Interpretare la realtà

Statistica descrittiva di base

A cura di
Silvia Monica
Paola Pezzani

Ringraziamenti

Un ringraziamento va rivolto alle classi seconde A e B e alle terze A, B e C per il contributo apportato in vista della realizzazione dell'indagine statistica svolta presso il Liceo Attilio Bertolucci di Parma nel corso dell'anno scolastico 2014-15 e a tutti gli studenti per la proficua collaborazione prestata fornendo risposte sulle quali si basano diverse sezioni ed esempi della dispensa.

© Liceo Attilio Bertolucci Editore

ISBN 9788898952106

Editato in Parma, aprile 2018

Tavola dei contributi

Coordinamento redazionale, revisione dei contenuti: Silvia Monica, Paola Pezzani

Composizione e progetto grafico: Luca Cantoni, Silvia Monica

Curatori delle figure: Alessia Alinovi, Massimo Buzzi

Autori

Alessia Alinovi
Massimo Buzzi
Luca Cantoni
Lucrezia Dolfi
Luisa Garripoli
Federico Mori
Davide Petrolini

Indice

Premessa

Il presente e-Book è stato realizzato da un team composto da studenti e docenti di diverse classi del Liceo Scientifico Attilio Bertolucci di Parma, nell'ambito di un progetto didattico di durata biennale.

Il progetto è stato ideato allo scopo di consentire agli studenti l'apprendimento collaborativo ed autentico del *metodo statistico*, svolgendo un'indagine sull'impiego dei device e dei dispositivi elettronici di ogni tipo, nell'uso scolastico e personale, all'interno di una scuola 2.0. Molti esempi del presente e-Book si riferiscono all'indagine condotta a scuola.

Il progetto si è articolato nelle seguenti fasi (tra parentesi gli autori):

- individuazione dell'oggetto dell'analisi statistica (docenti);
- stesura del progetto con definizione di compiti, tempi e attività (docenti);
- elaborazione questionario per l'indagine statistica e interviste (team);
- elaborazione e analisi dati (team);
- creazione della dispensa (team).

La collaborazione tra docenti e studenti e tra quest'ultimi è stata realizzata mediante la creazione con *Google Classroom* di un'apposita classe 'virtuale' del team. Tale strumento ha permesso di condividere documenti, attività, scadenze, revisioni, ...
L'indagine statistica ha coinvolto 129 alunni appartenenti a otto classi dalla prima alla quinta. Le interviste sono state effettuate personalizzando un modulo online di *Moduli di Google*, predisposto dagli alunni delle classi IIA, IIB, IIIA, IIIB, IIIC. Il campione statistico corrisponde a circa il 22% della popolazione scolastica del Liceo Scientifico Attilio Bertolucci di Parma degli anni 2015/16.

I materiali teorici di questa dispensa sono stati tratti dalle fonti citate in bibliografia e sitografia.

Prerequisiti di carattere teorico per la lettura di questa dispensa. Il lettore deve conoscere la terminologia statistica di base: carattere, modalità, classificazione dei caratteri, campione e metodo di campionamento, indagine, tipologie di grafici, frequenza relativa, assoluta e percentuale, anche cumulata.
E' richiesta la capacità di costruire, leggere ed interpretare tabelle di frequenza e grafici e di utilizzare metodi di campionamento.

Introduzione

Nel presente e-Book sono trattati gli argomenti di statistica descrittiva utili allo studio dei dati ricavati da un'indagine statistica. La dispensa è suddivisa in due parti.

La prima parte contiene lo studio degli indici di posizione e di variabilità riferiti a un carattere, la cosiddetta statistica **univariata**. La seconda parte riguarda lo studio di statistiche che elaborano più dati, in particolare due serie di dati considerate contemporaneamente e le loro relazioni, detta statistica **bivariata**.

I Gli indici

Nel momento in cui si raccolgono i dati di una statistica essi si raggruppano secondo le diverse modalità che il carattere assume.
Questa rappresentazione prende nome di distribuzione: quindi una **distribuzione statistica** è costituita dall'insieme di tutti i valori ottenuti nell'indagine.

Distribuzione statistica

Esempio 0.0.1

Nell'indagine svolta presso il liceo, molte domande hanno come risposta il gradimento, espresso con numeri da 1 a 5. Quindi il carattere studiato assume le modalità da 1 a 5. La distribuzione statistica quindi è l'insieme dei valori

$$\{1, 2, 3, 4, 5\}.$$

Ciascun valore viene ottenuto tante volte durante l'indagine, perché molte persone diverse possono dare la stessa risposta.

Due distribuzioni di dati sono **omogenee** quando le unità della statistica, rispetto al carattere dato, mostrano le stesse modalità.

Terminata la raccolta dei dati in una distribuzione statistica, si costruisce una tabella, detta tabella di frequenza, che consente di comprendere come le frequenze assolute o relative sono distribuite tra le varie classi di misura. La tabella può essere utile a realizzare dei grafici.

Nel contempo è opportuno procedere con l'analisi e lo studio dei dati stessi al fine di ricavarne informazioni utili, confrontandoli eventualmente con altre distribuzioni simili: per esempio, si vogliono studiare e confrontare i voti conseguiti all'esame di maturità degli alunni di due differenti licei.

Gli indici sono i valori che permettono una sintesi e uno studio dei dati. Essi si suddividono approssimativamente in due categorie: gli indici di posizione e gli indici di variabilità. Fra gli indici di posizione si hanno ad esempio la media aritmetica, geometrica, armonica e quadratica, le relative medie ponderata, la moda e la mediana.

1 Indici di posizione

Gli indici di posizione consentono di sintetizzare in un singolo valore numerico informazioni rilevanti riguardo un'intera distribuzione.
Come il nome suggerisce, aiutano a capire verso quale *posizione* tendono tutti i valori. Possiamo considerare un indice di posizione come *baricentro* della distribuzione data.
Gli indici di posizione presi in considerazione in questa dispensa sono le varie medie.

1.1 Media aritmetica

Si chiama media aritmetica di n valori $x_1, x_2, \ldots, x_n$ il numero:

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}, \tag{1.1.1}$$

che si può scrivere in forma compatta attraverso il simbolo di sommatoria:

Media aritmetica

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Esempio 1.1.1

La spesa settimanale (espressa in €) di quattro famiglie per i generi alimentari è riassunta nella tabella sottostante.

Famiglia 1	Famiglia 2	Famiglia 3	Famiglia 4
250	270	230	720

Calcoliamo la spesa media settimanale delle famiglie.

$$\mu = \frac{250 + 270 + 230 + 720}{4} = 367.5.$$

La media dell'esempio 2, è pari a 367.5*e* e non sintetizza in modo efficace i dati per la presenza del valore 720*e* che si discosta notevolmente dagli altri valori della serie. In questo caso la media aritmetica non è un buon indice per indicare la spesa media delle quattro famiglie prese in considerazione.

La media aritmetica è significativa soprattutto se i valori della distribuzione sono diffusi in modo bilanciato e non presentano grandi variazioni gli uni dagli altri. Non è un buon indice dei dati se sono presenti valori estremi, evidentemente anomali. In questo caso si possono usare come medie la mediana o la moda.

Per comprendere questo aspetto ancora meglio si considerino i voti ottenuti da uno studente in inglese: 6, 7, 8. La media, 7, corrisponde al voto che avrebbe preso se tutti e tre i voti presi fossero stati uguali.

