# Compendio di
# **Statistica Descrittiva**
# in preparazione all'esame di stato

Simone Zuccher

21 aprile 2013

## Indice



---

## 1 Indagine statistica

Il termine *statistica* significa *scienza dello stato*. Questo termine venne usato per la prima volta nel XVI secolo per indicare lo studio dei dati utili al governo degli stati, prevalentemente relativi a fenomeni di carattere demografico (nascite, morti, etc). Negli anni, poi, la statistica si è estesa ai campi più disparati: dalla fisica alla psicologia, alla ricerca di mercato. È nata essenzialmente con lo scopo di descrivere in maniera *sintetica* fenomeni relativi ad un certo gruppo di persone, animali o oggetti tramite la raccolta e l'analisi dei dati (*statistica descrittiva*) ed è successivamente divenuta uno strumento utile per fare previsioni sull'andamento futuro (*statistica inferenziale*).

In questo compendio ci occuperemo unicamente della *statistica descrittiva* il cui scopo è di riassumere in pochi numeri significativi grandi moli di dati. Croce e delizia della statistica descrittiva è proprio questo: voler riassumere in pochi numeri grandi quantità di dati implica necessariamente la *perdida di informazione* sulla provenienza e sulla diversità del dato. Ad esempio, dire che la media di uno studente, in pagella è 7, non dice nulla su come vada in matematica scritta piuttosto che in inglese orale.

**Definizione 1.1** *L'insieme di elementi oggetto dell'indagine statistica è detta* popolazione o universo*, mentre ciascun elemento della popolazione è detto* unità statistica*.*

Sono esempi di popolazioni gli abitanti di una città in un certo anno, i prezzi di un determinato bene, le temperature massime registrate in una giornata in un particolare luogo, i ciclomotori circolanti in Italia, gli alunni di una scuola.

**Definizione 1.2** *Per ogni unità statistica si possono studiare una o più caratteristiche ed ognuna di tali caratteristiche costituisce un* carattere *della popolazione.*

Esempi di caratteri su un campione di persone sono: l'altezza, l'età, il colore degli occhi, il genere, il segno zodiacale, il credo religioso. Si osservi che alcuni caratteri sono *quantitativi*, ossia esprimono una quantità attraverso dei valori numerici (l'altezza, il numero di scarpe, l'età, ecc.) mentre altri sono *qualitativi* in quanto esprimono una qualità attraverso dei valori non numerici (il genere o il credo religioso). Mentre i caratteri quantitativi sono naturalmente *ordinabili*, quelli qualitativi possono essere *ordinabili* (si pensi ad una scala che esprime un giudizio "pessimo", "cattivo", "mediocre", "buono" e "ottimo") oppure *non ordinabili* (le malattie o il colore degli occhi). Inoltre, i caratteri *quantitativi* possono essere di tipo *discreto*, quando assumono solo valori puntuali, oppure di tipo *continuo*, quando possono assumere tutti gli infiniti valori compresi (o meno) in un determinato intervallo. Sono esempi di caratteri quantitativi discreti il numero di figli in una famiglia o i pezzi prodotti in una catena di montaggio, mentre sono esempi di caratteri continui l'altezza di una persona, il peso di una persona e la lunghezza di un fiume.

Se l'indagine statistica riguarda l'intera popolazione si parla di *censimento*, se riguarda solo una sua parte si parla di *indagine a campione.*

## 2 I dati e le loro rappresentazioni

Nella tabella 1 sono riportati i dati raccolti in una classe di 20 alunni, ai quali si è chiesto la misura del numero di scarpe e l'altezza (in metri). I cognomi sono i 20 più diffusi in Italia (non in ordine di diffusione ma in ordine alfabetico,

come in ogni elenco). Questa tabella descrive tutti i dettagli della raccolta dei dati e permette di risalire alla misura delle scarpe e all'altezza dei singoli alunni. Tuttavia è *poco riassuntiva*.

| Cognome | Numero di scarpe | Altezza [m] |
|---|---|---|
| Barbieri | 46 | 1.94 |
| Bianchi | 38 | 1.67 |
| Bruno | 44 | 1.84 |
| Colombo | 44 | 1.82 |
| Conti | 38 | 1.68 |
| Costa | 39 | 1.70 |
| De Luca | 39 | 1.71 |
| Esposito | 39 | 1.69 |
| Ferrari | 44 | 1.86 |
| Gallo | 41 | 1.76 |
| Giordano | 41 | 1.75 |
| Greco | 40 | 1.73 |
| Lombardi | 45 | 1.87 |
| Mancini | 39 | 1.69 |
| Marino | 42 | 1.81 |
| Moretti | 40 | 1.75 |
| Ricci | 44 | 1.86 |
| Romano | 42 | 1.79 |
| Rossi | 39 | 1.70 |
| Russo | 38 | 1.69 |

Tabella 1: Numero di scarpe ed altezza degli alunni di una classe

Se ci concentriamo sul numero di scarpe, un primo passo verso la compattezza della rappresentazione dei dati è quello di ordinare, in un'altra tabella, la misura del numero di scarpe, ossia il *carattere*, con il corrispondente numero di alunni che portano quel numero di scarpe, ossia la *frequenza assoluta* di quel carattere. Questo è fatto in tabella 2.

| Misura | Numero di alunni |
|---|---|
| 38 | 3 |
| 39 | 5 |
| 40 | 2 |
| 41 | 2 |
| 42 | 2 |
| 44 | 4 |
| 45 | 1 |
| 46 | 1 |

Tabella 2: Tabella delle frequenze assolute, ossia del numero di alunni che portano un determinato numero di scarpe

Prima di tutto si osservi che in questa tabella ci sono solo 2 colonne di 8 elementi ciascuna, a differenza della tabella 1 in cui vi erano 2 colonne di 20 elementi ciascuna. La prima colonna indica il *carattere*, che in questo caso è *quantitativo ed ordinabile*, mentre la seconda colonna indica la *frequenza assoluta*, ossia il numero di volte che il carattere si ripete nella popolazione in esame. Per questo motivo essa viene detta *tabella delle frequenze assolute*.

