19 Bar chart, istogrammi, e altri diagrammi

19.1 Diagrammi a torta

Il *diagramma a torta* ha un ovvio significato: gli angoli o equivalentemente le aree sono proporzionali contemporaneamente ai valori assoluti o alle frazioni percentuali da rappresentare. (La proporzionalità agli uni è equivalente alla proporzionalità alle altre). Si veda in questo link-> un diagramma a torta di interesse farmaceutico coi valori assoluti riportati (le percentuali no, sono lasciate all'occhio del lettore che vede le fette della torta; può essere un'utile esercizio calcolarne qualcuna).
(Più complessa è la situazione in cui oltre alle percentuali relative a 2 o più casi, il che necessita di 2 diagrammi a torta, si voglia rappresentare anche un altro dato per ciascuno dei casi, e allora si faranno diagrammi a torta di diverse dimensioni).
Gli angoli si misurano col goniometro; ma anche a occhio si può fare qualcosa.
Il diffuso software Excel fa i diagrammi a torta.
Funzionano veramente bene solo se i valori da rappresentare sono molto pochi, magari 2 o 3, e specialmente se sono molto diversi fra loro, come 66% e 32% e 2% – ma non di più di 2 ordini di grandezza: valori come 90%, 9.9%, 0.07% e 0.03% (l'ultimo ha solo circa 1 decimo di grado e il penultimo 2 e mezzo) non saranno concretamente apprezzabili su un normale foglio a stampa, o schermo di computer.

Spesso si raggruppano in una classe i valori più piccoli, o più grandi.
Si veda un esempio in https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia-2 dove in una stessa classe sono i casi "3 o più".

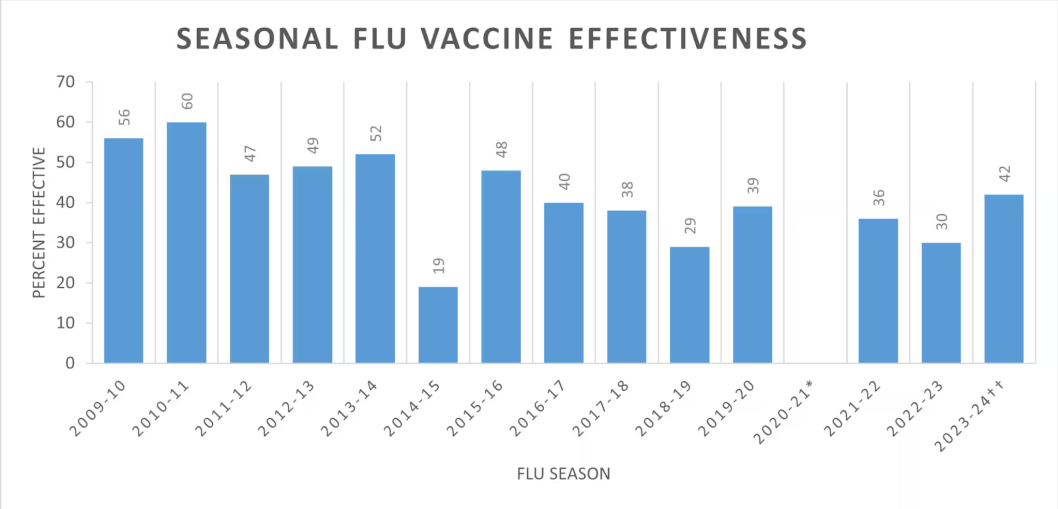
Si evitino come la peste gli ingannevoli diagrammi a torta tridimensionali in rappresentazione prospettica.

19.2 Catastrofe Statistica: dati mancanti e drop out

Una catastrofe della Statistica è il fenomeno dei dati mancanti.
Si veda un esempio nella prima figura del paragrafo seguente.