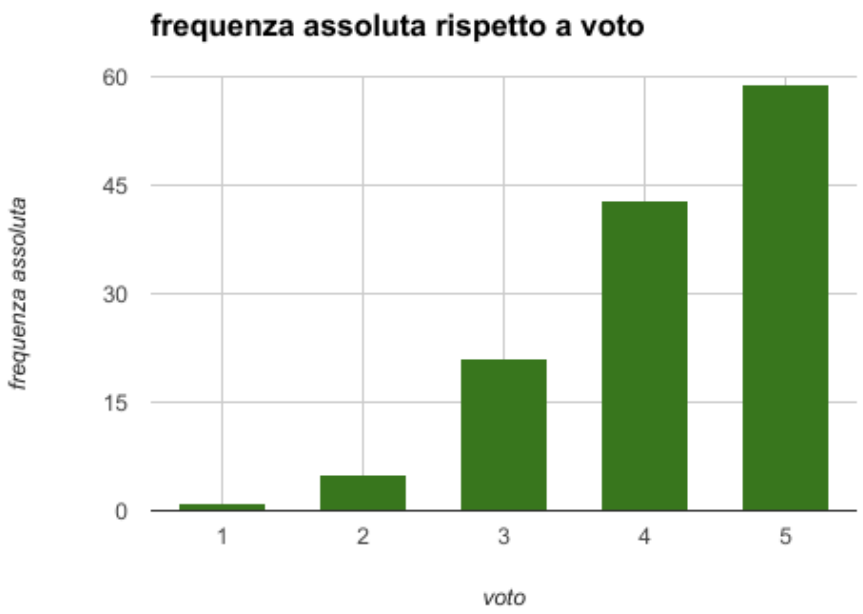
Quando usare la media aritmetica

In conclusione, *la media aritmetica viene usata tutte le volte che ha senso addizionare i dati* ed è il valore che, sostituito ai termini della somma, lascia invariato il risultato. Infatti per i dati quantitativi, la media M secondo Oscar Chisini è quel valore che, se sostituito ad ogni dato x, rende vera l'uguaglianza $f(x_1, x_2, \ldots, x_n) = f(M, M, \ldots, M_n)$, essendo f una funzione scelta dal matematico. Nel caso della media aritmetica questa funzione è la somma.

Esempio 1.1.2

L'indagine portata avanti all'interno del Liceo Attilio Bertolucci comprende la domanda: "Ritieni che utilizzare strumenti tecnologici fin da ora ti possa aiutare anche in futuro in campo lavorativo?".
Alla seguente domanda i 129 studenti intervistati dovevano rispondere con un voto da 1 (per niente) a 5 (assolutamente sì).
I risultati ottenuti sono sintetizzati nella seguente tabella.

voto	1	2	3	4	5	totale
frequenza assoluta	1	5	21	43	59	129



In questo caso la formula (1.1.1) ci permette di calcolare la media aritmetica dei valori ricavati, che risulta 4,186 (~ 4).
Questo significa che gli studenti a cui è stato sottoposto il questionario tendono a credere che gli strumenti tecnologici potranno aiutare molto in futuro: la media aritmetica risulta infatti essere compresa fra i due valori più alti (4 e 5).

L'utilizzo di una tabella di frequenza, inoltre, rende molto più semplice ed immediato il calcolo della media, infatti la formula precedente può essere riscritta come:

$$\mu = \frac{1 \cdot f_1 + 2 \cdot f_2 + \cdots + 5 \cdot f_5}{129}.$$

Maggiore è la frequenza di una classe di misura, maggiore è il suo' 'peso' nella costruzione della media.

1.1.1 Proprietà fondamentali della media aritmetica

Vengono qui riportate due delle proprietà fondamentali della media aritmetica.

Proprietà 1.1.1. La somma degli scarti dalla media è nulla:

$$(x_1 - \mu) + (x_2 - \mu) + \cdots + (x_n - \mu) = \sum_{i=1}^{n} (x_i - \mu) = 0.$$

Proprietà 1.1.2. Comunque si scelga un numero $c \in \mathbb{R}$, la somma dei quadrati degli scarti dalla media è sempre minore o uguale alla somma dei quadrati degli scarti da un qualunque numero c. La media aritmetica è il valore che rende minima la somma dei quadrati degli scarti.

$$\sum_{i=1}^{n} (x_i - \mu)^2 \leq \sum_{i=1}^{n} (x_i - c)^2, \ \forall c \in \mathbb{R}.$$

Proprietà 1.1.3. Se tutti i termini di una serie subiscono un incremento (o un decremento) $b \in \mathbb{R}$, allora la loro media aritmetica risulta aumentata (o diminuita) dello stesso incremento.

Dati $x_1, \ldots, x_i$ e i relativi termini incrementati $y_1, \ldots, y_i$ tali che $y_i = x_i + b$ si verifica

$$\mu_Y = \mu_X + b.$$

Proprietà 1.1.4. Se tutti i termini di una serie vengono moltiplicati per lo stesso numero $a \in \mathbb{R}$, allora la loro media aritmetica risulta moltiplicata per lo stesso incremento.

Dati $x_1, \ldots, x_i$ e i relativi termini maggiorati $y_1, \ldots, y_i$ tali che $y_i = ax_i$ si verifica

$$\mu_Y = a\mu_X.$$

1.1.2 Media aritmetica ponderata

A volte si attribuisce ad ogni valore $(x_1, x_2, \ldots, x_n)$ ottenuto nella distribuzione un corrispondente peso $(p_1, p_2, \ldots, p_n)$ che rappresenta numericamente l'importanza attribuita ai valori stessi; in tal caso si definisce **media ponderata** il valore:

$$\mu = \frac{x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n}{p_1 + p_2 + \cdots + p_n}. \tag{1.1.2}$$

Esempio 1.1.3

All'interno dell'indagine statistica compaiono le domande:

- "Quanto ritieni importante l'utilizzo dei device in ambito scolastico in una scala da 1 (per niente importante) a 5 (fondamentale)?"
- "Ritieni che utilizzare strumenti tecnologici fin da ora ti possa aiutare

anche in futuro in campo lavorativo?"

Consideriamo uno studente che ha risposto 2 (poco importante) alla prima domanda e 4 (importante) alla seconda.
Se riteniamo la seconda domanda di importanza doppia rispetto alla prima, la media aritmetica ponderata è:

$$\mu = \frac{2 \cdot 1 + 4 \cdot 2}{1 + 2} = 3.33$$

Il risultato 3.33 propende, anche se di poco, verso la maggior importanza e lo possiamo interpretare come una complessiva utilità derivante dall'uso dei device.
Se riteniamo le due domande ugualmente importanti la media è:

$$\mu = \frac{2 + 4}{2} = 3$$

Il risultato 3, ovvero "indifferente", non valorizza l'eventuale scelta dell'utilizzo dei device in ambito scolastico.

Quando usare la media ponderata

In generale la media aritmetica non ponderata considera i dati tutti allo stesso livello, e con ciò non attribuisce un diverso peso agli stessi; al contrario la media ponderata è utile quando i dati a disposizione non hanno un'uguale importanza e vanno, pertanto, *pesati* diversamente tra loro. In tal caso il peso è stabilito da chi analizza i dati, tenendo anche conto del significato delle diverse domande nel complesso dell'intera indagine.

La media ponderata si utilizza ad esempio per calcolare la media dei voti degli esami universitari con crediti differenti.

Esempio 1.1.4

Se uno studente iscritto al primo anno del corso di laurea in matematica ha superato i seguenti esami riportando le votazioni indicate

Esame	Punteggio in trentesimi	Crediti
Laboratorio di Matematica	25	9
Analisi Matematica	24	12
Geometria	21	6
Algebra	27	6
Calcolo delle probabilità	23	9
Fisica generale	24	9
Lingua inglese	30	3
Fondamenti di Informatica	28	3
Abilità relazionali	30	3

La media ponderata è

$$\mu = \frac{25 \cdot 9 + 24 \cdot 12 + 21 \cdot 6 + 27 \cdot 6 + 23 \cdot 9 + 24 \cdot 9 + 30 \cdot 3 + 28 \cdot 3 + 30 \cdot 3}{9 + 12 + 6 + 6 + 9 + 9 + 3 + 3 + 3} = 24.8.$$

La media aritmetica calcolata con le frequenze assolute è anch'essa una media pesata (vedi l'esempio 1.1.3). In quest'ultimo caso l'uso della media ponderata non è una necessità, ma un'opportunità di calcolo qualora sia presente una considerevole quantità di dati.

Esempio 1.1.5

Un alunno ha ottenuto come voti 6, 6, 7, 8 in latino. La media può essere calcolata in due modi: $\frac{6+6+7+8}{4}$ oppure considerando le frequenze $\frac{6 \cdot 2 + 7 + 8}{4}$, evitando di ripetere il numero 6 due volte.

1.1.3 Le differenze o scarti dal valore medio

Le differenze:

$$x_1 - \mu, x_2 - \mu, \ldots, x_n - \mu$$

tra i singoli valori $x_1, x_2, \ldots, x_n$ e il loro valore medio μ, si chiamano scarti della serie di valori dalla media. Questi valori possono essere calcolati con qualunque tipo di indice di posizione.