La tabella permette di capire, a colpo d'occhio, che il numero più basso è il 38 e quello più alto è il 46, mentre quello più frequente è il 39. Evidentemente, compattando i dati in questa tabella si è persa l'informazione su chi è lo studente con il numero 46 o chi sono quelli che portano il 44.

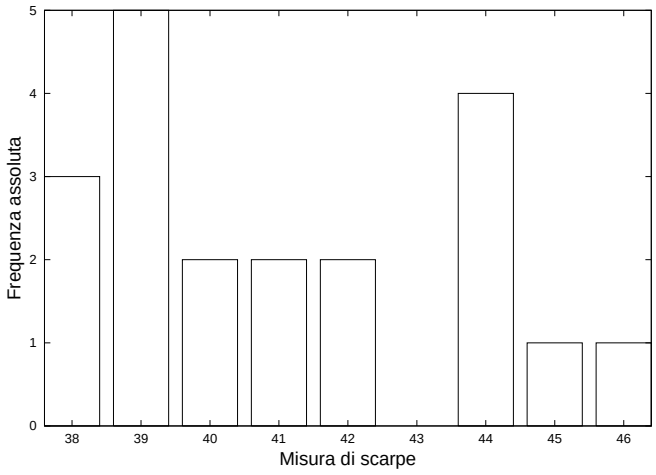



Figura 1: Istogramma che riporta il numero di studenti aventi un determinato numero di scarpe

Anziché utilizzare una tabella, i dati possono essere riportati in una figura in cui in ascissa si mette il carattere (nel nostro caso la misura di scarpe) e in ordinate la frequenza assoluta con la quale quel carattere compare nella popolazione (il numero di studenti che portano quella misura). Invece di unire i dati con una spezzata, vengono utilizzati dei rettangoli le cui altezze sono proprio le *frequenze* con le quale le misura si ripetono, come riportato in figura 1. Si osservi che una figura di questo tipo può essere fatta anche se i caratteri sono qualitatitivi e anche se non sono ordinabili.

Invece di utilizzare le frequenze assolute, che non dicono quanto un carattere si presenta, percentualmente, all'interno di una popolazione, si possono usare le *frequenze relative* o le *frequenze percentuali*.

**Definizione 2.3** *La* frequenza assoluta *è il numero di volte con cui si presenta una modalità del carattere indagato.*
*La* frequenza relativa *è il rapporto tra la frequenza assoluta ed il numero totale dei casi presi in esame.*
*La* frequenza percentuale *è la frequenza relativa moltiplicata per 100.*

Le frequenze assolute, relative e percentuali per i dati della tabella 2 sono riportate in tabella 3.

| Misura | Frequenza assoluta | Frequenza relativa | Frequenza percentuale |
|---|---|---|---|
| 38 | 3 | 0.15 | 15% |
| 39 | 5 | 0.25 | 25% |
| 40 | 2 | 0.10 | 10% |
| 41 | 2 | 0.10 | 10% |
| 42 | 2 | 0.10 | 10% |
| 44 | 4 | 0.20 | 20% |
| 45 | 1 | 0.05 | 5% |
| 46 | 1 | 0.05 | 5% |
| Totali: | 20 | 1 | 100% |

Tabella 3: Tabella delle frequenze assolute, relative e percentuali

Evidentemente, se anziché diagrammare le frequenze assolute si diagrammano le frequenze relative o percentuali l'istrogramma di figura 1 non cambia forma, cambiano solo i valori sull'asse verticale (il massimo è 0.25 per le frequenze relative, e 25% per quelle percentuali).

Fino ad ora abbiamo analizzato il carattere "numero di scarpe", che è quantitativo, ordinabile e discreto. Analizziamo adesso l'altezza dei singoli studenti che, pur essendo sempre un carattere quantitativo ed ordinabile, è però *continuo*. A rigor di logica non è proprio continuo, perché assume tutti i valori tra 1.67 m e 1.94 m, quindi se si considerano intervalli di un centrimetro si hanno 28 possibili valori. Per dati come l'altezza di una persona (che può andare da circa 40 cm negli infanti a 2.20 m in individui eccezionalmente alti) o l'età (che può andare da meno di un anno per i neonati a più di cent'anni per anziani centenari) è più opportuno suddividere in dati in *classi*, ossia in particolari *intervalli non necessariamente equispaziati*.

| Altezza | Numero di alunni |
|---|---|
| 1.65 – 1.70 | 5 |
| 1.70 – 1.75 | 4 |
| 1.75 – 1.80 | 4 |
| 1.80 – 1.85 | 2 |
| 1.85 – 1.90 | 4 |
| 1.90 – 1.95 | 1 |

Tabella 4: Tabella delle classi di altezza di intervallo costante (5 cm)

Ad esempio, per i dati sulle altezze degli studenti, potremmo suddividere in intervalli di 5 cm, da 1.65 m a 1.70 m (estremo inferiore compreso, estremo superiore escluso), da 1.70 m a 1.75 m (estremo inferiore compreso, estremo superiore escluso), e così via, ottenendo la tabella 4.