Un caso particolare di dati mancanti è quello dei *drop out*, partecipanti che abbandonano un trial clinico prima del suo completamento: sono una catastrofe della Statistica Medica. Ragioni il Lettore: se un cosmetico sperimentale è stato dato a 1000 persone, e dopo un mese rispondono a un questionario 400 persone, e magari risulta che mediamente sono soddisfatte, che valore ha quel risultato statistico? Magari sono alle Seichelles a godersi i risultati, o sono in gravi condizioni e non ci rispondono più al telefono; o sono morti?

19.3 Istogrammi a barre o bar chart, e istogrammi



https://www.cdc.gov/flu-vaccines-work/php/effectiveness-studies/index.html

* 2020-2021 flu vaccine effectiveness was not estimated due to low influenza virus circulation during the 2020-2021 flu season.

I diagrammi a colonne ovvero bar chart ovvero istogrammi a barre, e gli istogrammi – purtroppo con ambiguità nominalistiche nei vari testi – sono diagrammi per la visualizzazione di dati. Nei bar chart l'altezza di ogni colonna – o la lunghezza se disposta orizzontalmente – rappresenta un valore, negli istogrammi l'area rappresenta un valore:

bar chart – altezza ovvero lunghezza – valore

istogramma – area – valore

Spesso in italiano chiamano istogrammi quelli che in questo testo più precisamente chiamiamo bar chart.

Mostriamo un istogramma, nel senso di questo testo (area=valore), e poi ci occuperemo maggiormente di bar chart: LINK->
Si noti che la grandezza sull'asse delle ascisse è continua e non potrebbe essere che così.

Torniamo dunque ai bar chart.

Esempio. Dati rilevati a Chimica e Tecnologie Farmaceutiche di Trieste il 20/11/2023, sul numero di gatti e di cani che gli studenti (presenti in aula) hanno:

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

Istogramma a barre delle frequenze per il dataset dei gatti di CTF:

```
  0 gatti:  studenti 22 | ......................
   1 gatto:  studenti 6 | ......
   2 gatti:  studenti 2 | ..
   3 gatti:  studenti 2 | ..
```

Si noti che è possibile che una colonna rappresenti, per così dire, un valore medio di dati già considerati in altre colonne, per esempio in questo LINK->

Le varie barre possono avere colori che le associano fra loro: LINK->

Ma si faccia attenzione nell'interpretare le statistiche perché molto dipende da come si raggruppano i dati. Nel linkato diagramma sulle cause di morte il cancro compare poco perché è stato disaggregato in vari tipi di cancro. Considerandoli tutti, il cancro è la seconda causa di morte, dopo le malattie cardiocircolatorie. Addirittura qua disaggregando i tumori in tanti tipi diversi fa apparire il covid come prima causa di morte in Italia, datato 3 maggio 2020: LINK->

If you torture the data long enough, it will confess to anything.
(In questo LINK-> è in articolo su PubMed).

Si faccia attenzione nell'interpretare i bar chart in scala logaritmica perché le altezze rappresentano sì i valori ma non in modo proporzionale ad essi: LINK->

I bar chart possono rappresentare anche valori negativi: LINK->

Naturalmente si faccia attenzione a distinguere i valori assoluti dai valori relativi; per esempio di vari stati o regioni si possono considerare i morti di covid come numero assoluto, o come morti per milione o centomila abitanti. Vediamo un esempio in cui i valori sono già opportunamente scalati sul migliaio di abitanti in questi spettacolari diagrammi a bolle interattivi: LINK->

Si vedano vari tipi di bar chart di interesse specificamente farmaceutico in questo LINK->
Quella sorta di pelucchi che si vedono su certi bar chart rappresentano le deviazioni standard, concetto che vedremo, e che in sostanza rappresentano la variabilità della grandezza rappresentata, all'interno del campione da cui proviene.