Ulteriori informazioni sugli scarti si trovano nel successivo capitolo 2.2 sugli indici di variabilità.

1.2 Media geometrica

Si chiama media geometrica degli n numeri positivi $x_1, x_2, \ldots, x_n$ il numero positivo:

$$\mu_g = (x_1 x_2 \ldots x_n)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \ldots x_n}. \tag{1.2.1}$$

Media geometrica

La media geometrica viene utilizzata tutte le volte che deve rimanere invariato il prodotto dei valori.

In questo caso, la media geometrica μ_g secondo Oscar Chisini è quel valore tale per cui:

$$(x_1 x_2 \ldots x_n) = (\underbrace{\mu_g \mu_g \ldots \mu_g}_{n \text{ volte}}) = \mu_g^n.$$

Esempio 1.2.1

Supponiamo di voler stimare di quanto varia la nostra energia potenziale ($U = mgh$) salendo in cima ad una montagna di media altitudine.
Nel calcolo di una stima conviene utilizzare cifre intere, per l'esattezza potenze di 10. Ipotizziamo quindi che l'accelerazione di gravità g sia $10\,\mathrm{m/s^2}$ e che la nostra massa m sia $100\,\mathrm{kg}$.

Stimiamo poi l'altezza h di una montagna mediante la media geometrica, perché vogliamo che il valore stimato disti dello stesso fattore dai due estremi dell'intervallo considerato: una montagna media è più alta di un grattacielo ($3 \times 10^2\,\mathrm{m}$) e più bassa dell'Everest ($1 \times 10^4\,\mathrm{m}$), quindi

$$h = \mu_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt{(3 \times 10^2\,\mathrm{m})(1 \times 10^4\,\mathrm{m})} = 2 \times 10^3\,\mathrm{m}.$$

Sostituendo la massa, l'accelerazione di gravità e l'altezza media stimata, possiamo ricavare la grandezza da noi cercata. Il resto del calcolo è lasciato al lettore.
Esempio tratto da: L. Weinstein, J.A. Adam, *Più o meno quanto?*, Zanichelli.
Si può notare che il valore ottenuto con questo metodo ($2 \times 10^3\,\mathrm{m}$) è un valore decisamente ragionevole come stima, pensiamo ad esempio che le cime del nostro Appennino si aggirano all'incirca attorno a tale valore.

1.2.1 Proprietà della media geometrica

La media geometrica gode di alcune proprietà notevoli.

Proprietà 1.2.1. Il logaritmo della media geometrica è la media aritmetica dei logaritmi dei valori della serie di cui si calcola la media geometrica.

$$\log \mu_g = \log \left(x_1 x_2 \dots x_n\right)^{\frac{1}{n}} = \frac{\log x_1 + \log x_2 + \dots + \log x_n}{n}.$$

Proprietà 1.2.2. La media geometria delle potenze p-esime dei valori della serie è uguale alla potenza p-esima della media geometrica dei valori della serie.

$$\left(x_1^p x_2^p \dots x_n^p\right)^{\frac{1}{n}} = \left(x_1 x_2 \dots x_n\right)^{\frac{p}{n}} = \left(\left(x_1 x_2 \dots x_n\right)^{\frac{1}{n}}\right)^p = \left(\mu_g\right)^p.$$

Proprietà 1.2.3. La media geometrica è sempre minore o uguale della media aritmetica.

1.3 Media armonica

Si chiama media armonica degli n numeri positivi $x_1, x_2, \dots, x_n$ il numero positivo:

$$\mu_a = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}. \tag{1.3.1}$$

Media armonica

si può notare che la media armonica è il reciproco della media aritmetica dei reciproci degli n valori dati:

$$\frac{1}{\mu_a} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n}.$$

quando usare la media armonica

La media armonica è usata quando deve rimanere invariata la somma dei reciproci dei valori ed è importante notare che si può utilizzare *solo quando nessuno dei valore della serie è nullo*. La media armonica è utile soprattutto quando i valori della serie sono reciproci di altri dati ad essi collegati e rilevanti per la comprensione del fenomeno, oppure sono inversamente proporzionali ad altri dati o informazioni. Ad esempio, si usa la media armonica se i dati della serie sono velocità (e quindi inversamente proporzionali al tempo impiegato).
Anche questa media, quindi si può definire con la regola di Chisini e la grandezza da mantenere invariata è proprio la somma dei reciproci dei valori della serie di dati.

Si definiscono **scarti relativi** i valori *Scarti relativi*

$$\frac{x_i - \mu}{x_i},$$

dove μ è un indice di posizione. Lo scarto relativo quindi è il rapporto fra lo scarto e il valore x_i stesso e ad ogni valore della statistica è associabile il proprio scarto relativo.
Con questa definizione gli scarti relativi dalla media armonica sono i valori

$$\frac{x_i - \mu_a}{x_i}.$$

Esempio 1.3.1

Per chiarire il concetto, si pensi ad una statistica costituita da due soli valori: ad esempio i due voti ottenuti in una disciplina in due differenti verifiche: 9 e 7.5. La media armonica dei due valori è:

$$\mu_a = \frac{2 \cdot 9 \cdot 7.5}{9 + 7.5} = \frac{90}{11} \approx 8.18$$

e i due scarti relativi sono rispettivamente

$$\frac{x_1 - \mu_a}{x_1} \text{ e } \frac{x_2 - \mu_a}{x_2} \text{ cioè } \frac{9 - \frac{90}{11}}{9} \approx 0.09 \text{ e } \frac{7.5 - \frac{90}{11}}{7.5} \approx -0.09.$$

La somma dei due scarti relativi dalla media armonica è nulla.

Quanto visto nell'esempio precedente non è un caso particolare, bensì una regola generale. Si tratta di una proprietà della media armonica, vale sempre e comunque tutte le volte che si utilizza la media armonica. Si tratta quindi di una proprietà della media armonica e si può dimostrare (la dimostrazione per il caso di due soli valori è lasciata al lettore come esercizio).

Proprietà 1.3.1. La somma degli scarti relativi dei singoli valori dalla media armonica è nulla:

$$\frac{x_1 - \mu_a}{x_1} + \frac{x_2 - \mu_a}{x_2} + \frac{x_3 - \mu_A}{x_3} + \cdots + \frac{x_n - \mu_a}{x_n} = \sum_{i=1}^{n} \frac{x_i - \mu_a}{x_i} = 0.$$

Proprietà 1.3.2. La media armonica è invariante per conversioni dell'unità di misura. Significa che se si effettuano conversioni, si otterrà come media armonica la conversione della media di partenza.

Proprietà 1.3.3. La media armonica è sempre minore o uguale di quella geometrica, $\mu_a \leq \mu_g$, e complessivamente

$$\mu_a \leq \mu_g \leq \mu.$$

1.4 Media quadratica

Si chiama media quadratica degli n numeri positivi $x_1, x_2, \ldots, x_n$ il numero positivo:

$$\mu_q = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}. \tag{1.4.1}$$

La media quadratica viene utilizzata quando deve rimanere invariata la somma dei quadrati dei valori. Facendo riferimento alla definizione secondo Chisini, la funzione f della media quadratica è la somma dei quadrati dei numeri:

$$x_1^2 + x_2^2 + \cdots + x_n^2 = \underbrace{\mu_q^2 + \mu_q^2 + \cdots + \mu_q^2}_{n \text{ volte}} = n\mu_q^2.$$

Esempio 1.4.1

Secondo la teoria cinetica dei gas, un gas è costituito da un elevato numero di particelle che si muovono in modo casuale. La formula che descrive l'energia cinetica media di una molecola del gas è:

$$\overline{K} = \frac{1}{N}\sum_{i=1}^{n} \frac{1}{2} m v_i^2 = \frac{1}{2} m \overline{v}^2,$$

dove N è il numero delle particelle, v_i è la velocità delle differenti particelle e

$$\overline{v} = \mu_q = \sqrt{\frac{v_1^2 + v_2^2 + \cdots + v_n^2}{n}}$$

è la velocità quadratica media delle particelle.
Come si evince dalla formula, l'energia cinetica risulta una quantità positiva poiché è il prodotto di due fattori positivi, una velocità media al quadrato e una massa. L'utilizzo della velocità quadratica media è preferibile a quella della velocità aritmetica media. Infatti quest'ultima avrebbe verosimilmente un valore nullo dato che le particelle si muovono in ogni possibile verso e direzione.