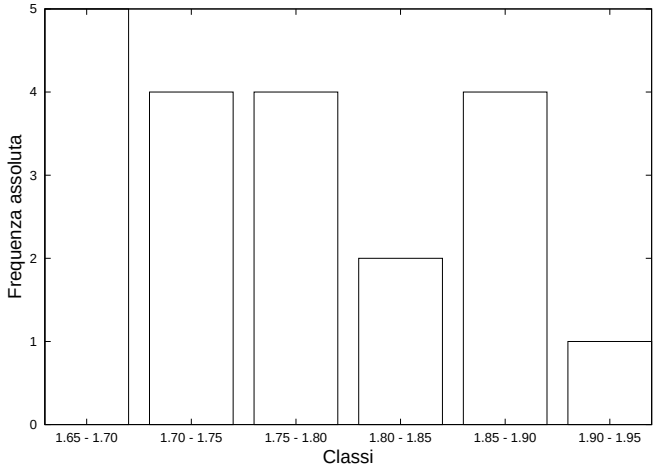



Figura 2: Istogramma che riporta le altezze degli studenti divise in 6 classi ciascuna di ampiezza 5 cm

| Altezza | Frequenza assoluta | Frequenza relativa | Frequenza percentuale |
|---|---|---|---|
| 1.65 – 1.70 | 5 | 0.25 | 25% |
| 1.70 – 1.75 | 4 | 0.20 | 20% |
| 1.75 – 1.80 | 4 | 0.20 | 20% |
| 1.80 – 1.85 | 2 | 0.10 | 10% |
| 1.85 – 1.90 | 4 | 0.20 | 20% |
| 1.90 – 1.95 | 1 | 0.05 | 5% |
| Totali: | 20 | 1 | 100% |

Tabella 5: Tabella delle classi di altezza di intervallo costante (5 cm), frequenze assolute, relative, e percentuali

Come fatto in precedenza, i dati di questa tabella possono essere diagrammati in un istogramma come quello riportato in figura 2 o riarrangiati in una tabella che riporta le frequenze relative e percentuali come la tabella 5.

Rappresentazione dei dati statistici, come visto, può avvenire sia sotto forma di tabelle sia sotto forma grafica. Le rappresentazioni grafiche più diffuse sono

- *diagramma cartesiano*: rappresentazione nel piano cartesiano dei valori della variabile sull' asse orizzontale e della relative frequenze sull'asse verticale;
- *ideogramma*: si rappresenta un certo numero di dati con un simbolo;
- *diagramma a nastri o a bastoni*: utilizzata prevalentemente per addetti ai lavori;
- *areogramma*: grafico a forma di cerchio composto da settori circolari con aree direttamente proporzionali alle frequenze;
- *istogramma*: grafico composto da rettangoli aventi area proporzionale alla frequenza.

# 3 Le medie statistiche

Abbiamo visto come ridurre i dati da una tabella molto dettagliata a tabelle più compatte e a grafici più o meno significativi.

La domanda che ci poniamo ora è: è possibile ridurre i dati ad un *solo valore* indicativo dei dati in questione?

Esistono vari modi per identificare un *valore significativo*, che potremmo anche chiamare *valore centrale* o *valore medio*. Inziamo con la media, che può essere di vario tipo.

**Definizione 3.4** *Siano $x_1, x_2, \dots, x_N$ i valori assunti da un particolare carattere (ossia le sue* modalità*), definiamo*

- *media aritmetica*
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
- *media geometrica*
$$\bar{x}_g = \sqrt[N]{x_1 \cdot x_2 \cdots x_N}$$
- *media armonica*
$$\bar{x}_a = \frac{1}{\frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}}{N}} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}}$$
- *media quadratica*
$$\bar{x}_2 = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_N^2}{N}}$$

Si può dimostrare che

$$\bar{x}_a < \bar{x}_g < \bar{x} < \bar{x}_2.$$

**Definizione 3.5** *Siano $x_1, x_2, \dots, x_N$ i valori assunti da un particolare carattere (ossia le sue* modalità*), e $p_1, p_2, \dots, p_N$ dei numeri reali positivi ad essi associati che chiamiamo* pesi*, definiamo*

- *media aritmetica ponderata*

$$\bar{x}_p = \frac{p_1x_1 + p_2x_2 + \cdots + p_Nx_N}{p_1 + p2 + \ldots p_N} = \frac{\sum_{i=1}^{N} p_ix_i}{\sum_{i=1}^{N} p_i}$$

- *media geometrica ponderata*

$$\bar{x}_{gp} = \sqrt[N]{x_1^{p_1} \cdot x_2^{p_2} \cdots x_N^{p_N}}$$

- *media armonica ponderata*

$$\bar{x}_{ap} = \frac{1}{\frac{\frac{p_1}{x_1} + \frac{p_2}{x_2} + \cdots + \frac{p_1}{x_N}}{p_1 + p_2 + \ldots p_N}} = \frac{p_1 + p_2 + \ldots p_N}{\frac{p_1}{x_1} + \frac{p_2}{x_2} + \cdots + \frac{p_1}{x_N}}$$

- *media quadratica ponderata*

$$\bar{x}_{2p} = \sqrt{\frac{p_1x_1^2 + p_2x_2^2 + \cdots + p_Nx_N^2}{p_1 + p_2 + \ldots p_N}}$$

Altre misure *centrali* di una serie di valori $x_i$ sono la *moda* e la *mediana*