ESERCIZIO$_{\mu 2019}$
≈ In un articolo scientifico[107] della rivista Neuropsychopharmacology leggiamo

> Patients were treated two to three times per week with high-frequency MST (i.e., 100 Hz) (N=24), medium frequency MST (i.e., 60 or 50 Hz) (N=26), or low-frequency MST (i.e., 25 Hz MST) (N=36)

Si disegni un istogramma a barre (bar chart) coi valori N_i denominando N l'asse delle ascisse, con barre denominate (lo si scriva entro esse)

high-frequency MST

medium frequency MST

[107] *Magnetic seizure therapy (MST) for major depressive disorder*, Neuropsychopharmacology (5 September 2019), Zafiris J. Daskalakis, Julia Dimitrova, Shawn M. McClintock, Yinming Sun, Daphne Voineskos, Tarek K. Rajji, David S. Goldbloom, Albert H. C. Wong, Yuliya Knyahnytska, Benoit H. Mulsant, Jonathan Downar, Paul B. Fitzgerald & Daniel M. Blumberger

low-frequency MST
e all'estremità di ogni barra si scriva la frequenza relativa, cioè N_i/N_{tot}, espressa percentualmente, con 1 decimale.

- Considerato il dataset

$$2, 3, 1, 4, 1, 5, 3, 4, 7, 1, 2, 4, 5, 9, 4, 8$$

se ne rappresenti il bar chart. Poi si rappresenti l'istogramma con intervalli $[0, 2.5[$, $[2.5, 5[$, $[5, 7.5[$, $[7.5, 10[$, e poi con intervalli $[0, 2.5[$, $[2.5, 5[$, $[5, 10[$.

19.4 Eventuali asimmetrie nei dataset: skewness

Una distribuzione coi dati più "addensati" verso i valori bassi che quelli alti si dice *right skewed*, e si intuisce com'è una distribuzione *left skewed*. (Queste *non* sono definizioni rigorose[108] ma permettono di decidere nella generalità dei casi non particolarmente "capricciosi").
Si provi con un diagramma a colonne coi dati

12.2 20%,
12.4 30%
12.6 25%
12.8 12.5%
13.0 5%
13.2 7%
13.4 3.5%
13.6 2%

19.5 Funzioni a campana varie

Facendo un bel po' di istogrammi a barre di dati presi dalla realtà sensibile, si troverà che spesso le barre si dispongono a formare più o meno una campana.

Praticamente dietro quasi ogni fenomeno della realtà sensibile c'è in qualche modo una funzione più

[108] Il lettore interessato cercherà sulla rete la *formula*, piuttosto complessa, che quantifica la skewness.

o meno a campana. Imparare a riconoscere queste configurazioni a campana aumenta enormemente la comprensione della realtà.

La campana può avere 2 significati principali: una densità, la classica
“distribuzione più o meno a campana",
coi suoi casi estremi rari e quelli medi più comuni,
oppure,
se in ascissa abbiamo il tempo, talvolta rappresenta (quantitativamente) la classica
“parabola della vita",
valida anche per una comunità microbica, per la potenza dell’Impero Romano, per la concentrazione di un farmaco immesso nel sangue, il numero di malati in un’epidemia, o quant’altro: sorgere, ascendere, declinare, finire.