La media quadratica è utilizzata in tantissimi contesti.
Nella sezione sugli indici di variabilità, si vedrà come la media quadratica venga utilizzata per descrivere la statistica degli scarti dalla media (capitolo 2.3).

Proprietà 1.4.1. La media quadratica è maggiore o uguale dell'aritmetica

$$\mu \leq \mu_q.$$

1.4.1 Considerazioni conclusive sulle medie definite secondo Chisini

Quando si vogliono costruire indici che utilizzano tutti i valori della distribuzione considerata si sceglie generalmente fra media aritmetica, geometrica, armonica e quadratica. Qualora non si vogliono utilizzare i dati "estremi" della distribuzione si possono impiegare la moda o la mediana (spiegate nei paragrafi successivi).
In alcuni testi si trova una netta distinzione fra questi due tipi di indice. In generale complessivamente vengono chiamati indici centrali, perché si riferiscono al valore che più o meno rappresenta tutti i dati della serie considerati nel loro complesso e fra questi si distinguono:

- le medie di calcolo, che vengono calcolate con la definizione secondo Chisini e quindi alle quali è sempre associabile una funzione che si mantiene costante sostituendo ogni valore della serie con la media (rientrano in questa categoria le medie aritmetica, armonica, quadratica, ecc.)

- le medie di posizione propriamente dette, che si riferiscono unicamente alla posizione che i valori occupano una volta messi in ordine in qualche modo (ad esempio moda e mediana). In questa dispensa però si utilizza il termine "indice di posizione" in tutti i casi.

Relazione fra le varie medie

Fra le medie esiste una gerarchia che ne stabilisce l'ordine secondo uno schema prestabilito. Senza approfondire questo aspetto, comunque, affermiamo che qualunque sia la statistica che si sta analizzando vale sempre

$$\mu_a \leq \mu_g \leq \mu \leq \mu_q.$$

1.5 Moda

Si chiama moda degli n elementi $x_1, x_2, \ldots, x_n$, l'elemento (o gli elementi) che si presenta più frequentemente. Si conta la frequenza con cui si presentano i valori della serie e si sceglie quello con la maggiore frequenza: è la moda. La moda è

quindi *sempre* uno dei valori della serie ed è significativa solo se esiste il massimo delle frequenze.

Esempio 1.5.1

Prendiamo in considerazione il quesito che riguarda il numero di device posseduti da ogni alunno intervistato. Le risposte ottenute sono riportate qui sotto.

numero di device	1	2	3	4	5	> 5	totale
frequenza assoluta	1	14	43	40	16	15	129
frequenza percentuale	0%	11%	33%	31%	13%	12%	100%

La moda è il valore 3 (presente 43 volte su 129) e ciò dimostra che la maggior parte degli studenti sottoposti al questionario possiede 3 device.

1.6 Mediana

Mediana

Si chiama mediana di una successione ordinata di n numeri, quel valore x_m tale che i numeri della successione minori di x_m sono tanti quanti quelli maggiori di x_m. Nel caso di dati numerici, la mediana x_m è il valore preso in modo tale che i numeri che la precedono sono tanti quanti quelli che la seguono. Quindi, dopo aver ordinato i dati (crescente o decrescente):

- se n è dispari la mediana è il valore centrale della serie di dati;
- se n è pari la mediana è la media aritmetica tra i due valori della serie di dati.

Esempio 1.6.1

L'indagine statistica svolta nel Liceo A. Bertolucci comprendeva la domanda: 'Qual è stata la valutazione media ottenuta nelle materie scientifiche nel primo periodo?' L'elenco delle risposte fornite dai 129 ragazzi intervistati è:

voto approssimato	5	5.5	6	6.5	7	7.5	8	8.5	9	10	tot
frequenza assoluta	8	6	39	3	48	2	20	1	4	1	129

Utilizzando una tabella di frequenze, appare chiaro che la moda dei valori sia 7, che in questo caso coincide con la mediana.

La media aritmetica risulta invece essere 6,793 (~ 6.8), un valore comunque molto vicino a quelli della moda e della mediana.

La mediana è calcolabile mettendo in ordine crescente tutti i valori (con l'aiuto di un apposito programma) e scegliendo quello centrale, quindi operativamente:

- se gli n elementi sono dispari, la mediana corrisponde al valore che occupa la posizione $(n+1)/2$.
- se gli n elementi sono pari (come in questo caso) la mediana corrisponde alla media aritmetica tra i due valori nelle posizioni $n/2$ e $(n+1)/2$.

1.6.1 Frequenza cumulata e mediana

La frequenza cumulata è la somma delle frequenze delle modalità inferiori e uguali di una data modalità. La mediana corrisponde al valore con frequenza del 50% all'interno della distribuzione di frequenza, cioè quel valore che ha il 50% dei casi prima e il 50% di quelli dopo. Particolare attenzione bisogna porre quando calcolando le frequenze cumulate si arriva a superare il 50%, ad esempio quando si passasse dal 45% al 60%. In questi casi infatti il valore corrispondente rientra in questo intervallo.

Relazione fra frequenza cumulata e mediana

1.7 Casi particolari

In una distribuzione simmetrica, media, mediana e moda coincidono ed è possibile individuare un asse che suddivida in due parti speculari la distribuzione.
In una distribuzione asimmetrica, invece, la media si posiziona nella direzione dell'asimmetria.

Se la media supera la mediana si parla di asimmetria positiva (obliqua a destra), mentre se la mediana supera la media si parla di asimmetria negativa (obliqua a sinistra).

Esempio 1.7.1

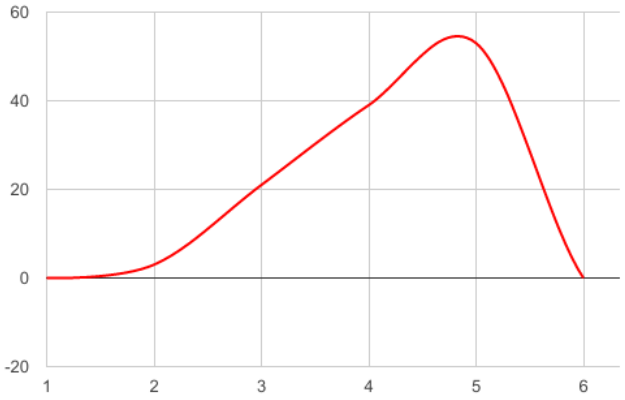
La tabella nel paragrafo precedente (riferita alla domanda "Qual è stata la valutazione media ottenuta nelle materie scientifiche nel primo periodo?") presenta moda, media e mediana quasi coincidenti e pertanto si tratta di una distribuzione quasi simmetrica.

Esempio 1.7.2

Nel caso delle risposte alla domanda (riportate sotto) "Ritieni che utilizzare strumenti tecnologici fin da ora ti possa aiutare anche in futuro in campo lavorativo?"

importanza tecnologia in futuro	1	2	3	4	5	totale
frequenza assoluta	1	5	21	44	59	130

La media aritmetica è di circa 4.192 307 692 3, la moda è 5 e la mediana è 4. Quindi siamo in presenza di una distribuzione asimmetrica positiva.



2 Indici di variabilità

Nello studio dei dati statistici è importante essere in grado di valutare la variabilità delle misure, detta anche dispersione. Gli indici di variabilità sono quei valori che rappresentano il grado di dispersione di una serie di dati e forniscono informazioni su quanto i dati sono *diversi* dal valore centrale (cioè dal baricentro scelto, l'indice di posizione scelto). Per questo motivo sono anche detti indici di dispersione.
Le misure di dispersione di una statistica sono funzioni non negative che hanno due proprietà fondamentali:

1. Valgono zero se gli elementi della statistica sono tutti uguali.
2. Sono positive se gli elementi della statistica non sono tutti uguali.

Gli indici di dispersione esaminati in questa dispensa sono:

- il range o campo di variazione,
- lo scarto semplice medio,
- lo scarto quadratico medio o deviazione standard,
- lo scarto interquartile.