**Definizione 3.6** *Chiamiamo* moda *(o termine modale) di una serie di dati il termine che compare con la massima frequenza.*

**Definizione 3.7** *Chiamiamo* mediana *il termine che occupa la posizione centrale in una serie di dati quando essi sono disposti in ordine crescente. Se il numero di dati è pari, la mediana è la media aritmetica dei due dati centrali.*

Per esempio, nel caso dell'istogramma in figura 1, la moda è facilmente riconoscibile ed è il numero di scarpe 39, mentre per determinare la mediana è necessario disporre tutti i numeri in ordine crescente: 38, 38, 38, 39, 39, 39, 39, 39, 40, 40, 41, 41, 42, 42, 44, 44, 44, 44, 45, 46, da cui la mediana $40.5 = \frac{40+41}{2}$. Si osservi che la media aritmetica dei numeri di scarpe è $\bar{x} = 41.1$, quindi la media aritmetica, la moda e la mediana non sono necessariamente valori concidenti (sperabilmente non sono molto diversi se i dati sono distribuiti in modo non troppo disuniforme).

# 4 Variabilità e concentrazione dei dati statistici

Nella sezione precedenti ci si è occupati di descrivere una serie di dati tramite un solo valore, che in qualche modo li riassuma. Spesso, tuttavia, questo non è sufficiente. In figura 3 sono riportati i voti di due alunni, Crocette e Pallini, in funzione del numero di prove di matematica sostenute durante l'anno scolastico.

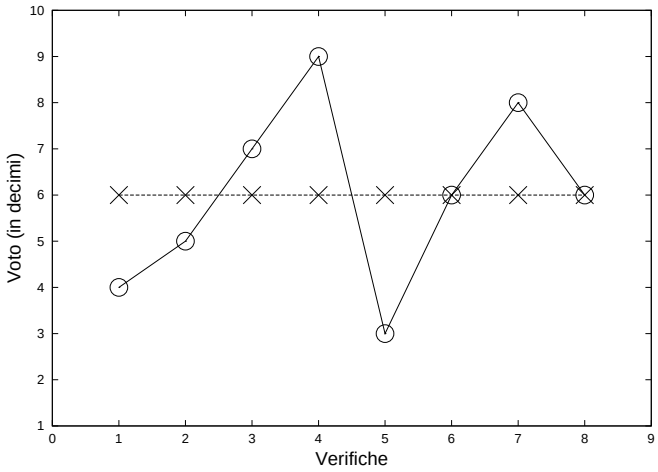



Figura 3: Andamento dei voti di due ragazzi, Crocette e Pallini, nelle verifiche di matematica sostenute durante l'anno scolastico

Evidentemente entrambi hanno la media del 6, quindi saranno promossi. Tuttavia Crocette è uno studente molto costante e in tutte le prove ha sempre preso 6, mentre Pallini ha un andamento molto oscillatorio: è partito da voti negativi, si è via via ripreso, poi ha avuto un calo ma, infine, è riuscito a recuperare. La domanda è: in che modo si può *misurare* lo scostamento dei dati dal valor medio?

**Definizione 4.8** *Chiamiamo scarto dalla media la differenza* $s_i = x_i - \bar{x}$, *dove* $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

Una misura dello scostamento dei dati potrebbe essere il massimo scarto preso in valore assoluto,

$$|x_i - \bar{x}|_{\max}.$$

Un'altra possibilità che misura una *distanza media dal valore medio* potrebbe essere la media degli scarti,

$$\frac{1}{N}\sum_{i=1}^{N} s_i = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}),$$

ma questa quantità è nulla, come si verifica facilmente dopo aver osservato che dalla definizione di media artimetica si ottiene immediatamente $\sum_{i=1}^{N} x_i = N\bar{x}$:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} s_i &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) \\ &= \frac{1}{N}\left[\sum_{i=1}^{N} x_i - \sum_{i=1}^{N}\bar{x}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}N\bar{x} \\ &= \bar{x} - \bar{x} = 0.\end{aligned}$$

Un'eventuale misura significativa di uno *scarto medio* potrebbe essere la media dei valori assoluti degli scarti:

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} |s_i| = \frac{1}{N}\sum_{i=1}^{N} |x_i - \bar{x}|,$$

tuttavia questa media dei moduli degli scarti non viene mai usata. Una misura molto utilizzata è invece lo *scarto quadratico medio*.

**Definizione 4.9** *Siano $x_1, x_2, \ldots, x_N$ i valori di una serie di $N$ dati e sia $\bar{x}$ la loro media aritmetica: chiamiamo* scarto quadratico medio*, e lo indichiamo con la lettere creca $\sigma$, la quantità*

$$\sigma = \bar{s}_2 = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}.$$

**Definizione 4.10** *Il quadrato dello scarto quadratico medio, ossia $\sigma^2$, prende il nome di* varianza*.*

Lo scarto quadratico medio è identicamente nullo se tutti gli $x_i$ sono proprio uguali al valor medio $\bar{x}$, come nel caso dello studente Crocette, ossia se non c'è nessuno scostamento dal valore medio. Per lo studente Pallini, invece, $\sigma = 1.87$ che significa che i suoi voti sono *mediamente* lontani quasi di 2 voti dal 6.

Si può dimostrare che lo scarto quadratico medio calcolato rispetto alla media aritmetica $\bar{x}$ è il minor scarto possibile, ossia uno scarto quadratico medio calcolato rispetto ad un valore medio diverso dalla media aritmetica dà una valore superiore. Inoltre, il calcolo dello scarto quadratico medio può essere fatto velocemente osservando che

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2},$$

che può essere facilmente ricordato come la *radice quadrata della differenza tra la media dei quadrati ed il quadrato della media*.

