20 Quartili e box-plot

Vogliamo riassumere molti numeri in un modo più ricco rispetto alla sola media o mediana: aggiungeremo alla mediana 4 numeri, o, corrispondentemente, con quei 5 numeri faremo 1 diagramma.

20.1 Quartili e riassunto dei 5 numeri

Si abbia il dataset numerico $x_1, .., x_n$ e lo si riordini in modo crescente: $y_1, ..., y_n$ *sono gli stessi numeri ma in ordine crescente*.

Il valore intermedio,

o la media dei 2 intermedi se n è pari,

sappiamo che è la mediana,

che da adesso chiameremo anche *secondo quartile*,

e lo indicheremo con

$q_{0.5}$ oppure

$q_{1/2}$ oppure

$q_{50\%}$.

Diviso così il dataset in 2 parti uguali,

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

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

$q_{0.25}$ oppure

$q_{25\%}$,

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

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

$q_{0.75}$ oppure

$q_{75\%}$.

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

q_0 oppure

$q_{0\%}$

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

q_1 oppure

$q_{100\%}$

Si hanno così 5 quartili. Ma purtroppo la loro definizione non è perfettamente univoca nei vari testi e software[111] e inoltre, di fatto, spesso si trova denominato *quartile* l'insieme di valori *fra* 2 quartili – come sopra definiti – e in questo senso, i quartili sono 4.

I 5 numeri detti, $q_{0\%}$,...,$q_{100\%}$, sono una statistica di sintesi (vettoriale[112]) che si chiama *riassunto dei 5 numeri* (o *five number summary*).

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

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

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

Applicazione: un valore è tanto o poco? Secondo Scimago, la rivista scientifica *Journal of Pharmacy and Pharmacology* ha – in un determinato momento – l'H index, che è un indicatore bibliometrico, di 107: è tanto o poco? Se avessimo la lista di tutte le riviste scientifiche catalogate per Pharmacology da Scimago, e ciascuna col suo H index, avremmo la risposta, salvo che sarebbe scarsamente leggibile per eccesso di numeri. Un riassunto dei 5 numeri ci permetterebbe di contestualizzare bene quel valore 107. Scimago fa una cosa del genere, magari non esattamente con l'H index (usa un suo indice interno, il *SCImago Journal Rank* ovvero *SJR indicator*) ma la sostanza è la stessa. Salvo che chiama primo quartile quello più alto, cioè dal terzo quartile (numero) al quarto quartile (numero). Scopriamo così che la rivista considerata, per la disciplina scientifica Pharmacology è solo nel terzo quartile nel 2018, non poi

[111] Leggiamo in questo link accademico: "Molti software hanno diversi algoritmi per calcolare i quantili."

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

così bene (e invece per la categoria Pharmaceutical Science è nel secondo, molto meglio). Invece la rivista *British Journal of Pharmacology* risulta nel primo quartile, il migliore, dal 1999 al 2018.

20.2 Box-plot ovvero diagramma a scatola e baffi

Il *box-plot* ovvero *diagramma a scatola e baffi* è un diagramma molto usato in ambito Biomedico per rappresentare tutti i quartili di un dataset, ovvero il riassunto dei 5 numeri, e magari anche gli outlier.

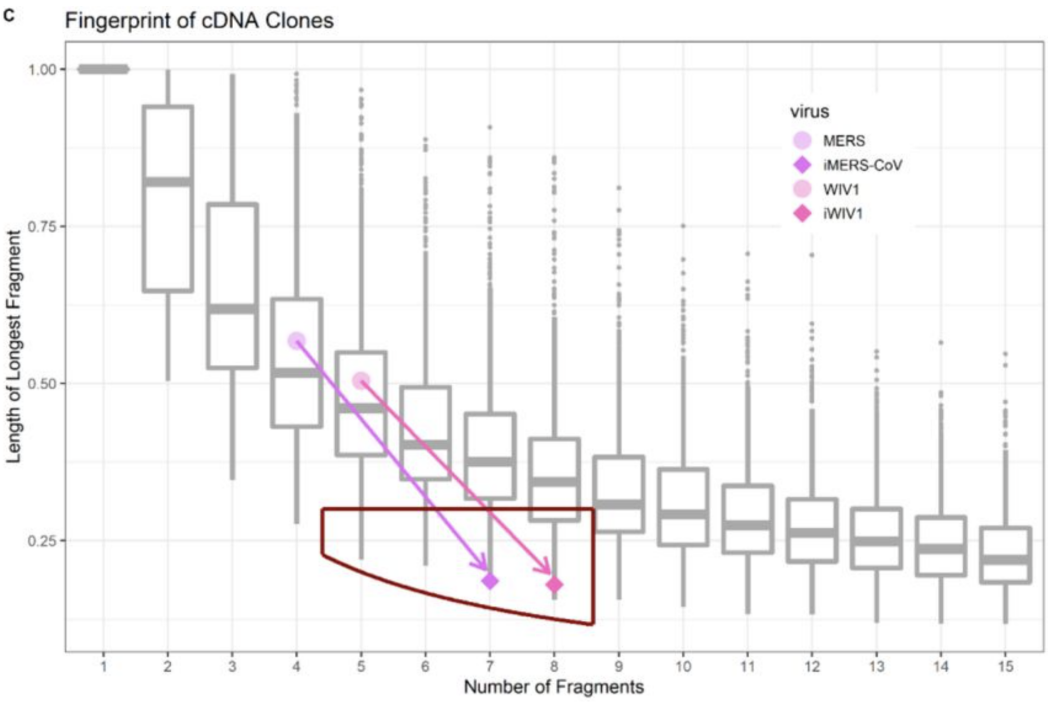


Figure 27: Figura 2 c in "Endonuclease fingerprint indicates a synthetic origin of SARS-CoV-2", by Bruttel, Valentin and Washburne, Alex and VanDongen, Antonius: in bioRxiv (2022), DOI 10.1101/2022.10.18.512756

Vediamo sulla rete alcuni esempi di pertinenza delle Scienze Mediche e Farmaceutiche seguendo questo `link->`, con immagini tratte da PubMed.