2.1 Range o campo di variazione

Si definisce range o campo di variazione di una statistica $x_1, x_2, \ldots, x_k$ la differenza tra il valore massimo e il valore minimo, ossia il numero:

$$d = \max\{x_1, x_2, \ldots x_k\} - \min\{x_1, x_2, \ldots x_k\}. \tag{2.1.1}$$ *Range*

Esempio 2.1.1

Alla domanda che riguarda quanto l'utilizzo di device in ambito scolastico incentiva a studiare con più coinvolgimento, i ragazzi hanno risposto scegliendo un valore compreso nella scala tra 1 e 5.

Almeno una volta tutti i valori della scala sono stati scelti. Il range di

questa statistica quindi risulta: $d = 5 - 1 = 4$

All'interno dell'indagine condotta dal liceo, le statistiche che assumono valori compresi tra 1 e 5 sono numerose. Il range in tutte queste è sempre lo stesso e vale 4.
Per esempio, si può notare il comune range nelle tre domande:

- "Quanto ritieni importante l'utilizzo dei device in ambito scolastico in una scala da 1 (per niente importante) a 5 (fondamentale)?"
- "Quanto la scuola e le attività svolte a scuola con i device hanno influenzato o modificato il tuo metodo di lavoro (da 1 a 5)?"
- "Quanto gli insegnanti della tua classe fanno uso degli strumenti tecnologici durante le lezioni (da 1 a 5)?"

Invece analizzando la domanda in cui è stato chiesto di svolgere la media aritmetica dei voti di matematica, fisica e scienze, l'insieme risultante è il seguente:

$$\begin{aligned} M = \{&8, 8, 8, 8, 7, 8, 8, 7, 7, 6, 7, 5, 7.07, 7, 7, 6, 8, 10, 6, 6, 7, 8, 9, 5, 7, 6, 7, 6, 7, \\ &6, 7, 8, 7, 6.3, 7, 9, 7, 5, 9, 8, 6, 6, 5, 6, 7, 7, 5, 7, 7, 7, 8, 6, 7, 7, 7, 7.7714, \\ &5, 7, 8, 6.2, 6, 9, 6, 6, 7, 5.6, 7, 7, 5.75, 6, 6, 7, 7, 5.6, 6, 7, 7.6, 8, 6, 6, 5.7, \\ &7, 6.1, 7, 8, 8, 6.9, 7, 8.3, 5.6, 6, 7, 7, 6, 6, 5.6, 6.5, 7, 6, 6, 7, 6, 6, 6, 7, 8, 6, \\ &5, 6, 5, 6, 6, 6, 7.5, 7, 7, 7, 5, 8, 6, 7, 7, 6, 7, 7, 7, 7, 6, 6, 6, 7\} \end{aligned}$$

e il range risulta uguale a

$$d = \max\{x_1, x_2, \ldots, x_k\} - \min\{x_1, x_2, \ldots, x_k\} = 10 - 5 = 5.$$

2.2 Scarto semplice medio

Assegnata la statistica $(x_1, x_2, \ldots, x_n)$, con media aritmetica μ e considerati i valori assoluti degli scarti della media

Scarti dalla media

$$|x_1 - \mu|, |x_2 - \mu|, \ldots, |x_n - \mu|$$

si definisce scarto semplice medio il numero non negativo:

Scarto semplice medio

$$S_1 = \frac{|x_1 - \mu| + |x_2 - \mu| + \cdots + |x_n - \mu|}{n} \tag{2.2.1}$$

ossia la media aritmetica dei valori assoluti degli scarti.

Esempio 2.2.1

Alla domanda "Quanto ritieni di essere capace nell'utilizzo e nella gestione degli strumenti tecnologici?" gli studenti del liceo Bertolucci hanno risposto secondo la seguente statistica, scegliendo un valore compreso tra 1 e 5:

$$A = \{3,2,3,3,3,3,3,3,4,3,5,4,5,4,4,3,3,4,4,4,2,4,4,3,4,5,4,3,4,$$
$$4,3,3,3,3,5,3,5,3,3,2,4,4,3,3,3,4,5,5,4,3,3,4,4,3,3,3,$$
$$2,3,3,5,1,4,5,3,4,5,5,4,4,5,3,3,4,3,3,3,4,5,4,2,5,4,2,4,$$
$$4,1,3,4,3,3,5,3,3,3,3,4,3,4,4,2,5,4,3,2,3,4,4,5,4,3,3,2,$$
$$3,2,5,3,3,3,3,3,5,2,3,4,2,3,3,3,4\}$$

importanza tecnologia in futuro	1	2	3	4	5	totale
frequenza assoluta	2	12	58	39	19	130

Calcoliamo lo scarto semplice medio.
Prima di tutto ricaviamo la media dei valori, per poter poi trovare gli scarti. Il valore della media aritmetica risulta uguale a $\mu = 3.47$. Gli scarti dalla media in valore assoluto per ciascun valore della scala risultano:

$|1 - 3.47| = 2.47;$
$|2 - 3.47| = 1.47;$
$|3 - 3.47| = 0.47;$
$|4 - 3.47| = 0.52;$
$|5 - 3.47| = 1.52.$

Successivamente si sommano tutti gli scarti per ogni valore dell'insieme della statistica; la somma è costituita da tanti addendi quanti sono i valori della statistica (pari alla numerosità del campione). La somma degli scarti dalla media in valore assoluto risulta:

$$S = 99.62.$$

A questo punto per trovare lo scarto semplice medio, dividiamo la somma S per il numero n degli elementi dell'insieme dato:

$$s_1 = \frac{|x_1 - \mu| + |x_2 - \mu| + \ldots + |x_n - \mu|}{n} = \frac{S}{n} = \frac{99.62}{130} = 0.77,$$

e quindi 0,77 è il valore dello scarto semplice medio.

2.3 Scarto quadratico medio o deviazione standard

Assegnata la statistica $(x_1, x_2, \ldots, x_n)$, con media aritmetica μ, lo scarto quadratico medio è il numero non negativo:

Deviazione standard

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}} \tag{2.3.1}$$

In statistica spesso si utilizza un ulteriore indicatore, detto **varianza**, pari al quadrato della deviazione standard, σ^2. Naturalmente, la varianza non è una grandezza omogenea alla statistica di partenza, in quanto la sua unità di misura sarà data dal quadrato dell'unità di misura dei valori x_i

Varianza

$$\mathbf{var}(X) = \sigma^2. \tag{2.3.2}$$

Si osservi che, con la definizione data, lo scarto quadratico medio (o deviazione standard) è la media quadratica dei valori assoluti degli scarti (capitolo 1.4).

2.4 Scarto interquartile

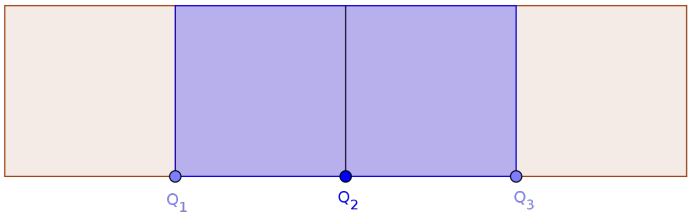
Data una distribuzione di un carattere quantitativo o qualitativo ordinabile (i cui valori possiedono un ordine *naturale*, si dicono quartili quei valori che ripartiscono la popolazione in quattro parti di uguale numerosità.
Generalmente si indicano con:

Q1 il valore che rappresenta il primo quartile, perciò il 25% dei valori della serie è minore di Q1 e il 75% è maggiore;

Q2 il valore che sta nella posizione centrale quando i valori della serie sono ordinati in modo crescente; quindi il 50% dei valori è minore di Q2 e il 50% dei valori è maggiore;

Q3 il valore che rappresenta il terzo quartile, perciò il 25% dei valori della serie è maggiore di Q1 e il 75% è minore.

Quindi i quartili dividono tutta la serie dei valori, ordinata in modo crescente, in quattro parti che contengono lo stesso numero di dati.