Infatti, ricordando che dalla definizione di media aritmetica si ha $\sum_{i=1}^{N} x_i = N\bar{x}$, sviluppando il quadrato si ottiene

$$\begin{aligned}
\sigma &= \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \\
&= \sqrt{\frac{\sum_{i=1}^{N}(x_i^2 - 2x_i\bar{x} + \bar{x}^2}{N}} \\
&= \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - 2\bar{x}\sum_{i=1}^{N} x_i + \sum_{i=1}^{N}\bar{x}^2}{N}} \\
&= \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - 2N\bar{x}^2 + N\bar{x}^2}{N}} \\
&= \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2}.
\end{aligned}$$

Se lo scarto quadratico medio misura lo scostamento di una serie di dati dal loro valor medio, a volte può essere interessante un indicatore opposto, ossia uno che misuri la *concentrazione* dei dati. Ad esempio, se la popolazione viene suddivisa in fasce di reddito, sarebbe interessante sapere se il reddito è equamente distribuito tra la popolazione oppure se è concentrato nelle mani di pochi ricchi.

Per ottenere una misura della concentrazione dei dati, utilizziamo l'altezza degli studenti divisa per classi come riportato in tabella 5. Prima di tutto *ordiniamo i dati in base alla loro frequenza (assoluta, relativa o percentuale, non ha nessuna importanza), da quello meno frequente a quello più frequente* ottenendo così la tabella 6.

| | Classe di altezza | Frequenza assoluta | Frequenza relativa | Frequenza percentuale |
|---|---|---|---|---|
| 1 | 1.90 – 1.95 | 1 | 0.05 | 5% |
| 2 | 1.80 – 1.85 | 2 | 0.10 | 10% |
| 3 | 1.70 – 1.75 | 4 | 0.20 | 20% |
| 4 | 1.75 – 1.80 | 4 | 0.20 | 20% |
| 5 | 1.85 – 1.90 | 4 | 0.20 | 20% |
| 6 | 1.65 – 1.70 | 5 | 0.25 | 25% |
| | Totali: | 20 | 1 | 100% |

Tabella 6: Ordinamento delle classi di altezza da quelle meno frequenti a quelle più frequenti

Poi costruiamo due serie di dati $(p_i, q_i)$ così ottenuti: $p_i$ è la *posizione percentuale* di quella particolare classe, mentre $q_i$ è la *frequenza percentuale cumulata* di quella particolare classe, come riportato in tabella 7

| Posizione assoluta | Posizione percentuale | Frequenza percentuale | Frequenza percentuale cumulata |
|---|---|---|---|
| | $p_i$ | | $q_i$ |
| 1 | 16.67% | 5% | 5% |
| 2 | 33.33% | 10% | 15% |
| 3 | 50.00% | 20% | 35% |
| 4 | 66.67% | 20% | 55% |
| 5 | 83.33% | 20% | 75% |
| 6 | 100.00% | 25% | 100 % |

Tabella 7: Ordinamento delle classi di altezza da quelle più frequenti a quelle meno frequenti

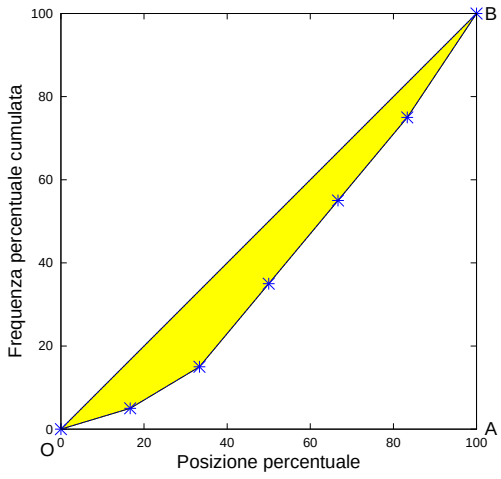



Figura 4: Curva di Lorenz costituita dalle coppie $(p_i, q_i)$ (pallini blu uniti da spezzate) e area di concentrazione (in giallo)

Infine, diagrammiamo le coppie $(p_i, q_i)$ (a partire dal punto iniziale $(0,0)$), ottenendo la *curva di Lorenz*, riportata in figura 4. L'area compresa tra la retta che unisce i punti $O(0,0)$ e $B(100,100)$ e la curva di Lorenz è detta *area di concentrazione* (nella figura è indicato in giallo). Si osservi che, nel caso in cui tutte le classi compaiano con la stessa frequenza (equidistribuzione del carattere), la curva di Lorenz è proprio la retta $OB$ e quindi l'area di concentrazione è nulla. Al contrario, nel caso limite in cui tutti i dati fossero concentrati in una sola classe, la curva di Lorenz sarebbe la spezzata $OAB$ e l'area di concentrazione sarebbe l'area del triangolo $OAB$.