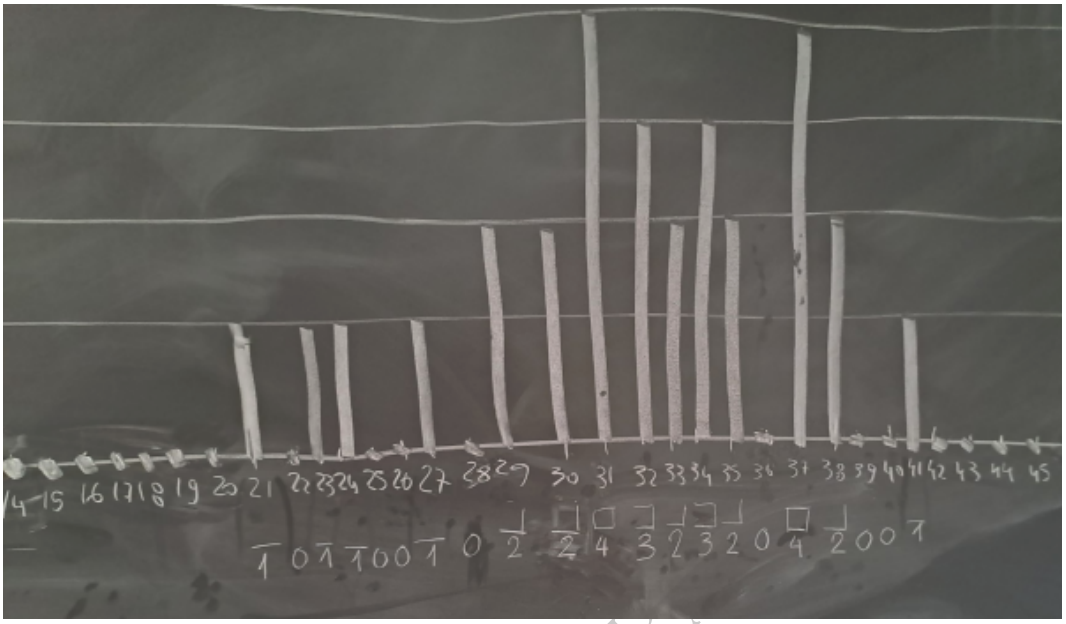


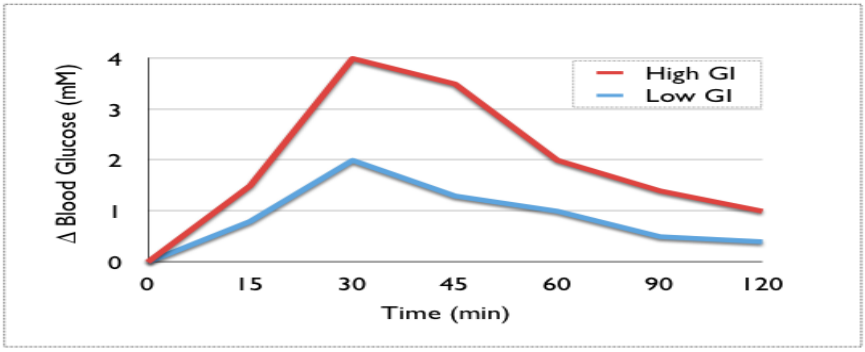
Figure 26: Qual era l'età della madre al momento della tua nascita? Anni compiuti, numero intero. Rilevazione in aula (2023) con questionario anonimo. Anni madre – frequenza: 21 1, 22 0, 23 1, 24 1, 25 0, 26 0, 27 1, 28 0, 29 2, 30 2, 31 4, 32 3, 33 2, 34 3, 35 2, 36 0, 37 4, 38 2, 39 0, 40 0, 41 1.

Qua stiamo parlando di curve "più o meno a campana" mentre con "curva a campana" *di solito* si intende proprio la *campana gaussiana*, grafico della *densità normale standard* $\frac{1}{\sqrt{2\,\pi}}\,\mathrm{e}^{-\frac{t^2}{2}}$, la forma più "pura" di curva a campana del primo tipo.

Curve più o meno a campana del primo tipo, cioè densità, (della Microbiologia) sono in questo link->.

Curve più o meno a campana del secondo tipo, cioè evoluzioni di una quantità nel tempo, sono in questa figura.[109]

[109] Immagine di pubblico dominio tratta da Wikimedia Commons.



Altre, relative ad epidemie di influenza, sono queste[110]
?????FIGURA MANCANTE
Un'altra, (della Microbiologia) è in questo link-> e un altro (della Farmacocinetica) in questo link-> e un altro a pagina 44 di questo link->.