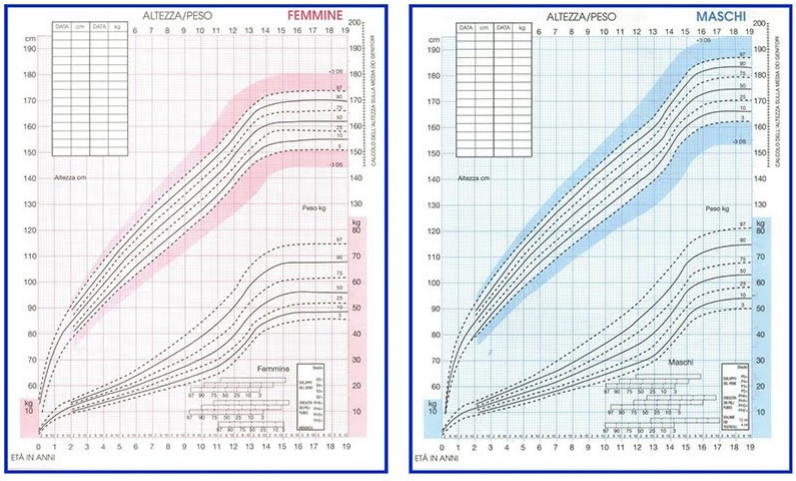


Esempio 2.4.1

In campo medico si utilizzano i percentili di crescita per valutare come procede la crescita di un bambino in peso e in altezza e se si possono escludere problemi relativi allo sviluppo. I centili funzionano come i quartili ma, anziché in quattro parti, la popolazione è divisa in 100 parti uguali.

I diagrammi percentili permettono di suddividere i bambini in base al loro peso o alla loro altezza, a partire dai più minuti (primo centile) fino a quelli più alti e grossi (nel centesimo). La maggior parte dei neonati si stabilisce di solito tra il 25° e il 75° centile.

Il seguente diagramma percentile, che studia il peso di bambini e bambine nei primi 3 anni di vita, proviene da una ricerca presente nel sito **www.medicitalia.it**, del resto online o presso i pediatri di base si trovano varie tabelle di questo tipo.



Tanner JM, Whitehouse RH. Atlas of Children's Growth, 1982

Tornando ai quartili, lo scarto interquartile, indicato solitamente IQR, è un indice di dispersione che permette di calcolare la differenza tra il terzo e il primo quartile, ossia l'ampiezza della fascia di valori che contiene la *metà centrale* dei valori osservati. Lo scarto interquartile rappresenta una misura di quanto i valori si allontanino dalla mediana ed è uguale a:

Scarto interquartile

$$IQR = Q_3 - Q_1. \tag{2.4.1}$$

Esempio 2.4.2

Calcoliamo lo scarto interquartile relativo alla media ottenuta nelle materie scientifiche fra i 130 studenti intervistati all'interno del liceo Attilio Bertolucci.

- Scriviamo l'insieme in ordine crescente.
- Determinare il punto medio che divide la serie a metà.
  Nel caso di un insieme composto da n elementi, con n pari, si tratta del punto tra i valori che si trovano nelle posizioni $n/2$ e $n/2+1$.
- Determinare la mediana della prima metà e della seconda metà in cui la serie di dati è stata suddivisa al punto 2 (ossia delle due metà dei dati create nel punto 2), i valori così trovati sono il primo e il terzo quartile: $Q1 = 6$ e $Q3 = 7$.
- Determinare lo scarto interquartile: $IQR = Q3 - Q1 = 7 - 6 = 1$.

Esempio 2.4.3

Ora calcoliamo i differenti indici di variabilità relativi ad uno stesso quesito, così che il lettore possa più facilmente confrontarli.

Consideriamo il carattere "Credi che la scuola ti incentivi ad utilizzare questi device (computer, cellulare, tablet, ...)?" e calcoliamo i relativi gli indici di variabilità.

Campo di variazione $\max x_i - \min x_i = 5 - 1 = 4$.

Scarto semplice medio Calcoliamo prima di tutto la media dei valori, risulta $\mu = 3.19$.

Utilizzando poi la formula

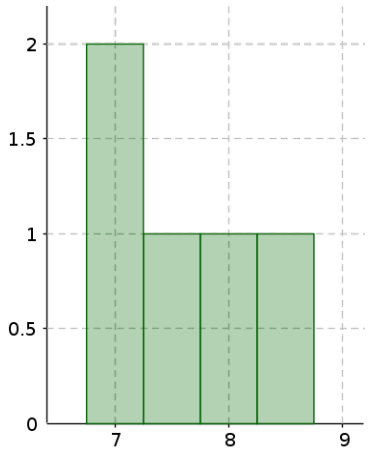
$$S_1 = \frac{|x_1 - \mu| + |x_2 - \mu| + \cdots + |x_n - \mu|}{n},$$

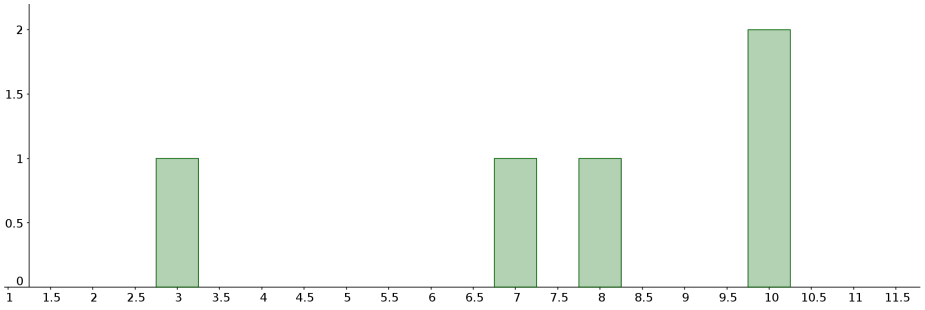
troviamo che lo scarto semplice medio è uguale a 0.73. Lo scarto semplice medio rappresenta la distanza media dei valori dalla media μ.
Lo scarto semplice medio si può calcolare anche a partire dalla mediana o da un'altra media, in questo esempio si è utilizzata la media aritmetica.

Deviazione standard Anche in questo caso occorre innanzi tutto la media aritmetica.
Utilizzando la formula

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_n - \mu)^2}{n}},$$

troviamo che la deviazione standard è uguale a 0.88.
La deviazione standard ci fornisce un'indicazione riguardo a quanto i valori si allontanano dal valore medio. Più il valore della deviazione standard è piccolo più la media è attendibile come indicatore globale della statistica.
Consideriamo la media di 5 voti effettuata tra due campioni diversi. Nel primo campione l'insieme dei valori è $I_1 = \{7, 7.5, 7, 8, 8.5\}$, nel secondo $I_2 = \{3, 7, 8, 10, 10\}$.
La media di entrambi è 7.6.
Calcoliamo le relative deviazioni standard: $\sigma_1 = 0.58$; $\sigma_2 = 2.58$.
Nel nostro caso 1 è minore di 2 ed infatti la media approssima in modo migliore i valori dell'insieme I_1 rispetto a quelli dell'insieme I_2.





Varianza In statistica si utilizza un altro indicatore, detto varianza, che è pari al quadrato della deviazione standard. Nel nostro caso $\sigma^2 = 0.78$.

Scarto interquartile Ordiniamo la serie di dati in ordine crescente e troviamo Q_1 e Q_3. Si ha $Q_1 = 3$ e $Q_3 = 4$ utilizzando la formula $IQR = Q3 - Q1$ troviamo che lo scarto interquartile relativo al carattere studiato è $IQR = 1$. Lo scarto interquartile ci fornisce un'informazione riguardo a quanto i valori si allontanano dalla mediana.

II Statistica bivariata

3 Tabelle a doppia entrata

Quando si vuole indagare la relazione tra due caratteri inerenti ad una stessa popolazione statistica si produce una statistica bivariata o a due dimensioni. Lo studio contemporaneo di due caratteri su n individui può essere comodamente visualizzato utilizzando una tabella a doppia entrata.

Per costruire una tabella a doppia entrata, occorre considerare la distribuzione doppia di frequenze dividendo in classi gli individui della popolazione n secondo le modalità dei due caratteri e rilevando le frequenze.