**Definizione 4.11** *Chiamiamo* rapporto di concentrazione $R$ *il rapporto tra l'area di concentrazione (area racchiusa tra la retta $OB$ e la curva di Lorenz) e l'area del triangolo $OAB$,*

$$R = \frac{\textit{area di concentrazione}}{\textit{area del triangolo } OAB}.$$

Evidentemente $0 \leq R \leq 1$, in paticolare $R = 0$ per dati equamente distribuiti e $R = 1$ per dati concentrati in una sola classe. Se con $A_\mathrm{T}$ indichiamo l'area del triangolo $OAB$, con $A_\mathrm{C}$ indichiamo l'area di concentrazione e con $A_\mathrm{L}$ l'area sotto alla curva di Lorenz, allora si ha

$$R = \frac{A_\mathrm{C}}{A_\mathrm{T}} = \frac{A_\mathrm{T} - A_\mathrm{L}}{A_\mathrm{L}} = 1 - \frac{A_\mathrm{L}}{A_\mathrm{T}}.$$

Nel caso della curva di Lorenz rappresentata in figura 4 il rapporto di concentrazione vale $R = 0.217$, ossia i dati sono relativamente distribuiti.

# 5 Interpolazione statistica

Supponiamo di conoscere $N$ coppie di dati $(x_i, y_i)$ con $i = 1, \ldots, N$. Ad esempio,il numero di scarpe e le altezze di ogni studente riportati nella tabella 1 sono, rispettivamente, $x_i$ e $y_i$.

Dal punto di vista matematico, interpolare questi dati significa determinare una funzione $y = f(x)$ che riproduca *esattamente* i dati noti, ossia nel caso di *interpolazione matematica* deve essere $f(x_i) = y_i$ per ogni $i = 1, \ldots, N$. Innanzitutto si osserva che per i dati della tabella 1 questo può essere problematico in quanto allo stesso numero di scarpe corrispondono altezze diverse (e quindi si avrebbe una funzione a più valori). Se invece si pensa che le coppie $(x_i, y_i)$ descrivano un grafico che dà la temperatura di una città ad ogni ora della giornata, allora la *funzione interpolante* può essere interessante per *estrapolare* dei dati mancanti come ad esempio la temperatura alle 16:30 del pomeriggio (che potrebbe essere semplicemente la media tra la temperatura alle 16:00 e quella alle 17:00) o la temperatura alle 16:45.

L'*interpolazione statistica*, al contrario dell'interpolazione matematica, si propone di determinare una funzione che descriva la relazione che intercorre tra l'insieme di dati $\{x_i\}$ e l'insieme di dati $\{y_i\}$. La *funzione interpolante*, in questo caso, *non passa necessariamente per i punti* $(x_i, y_i)$, ossia in generale si ha $f(x_i) \neq y_i$, anche se può succedere che si riesca a determinare una funzione tale che sia proprio $f(x_i) = y_i$ per ogni $i = 1, \ldots, N$.

**Interpolazione lineare**

La funzione più semplice che può descrivere la relazione che intercorre tra l'insieme di dati $\{x_i\}$ e l'insieme di dati $\{y_i\}$ è la retta

$$y = f(x) = mx + q.$$

Ci proponiamo di determinare $m$ e $q$ in modo che *la retta sia la migliore possibile* a descrivere i dati. Siccome non tutti i dati si troveranno allineati su questa retta, chiamiamo $\hat{y}_i$ il valore $f(x_i)$, ossia

$$\hat{y}_i = mx_i + q.$$

La *migliore retta possibile* è quella che si discosta meno, in senso *globale*, dalla nuvola di dati, ossia quella per la quale si ha che la quantità

$$\sum_{i=1}^{N} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{N} [y_i - mx_i - q]^2$$

è minima. In altre parole, si tratta di determinare la coppia $(m, q)$ che minimizzi la funzione $\varphi(m, q)$ così definita

$$\varphi(m, q) = \sum_{i=1}^{N} [y_i - mx_i - q]^2.$$

Siccome $\varphi(m, q)$ è una funzione sempre positiva e somma di quadrati, il minimo coincide con il punto in cui si annullano simultaneamente entrambe le *derivate parziali* di $\varphi(m, q)$, sia rispetto a $m$ che rispetto a $q$:

$$\min(\varphi(m,q)) \iff \begin{cases} \dfrac{\partial \varphi}{\partial m} &= 0 \\ \dfrac{\partial \varphi}{\partial q} &= 0 \end{cases}$$

Le due condizioni si traducono in

$$\begin{cases} \displaystyle\sum_{i=1}^{N} 2\,[y_i - mx_i - q] \cdot (-x_i) &= 0 \\ \displaystyle\sum_{i=1}^{N} 2\,[y_i - mx_i - q] \cdot (-1) &= 0 \end{cases}$$

che riscritte dopo aver diviso tutto per 2 diventano

$$\begin{cases} m \displaystyle\sum_{i=1}^{N} x_i^2 + q \sum_{i=1}^{N} x_i &= \displaystyle\sum_{i=1}^{N} x_i y_i \\ m \displaystyle\sum_{i=1}^{N} x_i + qN &= \displaystyle\sum_{i=1}^{N} y_i \end{cases}$$

la cui unica soluzione è

$$\begin{cases} m &= \dfrac{N \displaystyle\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N \displaystyle\sum_{i=1}^{N} x_i^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \\ q &= \dfrac{\displaystyle\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{N \displaystyle\sum_{i=1}^{N} x_i^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \end{cases}$$

In figura 5 sono riportati i dati della tabella 1 relativi al numero di scarpe ed altezza degli studenti.