Precisiamo subito che di questa rappresentazione grafica
• esistono varianti, sulla lunghezza dei "baffi"[113], ma non solo[114]
• considereremo (quasi) solo la versione più semplice
• chi pubblica un box plot dovrebbe specificare[115] i dettagli su come intenderlo.

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

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

Esercizio$_\mu$ Si disegni il box-plot dei primi 10 numeri primi, e poi dei primi 11.

Talvolta gli outlier vengono rappresentati isolati. Si faccia così per il dataset 1, $\frac{1}{2}$, $\frac{1}{3}$,..., $\frac{1}{11}$, $\sqrt{7}$; e poi senza $\frac{1}{11}$.

Su Nature.com, un sito web scientifico di altissimo livello, troviamo qui troviamo una descrizione dei box plot, in cui leggiamo

[113] Si veda per esempio questo link.
[114] Per esempio in questo link vediamo una crocetta: possiamo ragionevolmente ipotizzare che rappresenti un decesso.
[115] Per esempio in questo link leggiamo "the whiskers show the smallest and highest value within 1.5 box lengths from the box", che è una variante comune.

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

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

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

```
.......-------!OOOOOOO!OOOOOOO!-----
```

e una finale di questo primo tipo o quest'altro tipo:

```
.........------!OOOOOO!OOOOOO!---
.....-------!OOOOOOO!OOOOOOOOOO!------
```

con media ridotta e valori addensati nel primo caso

oppure

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

Una nota finale nella pagina successiva.

20.3 I dati? Quali dati?

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

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

> era nata la campagna #DatiBeneComune, «per chiedere al Governo italiano di pubblicare in maniera aperta i dati sulla gestione della pandemia di COVID-19». Oggi la campagna, che ha superato i 60.000 firmatari, si rivolge ancora alle istituzioni per disporre di dati trasparenti sui principali temi di interesse dei cittadini, a partire dal Piano nazionale di ripresa e resilienza (PNRR).
>
> I dati, sempre i dati, croce di questo paese. Non dico che si debba avere l'attenzione spasmodica per i numeri degli Stati Uniti, dove c'è una maniacale ossessione per le statistiche persino nel baseball. Ma i numeri, i dati trasparenti e con il massimo grado di dettaglio possibile, dovrebbero essere lo spirito guida delle decisioni pubbliche, soprattutto quando ci sono in gioco investimenti importanti e strategie per il futuro.
>
> (...) L'analisi delle fonti a disposizione, a livello nazionale, mostra lacune che si sarebbe tentati di definire sconfortanti.
>
> «Il problema, peraltro, non è solo di mancanza del dato, ma anche, nei casi in cui esiste, di una ritrosia a rilasciare i dati», scrivono Faiella e Lavecchia. (Enfasi aggiunta)

La "sconfortante" "ritrosia a rilasciare i dati" sulla mortalità e lo stato vaccinale impedisce (2025) a tutt'oggi (!) a più di 4 anni dalle vaccinazioni di calcolare quanti giorni o mesi o anni di vita hanno (sperabilmente) guadagnato con la vaccinazione covid-19 coloro che si sono vaccinati, rispetto ai no-vax – il che genera sospetti nei *malfidenti*.

ESERCIZI SULLA LEZIONE 20

Esercizio$_{\mu}$ Si disegni il box-plot del dataset

$$x_k := \frac{1}{\pi - k} \qquad 1 \leq k \leq 10$$

rappresentando con un pallino l'outlier, escludendolo quindi dal calcolo del massimo.
Similmente poi con $1 \leq k \leq 11$.

20.3.1 Esercizio risolto a – Media interquartile

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

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

Il dataset ordinato in modo crescente, e scritto senza separatore delle migliaia, è
0,016 0,023 0,112 2,150 2,345 3,410 19,097 54,071 117,512 282,012 630,013 2848 Esso ha $n = 12$ cioè $4 \cdot 3$ elementi

••• ••• ••• •••

ed eliminando il primo e ultimo "quartile" dà
2,150 2,345 3,410 19,097 54,071 117,512
e la media aritmetica dei soprastanti 6 numeri

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

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

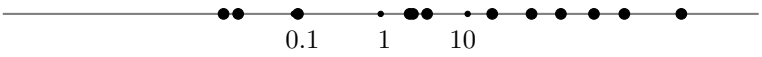
è la media interquartile cercata:

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

o con minor precisione

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

Nota 1. Salvo l'ordine il dataset è stato ottenuto da una variabile aleatoria log-normale di parametri $\mu = 1$ e $\sigma^2 = 25$, che non sono assolutamente media e varianza della variabile aleatoria. Precisamente sono stati tratti gli esponenziali dei valori della variabile aleatoria normale usata nell'esercizio 5: gli esponenziali dei valori di una variabile aleatoria normale sono distribuiti log-normalmente. Ecco un diagramma cartesiano in scala logaritmica dei 12 valori (ma 2 sono quasi sovrapposti). Si noti l'uguaglianza, salvo i numeri scritti all'asse delle ascisse, col diagramma dell'esercizo 37.6.1, che non è in scala logaritmica.



Il pallino più a destra corrisponde al 2848, il massimo del dataset,

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