		Y						
		y_1	y_2	...	y_j	...	y_t	$TOTALE$
X	x_1	$f_{1,1}$	$f_{1,2}$	...	$f_{1,j}$	...	$f_{2,t}$	X_1
	x_2	$f_{2,1}$	$f_{2,2}$	...	$f_{2,j}$	...	$f_{2,t}$	X_2
	...	...	...	...	...	...	...	...
	x_i	$f_{i,1}$	$f_{i,2}$	...	$f_{i,j}$	...	$f_{i,t}$	X_i
	...	...	...	...	...	...	...	...
	x_s	$f_{s,1}$	$f_{s,2}$	...	$f_{s,j}$	...	$f_{s,t}$	X_t
	$TOTALE$	Y_1	Y_2	...	Y_j	...	Y_t	N

Figura 3.1 Un esempio di tabella a doppia entrata

In generale, si considera la statistica bivariata riguardante lo studio di due caratteri X e Y sugli n individui di una popolazione. Consideriamo dunque il carattere

X che assume le s modalità $x_1, x_2, \ldots, x_s$ e il carattere Y che assume le t modalità $y_1, y_2, \ldots, y_t$.
La statistica relativa ad X e Y fornisce la matrice F a s righe e t colonne, i cui elementi $f_{i,j}$ rappresentano le frequenze degli individui della popolazione n che presentano il carattere X con modalità x_i e il carattere Y con modalità y_j.
La somma X_i delle frequenze della riga i della matrice in cui il carattere X si presenta con modalità y_i, si chiama frequenza marginale di X

Frequenza marginale di X

$$X_i = \sum_{j=1}^{t} f(i,j).$$

Mentre la somma Y_j delle frequenze della colonna della matrice in cui il carattere Y si presenta con modalità y_j, si chiama frequenza marginale di Y

Frequenza marginale di Y

$$Y_j = \sum_{i=1}^{s} f(i,j).$$

Si veda la tabella 3.1 per visualizzare le frequenze marginali.

Esempio 3.0.1

Consideriamo la relazione presente tra il numero di device posseduti e la valutazione media ottenuta nelle materie scientifiche.

X: "Quale è stata la valutazione media ottenuta nelle materie scientifiche?"
Y: "Quanti device possiedi?"

La tabella a doppia entrata che si utilizza per rappresentare la statistica bivariata X, Y è riportata di seguito.

		Y : *numero di device*							
		0	1	2	3	4	5	>5	*TOT*
X : *voto*	2								0
	3								0
	4								0
	5				1	5		2	8
	6		1	5	17	13	7	4	47
	7			8	14	13	9	5	49
	8			3	8	8		2	21
	9				1	1		2	4
	10				1				1
	TOT	0	1	16	42	40	16	15	130

Esempio 3.0.2

Consideriamo la relazione presente tra il sesso degli studenti X e due caratteri differenti:

carattere I: il tempo di utilizzo medio dei device per motivi scolastici durante la settimana;
carattere II: l'importanza dell'utilizzo dei device in una scala da 1 a 5.

Descriviamo i due caratteri separatamente, sempre rispetto al genere, in due differenti tabelle.

Carattere I

X: Genere
Y: "Quanto tempo utilizzi mediamente questi device per motivi scolastici durante la settimana?"

		Y : tempo di utilizzo settimanale (ore)				
		meno di 1	*1 – 2*	*3 – 6*	*più di 6*	*TOT*
X : genere	*Maschio*	13	32	24	2	71
	Femmina	9	21	23	5	58
	TOT	22	53	47	7	129

Carattere II

X: Genere
Y: "Quanto ritieni importante l'utilizzo dei device in ambito scolastico in una scala da 1 (per niente importante) a 5 (fondamentale)?"

		Y : importanza dell'utilizzo dei device					
		1	2	3	4	5	*TOT*
X : genere	*Maschio*	3	6	26	26	10	71
	Femmina		4	18	33	3	58
	TOT	3	10	44	59	13	129

Come si osserva dal totale della tabella (129 invece di 130), alcuni valori non sono stati ritenuti validi per l'indagine statistica e per questo motivo non sono stati tenuti in considerazione. I motivi per cui alcuni dati di un'indagine statistica vengono talvolta scartati sono numerosi e variano a seconda delle modalità con cui questa viene condotta: dati incoerenti, valori non accettabili, errori durante la compilazione, interpretazione errata della domanda, errori di carattere formale...

In questo caso si tratta di un errore nell'inserimento dei dati durante la compilazione del questionario.

4 Indipendenza e dipendenza statistica

Nel corso di un'indagine statistica basata sullo studio di più caratteri (in genere due per volta, per questo si parla di statistica *bivariata*) potrebbe risultare interessante indagare quanto i due caratteri considerati siano tra loro interdipendenti, ossia quanto uno dipenda dall'altro e viceversa. A questo proposito potremmo trovare, per esempio, che l'età degli studenti di una classe dipenda dalla classe frequentata (questo è abbastanza ovvio, considerato che in genere il numero di studenti ripetenti o anticipatari è, in percentuale, piuttosto basso): in questo caso è intuitivo il fatto che la dipendenza del carattere "età" e del carattere "classe frequentata" deve essere elevata.

Nell'indagine svolta al Liceo Attilio Bertolucci sarebbe di notevole interesse studiare la dipendenza che intercorre tra il possesso di strumenti digitali e l'andamento scolastico nelle materie scientifiche (misurato con la media dei voti in matematica, fisica e scienze naturali). Nel corso della trattazione proveremo a scoprire quanto l'utilizzo di device influenzi l'aspetto della pagella di una studentessa o di uno studente del Bertolucci.

Studiando una tabella a doppia entrata possiamo definire l'indipendenza statistica in questo modo:

Indipendenza statistica. Due caratteri A e B si dicono indipendenti se le frequenze relative del carattere A negli individui che possiedono il carattere B nella modalità y_1 sono le stesse di quelli che possiedono il carattere B con modalità y_2, con modalità y_3, ...

Indipendenza statistica

Quando prendiamo in esame due caratteri, di cui almeno uno è qualitativo, si parla di **connessione**, mentre se i due caratteri sono quantitativi si parla di **correlazione** statistica.

Contingenza
Connessione
Correlazione

Se si adopera una tabella a doppia entrata per studiare due caratteri, si parlerà di tabella di **contingenza**; se almeno uno dei caratteri esaminati è qualitativo si parla di tabella di **connessione**, se tutti i caratteri sono quantitativi si parla, invece, di tabella di **correlazione**.

Esempio 4.0.1

Analizziamo la relazione (se esiste e di che tipo si tratta) tra l'utilizzo settimanale di device elettronici e il genere di studenti e studentesse.

		tempo di utilizzo settimanale (ore)				
		meno di 1	1 – 2	3 – 6	*più di* 6	*TOT*
genere	*Maschio*	13	32	24	2	71
	Femmina	9	21	23	5	58
	TOT	22	53	47	7	129

In questo caso, avendo un carattere quantitativo e uno qualitativo, è possibile utilizzare una tabella di connessione che rende molto immediata la visualizzazione dei dati e la loro interpretazione. Non esiste alcuna relazione tra il tempo di utilizzo settimanale dei device elettronici e il genere degli studenti: o meglio i due caratteri sono indipendenti.

Esempio 4.0.2

Analizziamo ora, invece, analizziamo la relazione tra l'utilizzo settimanale di device elettronici e i voti in pagella nelle materie scientifiche degli studenti e studentesse intervistate.

		tempo di utilizzo settimanale (ore)				
		meno di 1	1 – 2	3 – 6	*più di* 6	*TOT*
media voti	5	1	5	2		8
	6	8	19	17	2	46
	7	11	18	15	5	49
	8	2	9	10		21
	9		2	2		4
	10			1		1
	TOT	22	53	47	7	129

In questo caso, avendo due caratteri quantitativi, è possibile utilizzare una tabella di correlazione. In questo caso, si nota una possibile relazione tra il tempo di utilizzo settimanale dei device elettronici e la media dei voti degli studenti. Occorre però capire quanto questi caratteri incidano l'uno sull'altro.