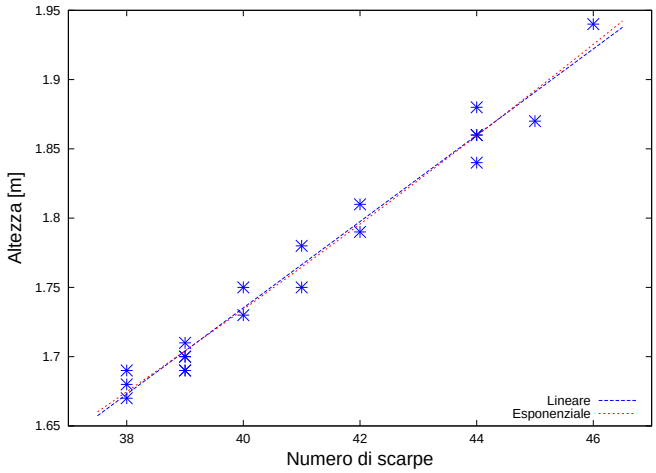



Figura 5: Coppie $(x_i, y_i)$ di numero di scarpe ed altezza degli studenti di una classe (punti con asterischi). Retta dei minimi quadrati $y = mx + q$ (linea blu) con $m = 0.03115$ e $q = 0.4892$; interpolazione esponenziale $y = a \cdot b^x$ (linea rossa) con $a = 0.86324$ e $b = 1.0176$

Applicando le formule precedentemente ricavate si ottengono $m$ e $q$, da quali la funzione lineare che descrive l'altezza in funzione del numero di scarpe

$$\begin{cases} m = 0.03115 \\ q = 0.4892 \end{cases} \qquad \Longrightarrow \qquad y = 0.03115x + 0.4892.$$

Si osservi che la retta ottenuta, detta *retta dei minimi quadrati*, è relativa ai dati a disposizione: aggiungendo o togliendo dei dati, $m$ e $q$, in generale, cambierebbero.

**Interpolazione esponenziale**

L'interpolazione lineare è certamente la più semplice, tuttavia i dati potrebbero essere distribuiti in altro modo, ad esempio come una parabola del tipo $y = ax^2 + bx + c$ o come una funzione esponenziale del tipo $y = a \cdot b^x$. Il metodo per determinare la *miglior parabola* segue gli stessi passi di quello visto per la miglior retta, quindi evitiamo i dettagli. Ci concentriamo, invece, sul determinare la *miglior funzione esponenziale*

$$y = a \cdot b^x \quad \text{con} \quad a > 0,\ b > 0,\ b \neq 1.$$

Il problema si semplifica di molto se, anziché considerare la funzione $y = a \cdot b^x$ prendiamo il suo logaritmo, per esempio in base 10: questo è possibile perché sia $a$ che $b$ sono positivi, per cui lo è anche $y$,

$$\operatorname{Log} y = \operatorname{Log} a \cdot b^x = \operatorname{Log} a + \operatorname{Log} b^x = \operatorname{Log} a + x \operatorname{Log} b,$$

ossia, se indichiamo con

$$z = \operatorname{Log} y, \quad m = \operatorname{Log} b, \quad q = \operatorname{Log} a,$$

si ha

$$z = mx + q,$$

che è lo stesso problema visto in precedenza. Pertanto, i coefficienti $m$ e $q$ si determinano dalle relazioni

$$\begin{cases} m = \dfrac{N \displaystyle\sum_{i=1}^{N} x_i \operatorname{Log} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} \operatorname{Log} y_i}{N \displaystyle\sum_{i=1}^{N} x_i^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \\ q = \dfrac{\displaystyle\sum_{i=1}^{N} \operatorname{Log} y_i \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i \operatorname{Log} y_i}{N \displaystyle\sum_{i=1}^{N} x_i^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \end{cases}$$

da cui si ottengono i coefficienti $a$ e $b$ dell'interpolazione esponenziale

$$a = 10^q \quad \text{e} \quad b = 10^m.$$

Per i dati della figura 5 si ha

$$\begin{cases} a = 0.86324 \\ b = 1.0176 \end{cases} \qquad \Longrightarrow \qquad y = 0.86324 \cdot 1.0176^x.$$

Dalla figura 5 è piuttosto difficile distinguere l'approssimazione esponenziale da quella lineare. Un indice che permette di valutare la bontà o meno della funzione interpolante $f(x)$ è una riscalatura della funzione $\varphi$ che si era minimizzata, ossia il coeffiente

$$E = \sqrt{\frac{1}{N} \sum_{i=1}^{N} [y_i - \hat{y}_i]^2} \quad \text{con} \quad \hat{y}_i = f(x_i).$$

Nel caso dei dati in figura 5, per il caso lineare e per il caso esponenziale si ottengono

$$E_{\text{lineare}} = 0.012818 \quad \text{e} \quad E_{\text{esponenziale}} = 0.012677,$$

da cui si evince che l'interpolazione esponenziale è leggermente migliore.

# 6 Esercizi

1. Una classe di studenti ha riportato, nell'ultima verifica di matematica, i seguenti voti: 2, 6, 3, 4, 8, 3, 10, 9, 4, 3, 7, 7, 10, 5, 2, 9, 3, 6, 3, 8. Determinare: (a) la media aritmetica; (b) la media geometrica; (c) la media armonica; (d) la media quadratica; (e) la moda; (f) la mediana; (g) lo scarto quadratico medio; (h) la varianza.

2. Uno studente, durante l'anno scolastico, ha ottenuto, nelle veriche scritte di matematica, i seguenti voti: 5.5, 7, 6, 4.5, 6, 5.5, mentre nelle veriche orali ha ottenuto i voti: 6, 7, 5.5, 6.6, 7, 8. Sapendo che l'insegnante assegna un peso doppio alle verifiche scritte rispetto a quelle orali, determinare: (a) la media aritmetica ponderata; (b) la media geometrica ponderata; (c) la media armonica ponderata; (d) la media quadratica ponderata.