Leggiamo nel Fedone, XXXIX (IV secolo a.C.), di Platone, citato in Wikipedia, l'enciclopedia libera, alla voce *Variabile casuale*:

> Credi forse che sia tanto facile trovare un uomo o un cane o un altro essere qualunque molto grande o molto piccolo o, che so io, uno molto veloce o molto lento o molto brutto o molto bello o tutto bianco o tutto nero? Non ti sei mai accorto che in tutte le cose gli estremi sono rari mentre gli aspetti intermedi sono frequenti, anzi numerosi?

Nota. Naturalmente nelle Scienze Applicate ricorrono anche grafici ben lontani dall'avere una forma a campana, pur così ubiqua. Già abbiamo visto sigmoidi e sinusoidi... Ma la realtà sensibile è ancora più ricca, per esempio si veda il grafico a questo LINK->.

[110] Tratte da Mid-season real-time estimates of seasonal influenza vaccine effectiveness in persons 65 years and older in register-based surveillance, Stockholm County, Sweden, and Finland, January 2017. Euro Surveill. 2017 Feb 23;22(8). pii: 30469. doi: 10.2807/1560-7917.ES.2017.22.8.30469. By Hergens MP et al. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5356437/. "This is an open-access article distributed under the terms of the Creative Commons Attribution (CC BY 4.0) Licence. You may share and adapt the material, but must give appropriate credit to the source, provide a link to the licence, and indicate if changes were made." LINK->

ESERCIZI SULLA LEZIONE 19

19.5.1 Esercizio risolto a – Bar chart

$_{\mu}$ *

Consideriamo il modello di un fenomeno, e per fissare le idee diciamo che è un'epidemia. (Anche se quello che considereremo non è il miglior modello proprio per le epidemie). Il numero di morti dell'epidemia sia modellizzato da

$$x_n = \left\lfloor \frac{20\, n^2}{n\,!} \right\rfloor \qquad n = 1, 2, \ldots$$

essendo n il numero di giorni dall'inizio ed indicando il simbolo $\lfloor \ldots \rfloor$ la parte intera (quella che "toglie i decimali" ai numeri positivi, per esempio $\lfloor 3.14 \rfloor = 3$). Fare un istogramma a barre fino all'ultimo giorno con morti (o per meglio dire fino al giorno precedente del primo giorno con 0 morti).

SVOLGIMENTO

Viene lo standard del punto decimale. (Nello svolgimento e già nel testo del quesito, in cui si trova il numero 3.14).

$$x_1 = \left\lfloor \frac{20 \cdot 1 \cdot 1}{1} \right\rfloor = 20$$

$$x_2 = \left\lfloor \frac{20 \cdot 2 \cdot 2}{1 \cdot 2} \right\rfloor = 40$$

$$x_3 = \left\lfloor \frac{20 \cdot 3 \cdot 3}{1 \cdot 2 \cdot 3} \right\rfloor = 30$$

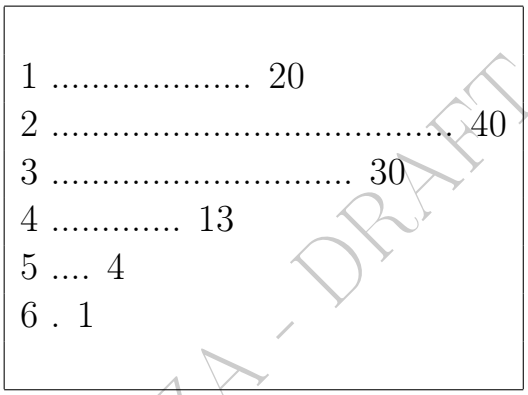
$$x_4 = \left\lfloor \frac{20 \cdot 4 \cdot 4}{1 \cdot 2 \cdot 3 \cdot 4} \right\rfloor = \left\lfloor \frac{40}{3} \right\rfloor = \lfloor 13.33\ldots \rfloor = 13$$

$$x_5 = \left\lfloor \frac{20 \cdot 5 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} \right\rfloor = \left\lfloor \frac{100}{24} \right\rfloor = \lfloor 4.16\ldots \rfloor = 4$$

