7+7=14 oppure 3+3=6.

Col primo schema Pinco figura meglio di Pallino, 15 a 14, e col secondo sono pari, 6 a 6.

18.10 Le medie creano illusioni percettive

In https://www.epicentro.iss.it/coronavirus/bollettino/Report-COVID-2019_4_ottobre.pdf leggiamo

> Dati al 7 settembre 2020 (...) L'età media dei pazienti deceduti e positivi a SARSCoV-2 è 80 anni

e coi dati al 5 ottobre 2021 è ancora 80 anni. Eppure dall'aprile 2021 al settembre 2021 l'età media dei morti è molto più bassa, come si vede in Figura 3 di https://www.epicentro.iss.it/coronavirus/bollettino/Report-COVID-2019_5_ottobre_2021.pdf. Addirittura, dal grafico vediamo che nel giugno 2020 l'età media era circa 84 anni e nel giugno 2021 (poco meno di) 76. Niente di strano: i morti del periodo estivo sono pochi nel 2020 e molti di più ma comunque pochi nel 2021, la media la fanno principalmente i morti di marzo-aprile 2020 e poi novembre-maggio.

ESERCIZI SULLA LEZIONE

ESERCIZIO$_{\mu 2018}$
$*$ Considerato il seguente dataset

19.68 19.20 19.63 18.94 18.81 18.10 18.63 18.85 0.01 19.51 19.54

che possiamo supporre misurazioni di parametri corporei, si determini la mediana dopo avere eliminato un outlier.

SVOLGIMENTO
Chiaramente 0.01 è l'outlier preannunciato. (Potrebbe ragionevolmente provenire da un momentaneo malfunzionamento di una macchina che ha prodotto i dati).
I 10 dati rimanenti riordinati in modo crescente sono

18.10 18.63 18.81 18.85 18.94 19.20 19.51 19.54 19.63 19.68.

I 2 centrali ovvero mediani sono il 5° e il 6°, eliminando 4 da ogni parte:

18.94 19.20

la cui media aritmetica è il risultato cercato:

$$\boxed{19.07}$$

(Se non si eliminasse l'outlier – talvolta non è poi così evidente che qualche valore vada scartato – la mediana sarebbe alquanto simile, 18.94, mentre nei 2 casi le medie aritmetiche sono molto più diverse fra loro, rispettivamente 19.089 e 17.3536; la mediana ha il pregio di risentire poco degli outlier, il che in casi complessi con migliaia o milioni di dati, e magari nessuna certezza su quanti e quali sarebbero da considerare outlier, è molto significativo).

ESERCIZIO$_{\mu 2019}$
$\approx$ Si supponga di avere questi dati di un ospedale in anni successivi, relativi ai consumi di un certo farmaco:

1907 4257
1908 3956
1909 3936
1910 4183
1911 4114
1912 4525
1913 4188
1914 4111

1915 4404
1916 4180
1917 0
1918 0
1919 4361
1920 4035

Dopo aver eliminato gli outlier, determinare la media interquartile dei dati sul consumo del farmaco. (Si approssimi all'intero più vicino).

SVOLGIMENTO

Gli outlier sono i due 0 (verosimilmente dovuti alla guerra).

Ordiniamo il dataset rimanente, che ha 12 elementi:

3936, 3956, 4035, 4111, 4114, 4180, 4183, 4188, 4257, 4361, 4404, 4525

Eliminiamo i primi 3 e gli ultimi 3 valori

$$\Big[3936, 3956, 4035, \Big]4111, 4114, 4180, 4183, 4188, 4257\Big[, 4361, 4404, 4525\Big]$$

e la media dei 6 rimanenti è

$$\approx 4172.17$$

e approssimiamo come richiesto:

$$\boxed{4172}$$

Nota. Mostriamo la straordinaria insensibilità della media interquartile agli outlier rispetto alla media aritmetica. I 12 valori non nulli provenivano da una variabile aleatoria più o meno normale, e poi erano stati aggiunti 2 zeri in corrispondenza ad anni di guerra. Ora, in questo caso è ben evidente che si tratta di outlier per una serie di motivi:

molto diversi dagli altri valori
valori esattamente 0
corrispondenza con anni di guerra.

Ma in altri casi non è così evidente quali valori considerare outlier, per eliminarli "a mano". Allora potremmo fare la media aritmetica di tutti i valori, che però verrebbe fortemente influenzata dagli outlier. Invece la media interquartile dei 12 valori degli ultimi 12 anni, dal 1909 al 1919, senza escludere gli zeri,

1909 3936
1910 4183
1911 4114
1912 4525
1913 4188

1914 4111
1915 4404
1916 4180
1917 0
1918 0
1919 4361
1920 4035

è 4135, alquanto simile a quella di prima. Gli zeri sono finiti nelle porzioni scartate.

Insoma la media interquartile ci mostra di cogliere il *vero valore "medio"* della variabile aleatoria retrostante al fenomeno quando esso avviene *normalmente*, e lo fa in un modo *automatico*, con una formula, che non richiede l'esclusione a mano degli outlier, cosa impossibile se i dati sono milioni.

La media aritmetica 3503 invece è molto minore (essenzialmente a causa dei 2 zeri) della media 4188 dei 12 valori non nulli iniziali (e qua si vede la sensibilità della media aritmetica agli outlier).

Sarebbe bello mostrare che similmente avverrebbe considerando tutti i 14 valori iniziali, ma sfortunatamente la definizione di media interquartile per campioni con un numero non quadruplo di elementi non è facilissima per un calcolo a mano.