3. Una classe di 20 studenti ha riportato, nell'ultima verifica di matematica, i seguenti voti: 9, 3, 7, 3.5, 9.5, 7.5, 2.5, 6.5, 10, 10, 3, 5, 7.5, 5.5, 4, 3.5, 4, 8.5, 2.5, 2.

(a) Suddividere i dati nelle seguenti 6 classi: $G$ (gravemente insufficiente, voto $< 5$), $I$ (insufficiente, voto $\in [5;6)$), $S$ (sufficiente, voto $\in [6;7)$), $D$ (dicreto, voto $\in [7;8)$), $B$ (buono, voto $\in [8;9)$), $E$ (eccellente, voto $\geq 9$).
(b) Costruire la tabella delle frequenze assolute, relative e percentuali per i dati suddivisi in classi.
(c) Riportare in un istogramma le frequenze assolute per i dati suddivisi in classi.
(d) Costruire la curva di Lorenz e calcolare il rapporto di concentrazione $R$.

4. In un esperimento si è fatta variare la forza applicata ad una molla misurando il relativo allungamento ottenendo una serie di coppie di dati $(F_i; \Delta x_i)$, dove le forze e gli allungamenti sono di seguito riportati: $F_i = 15, 41, 27, 16, 47, 47, 50, 27, 26, 13$ e $\Delta x_i = 0.28, 0.86, 0.50, 0.34, 0.98, 0.94, 0.99, 0.57, 0.54, 0.31$.
(a) Determinare i coefficienti $m$ e $q$ della miglior retta interpolante nel senso dei minimi quadrati $y = mx+q$.
(b) Determinare i coefficienti $a$ e $b$ della miglior approssimazione (nel senso dei minimi quadrati) della funzione esponenziale $y = a \cdot b^x$.
(c) Determinare il coeffiente

$$E = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{y}_i]^2} \quad \text{con} \quad \hat{y}_i = f(x_i)$$

nel caso di interpolazione lineare ed esponenziale.

**Soluzione/risoluzione degli esercizi**

1. (a) Media aritmetica: 5.6.
(b) Media geometrica: 4.93.
(c) Media armonica: 4.29.
(d) Media quadratica: 6.2
(e) Moda: 3
(f) Mediana: 5.5
(g) Scarto quadratico medio: 2.67.
(h) Varianza: 7.14.

2. (a) 6.06; (b) 6.00; (c) 5.93; (d) 6.13.

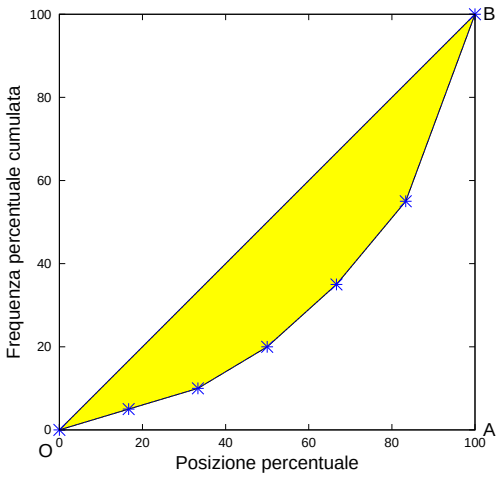
3. (a) e (b) Dividendo in classi si ottiene la seguente tabella

| Classe | Frequenza assoluta | Frequenza relativa | Frequenza percentuale |
|---|---|---|---|
| G | 9 | 0.45 | 45% |
| I | 2 | 0.10 | 10% |
| S | 1 | 0.05 | 5% |
| D | 3 | 0.15 | 15% |
| B | 1 | 0.05 | 5% |
| E | 4 | 0.20 | 20% |
| Totali: | 20 | 1 | 100% |

(c) Solito istogramma.
(d) Ordinando i dati in ordine di frequenza percentuale crescente si ha

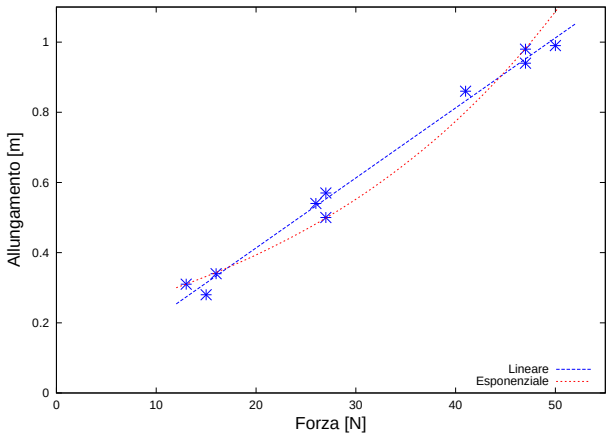
| | Classe | Posizione percentuale | Frequenza percentuale | Frequenza cumulata |
|---|---|---|---|---|
| | | $p_i$ | | $q_i$ |
| 1 | S | 16.67% | 5% | 5% |
| 2 | B | 33.33% | 5% | 10% |
| 3 | I | 50.00% | 10% | 20% |
| 4 | D | 66.67% | 15% | 35% |
| 5 | E | 83.33% | 20% | 55% |
| 6 | G | 100.00% | 45% | 100% |

da cui la figura sottostante




che fornisce $R = 0.4167$.

4. I dati e le due interpolazioni sono riportati nella seguente figura




(a) $y = 0.01996 * x + 0.014297$.
(b) $y = 0.19979 \cdot (1.0345)^x$.
(c) $E_{\text{lineare}} = 0.027941$, $E_{\text{esponenziale}} = 0.050754$.