$$x_6 = \left\lfloor \frac{20 \cdot 6 \cdot 6}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} \right\rfloor = 1$$

$$x_7 = \left\lfloor \frac{20 \cdot 7 \cdot 7}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7} \right\rfloor = \left\lfloor \frac{140}{720} \right\rfloor = \lfloor 0.19... \rfloor = 0$$

e abbiamo raggiunto lo 0 previsto. Il numero di morti x_n calcolato dal modello poi non si risolleva da 0 com'è dimostrabile con passaggi facili però laboriosi, ma questa osservazione non ci serve perchè nel testo è detto di fermarsi al giorno precedente del primo giorno con 0 morti, che è il giorno numero 7, come abbiamo trovato. Allora l'epidemia dura 6 giorni, e possiamo produrre l'istogramma a barre.



(Che molto meglio sarà rappresentare con le aste disposte verticalmente).

Nota. Ecco una soluzione online su WolframAlpha:

LINK -> (con l'istruzione `DiscretePlot`)

Rispetto a quello dato nella risoluzione dell'esercizio, questo risultato di WolframAlpha

– ha il pregio della precisione

– ha il pregio delle barre disposte verticalmente

– ha il difetto che non scrive sulle barre i 6 valori effettivi.

Ecco un'atra soluzione online su WolframAlpha:

LINK -> (con le istruzioni `BarChart` e `Table`)

Rispetto a quello dato nella risoluzione dell'esercizio, questo risultato di WolframAlpha

– di nuovo, ha il pregio della precisione

– di nuovo, ha il pregio delle barre disposte verticalmente

– ha il pregio che le barre hanno forma di rettangoli e non segmenti

– di nuovo, ha il difetto che non scrive sulle barre i 6 valori effettivi

– ha il difetto che è piuttosto complessa da trovare, richiedendo l'istruzione `Table`

19.5.2 Esercizio risolto b – Istogramma

$_{\mu 2018}$ [*disegno*]
In relazione all'elemento sodio (Na), definite le cellule

ipo-sodiche: con contenuto di sodio fra 0 e < 1 000 unità

normo-sodiche: con contenuto di sodio fra 1 000 e < 10 000 unità

iper-sodiche: con contenuto di sodio fra 10 000 e < 20 000 unità

(unità che non specifichiamo, essendo ininfluente) un macchinario esamina 52 000 cellule (tutti numeri arrotondati per semplicità) trovandovi

6 000 ipo-sodiche, 36 000 normo-sodiche, 10 000 iper-sodiche.

Rappresentare la situazione con un istogramma. (Non è il bar chart).

SVOLGIMENTO

Nell'istogramma – che purtroppo alcuni Autori scambiano col bar chart ma qua veniamo avvertiti della differenza – le grandezze sono proporzionali alle aree dei rettangoli (invece nei bar chart alle lunghezze delle aste, eventualmente strisce rettangolari).

Avremo quindi 3 rettangoli, con basi che si estendono su un asse orizzontale

da 0 a 1000 lunga 1 000

da 1 000 a 10 000 lunga 9 000

da 10 000 a 20 000 lunga 10 000.

(Non vi è alcuna questione problematica fra valori compresi o esclusi).

Le altezze le otteniamo dividendo le aree

6 000, 36 000, 10 000 (o numeri ad essi rispettivamente proporzionali)

per le lunghezze delle basi corrispondenti, ottenendo:

6 unità
4 unità
1 unità

Come unità si può prendere per esempio 1 centimetro, o su carta a quadretti 1 o 2 o 3 quadretti, o anche altre unità.
Ecco l'istogramma, coi 3 rettangoli ed una legenda:

```
A----------
A---------- A=ipo-sodiche, 6 000
ABBBBBBBBB------ B=normo-sodiche, 36 000
ABBBBBBBBB------ C=iper-sodiche, 10 000
ABBBBBBBBB------
ABBBBBBBBBCCCCCCCCCC
!!----!-----!
0 1000      10000          20000
```