4.1 Indici per la dipendenza e l'indipendenza statistica

Occorre a tale proposito individuare uno strumento matematico che consenta di quantificare la relazione di dipendenza (o indipendenza, perché no?) di due caratteri. Introduciamo dunque il parametro di **covarianza** che fornisce informazioni circa il grado di dipendenza di due caratteri:

Covarianza

$$\mathbf{cov}(X;Y) = \frac{(x_1 - \mu_X)(y_1 - \mu_Y) + \ldots + (x_n - \mu_X)(y_n - \mu_Y)}{n}. \tag{4.1.1}$$

Nella formula compare la somma dei prodotti dei rispettivi scarti dalla media di X (cioè μ_x) e Y (cioè μ_y), ovvero dei due caratteri. Definiamo dunque $(x_n - \mu_X)$ come x'_n e $(y_n - \mu_Y)$ come y'_n; riscriviamo ora la covarianza con la sommatoria:

$$\mathbf{cov}(X;Y) = \frac{\sum_n x'_n y'_n}{n}.$$

La covarianza può anche essere scritta utilizzando i valori medi delle distribuzioni marginali e μ_{xy}:

$$\mathbf{cov}(X;Y) = \mu_{XY} - \mu_X \mu_Y \tag{4.1.2}$$

(con $\mu_{XY} = \dfrac{x_1y_1 + x_2y_2 + \cdots + x_ny_n}{n}$, valore medio della variabile prodotto). La grandezza matematica che ci fornisce informazioni circa la dipendenza statistica è il coefficiente di correlazione, definito come:

$$\rho = \frac{\mathbf{cov}(X;Y)}{\sigma_X\sigma_Y} \tag{4.1.3}$$

Coefficiente di correlazione

in cui σ_X indica lo scarto quadratico medio dei valori x_i, mentre σ_Y lo scarto quadratico medio dei valori y_i.
Analogamente a quanto visto per la covarianza, possiamo infine riscrivere il coefficiente di correlazione come:

$$\rho = \frac{\mu_{XY} - \mu_X\mu_Y}{\sigma_X\sigma_Y}.$$

Si può dimostrare che in ogni caso: $-1 \leq \rho \leq 1$.
Essa fornisce informazioni circa la dipendenza dei caratteri e permette di comprendere se la dipendenza è di tipo **diretto** o **inverso**, ovvero se al crescere di uno dei due caratteri l'altro cresce o decresce. In particolare

- se $\rho = 0$ i due caratteri sono indipendenti;
- se $\rho \neq 0$ i due caratteri hanno una qualche forma di dipendenza che vale la pena studiare.

Se vi è una dipendenza di qualche tipo ($\rho \neq 0$):

- sussiste dipendenza diretta, se $\rho = 1$ (dipendenza lineare e diretta);
- sussiste dipendenza inversa, se $\rho = -1$ (dipendenza lineare e inversa).

Nei seguenti quattro esempi esaminiamo più nel dettaglio il concetto di dipendenza, applicandolo ad alcuni casi concreti.

Esempio 4.1.1

Correlazione Covarianza

Per mostrare in cosa consiste il concetto di correlazione statistica, consideriamo le due domande del questionario, rispettivamente una delle quali riporta il numero di device posseduti dagli studenti e l'altra la valutazione media ottenuta nelle materie scientifiche nel primo periodo. In base ai dati delle statistiche si ottiene la seguente tabella di correlazione.

"Quanti device possiedi?"	"Quale è stata la valutazione media ottenuta nelle materie scientifiche nel primo periodo?"
2	7
2	5
5	7,07
4	7
4	7
5	6
3	8
3	10
6	6
4	6

A titolo di esempio è stata riportata solo una parte dei valori delle statistiche. La tabella con tutti i valori può essere richiesta agli autori.

Osservando la tabella è possibile stabilire che tra i due caratteri esiste un rapporto di correlazione statistica, in quanto i due caratteri sono entrambi quantitativi.

Ora calcoliamo la covarianza. Per prima cosa si calcola le medie aritmetiche dei rispettivi insiemi statistici. Le rispettive medie sono: $\mu_X = 4$ dei device posseduti e $\mu_Y = 7$ dei voti.

Una volta calcolati gli scarti dalla media delle due statistiche, la somma dei prodotti calcolati tra i rispettivi scarti corrispondenti risulta uguale a 11.07. A questo punto abbiamo ottenuto il valore del numeratore della formula della covarianza, per cui:

$$\mathbf{cov}(X;Y) = \frac{(x_1 - \mu_X)(y_1 - \mu_Y) + \cdots + (x_n - \mu_X)(y_n - \mu_Y)}{n} = \frac{11.07}{n} = 1.107$$

Il valore della covarianza corrisponde a 1.107.

Esempio 4.1.2

Due caratteri A e B si dicono indipendenti se:

- A assume s modalità,
- B t modalità,
- Si calcolano le frequenze congiunte $f(i,j)$

- Qualsiasi frequenza congiunta calcolata assume lo stesso valore del prodotto delle frequenze marginali diviso per N $f_{i,k} = \dfrac{A_i B_k}{N}$

Osserviamo ora, attraverso una tabella, la relazione che esiste tra la classe frequentata dagli studenti e il numero di social network a cui sono iscritti. Per semplificare i calcoli sono stati utilizzati i dati forniti da un campione ristretti di 15 studenti.

		classe					
		I	II	III	IV	V	*TOT*
social networks	1		1	2			3
	2		1	1			2
	3		1	1		1	3
	> 3		2	3	1	1	7
	TOT	0	6	7	1	1	15

Esempio 4.1.3

Per mostrare la relazione di dipendenza diretta fra due caratteri differenti di una statistica, si utilizza un esempio dalla fisica e non riferito all'indagine statistica sui device, quindi i dati usati non risulteranno essere presi dal questionario su cui fino ad ora ci si è basati.
Dati i seguenti valori numerici di pressione e forza:

Pressione (x)	**Forza (y)**
20 kPa	102 N
633 Pa	50 N
1.5 kPa	95 N

La media delle due statistiche risulta rispettivamente uguale a $\mu_X = 7.38 \times 10^3$ e $\mu_Y = 82,3$.

Il valore di μ_{XY} è 7.38×10^5. Per poter calcolare il valore del coefficiente di correlazione è necessario procurarsi la deviazione standard delle due statistiche: si ottiene $\sigma_X = 1.1 \times 10^4$ e $\sigma_Y = 28$.

Inoltre è anche necessario calcolare la covarianza. Applicando la formula (4.1.2) si ottiene 1.3×10^5, ed infine è possibile calcolare il coefficiente di

correlazione:

$$\rho = \frac{\mathbf{cov}(X;Y)}{\sigma_X \sigma_Y} = 0.42.$$

Essendo ρ diverso da 0 si può affermare che i due caratteri sono dipendenti, inoltre essendo positivo la dipendenza è diretta.

Esempio 4.1.4

In questo esempio, invece, viene esaminato il caso della dipendenza inversa, ci aspettiamo quindi di trovare $\rho < 0$.
Consideriamo i valori di pressione e superficie espressi nella tabella seguente.

Pressione (x)	**Superficie (y)**
20 kPa	$5.1 \times 10^{-3}\,\mathrm{m}^2$
633 Pa	$7.9 \times 10^{-2}\,\mathrm{m}^2$
1.5 kPa	$6.4 \times 10^{-2}\,\mathrm{m}^2$

Calcoliamo la media dei due insiemi statistici $\mu_X = 7.4\,\mathrm{kPa}$, $\mu_Y = 5 \times 10^{-2}\,\mathrm{m}^2$.
Il valore μ_{XY} è uguale a 52.1.
Per calcolare il coefficiente di correlazione calcoliamo la deviazione standard dei due caratteri: $\sigma_X = 1.1 \times 10^4$, $\sigma_Y = 4 \times 10^{-2}$.
Una volta calcolata la deviazione standard possiamo ricavare la covarianza: $\mathbf{cov}(X;Y) = -2.8 \times 10^2$.
Infine troviamo il coefficiente di correlazione:

$$\rho = \frac{\mathbf{cov}(X;Y)}{\sigma_X \sigma_Y} = -0.7.$$

Si noti che, come anticipato, $\rho < 0$.

4.2 Interpolazione

Una distribuzione doppia è rappresentata nel piano cartesiano da una nuvola di punti le cui coordinate sono costituite da coppie ordinate del tipo $(x_n; y_n)$.

Nel grafico si posizionano i dati relativi al primo carattere lungo l'asse x, quelli relativi al secondo lungo l'asse y. La scelta di quale sia il primo e quale il secondo carattere è compito di chi analizza i dati.

Durante un'indagine statistica spesso i dati rilevati presentano lacune a causa o di errori di rilevazione o per la scarsità dei dati raccolti. Per questo motivo, durante

la rappresentazione grafica è possibile non non trovare un grafico lineare. Questo può avvenire anche quando la relazione fra i due caratteri è in realtà lineare o quasi.

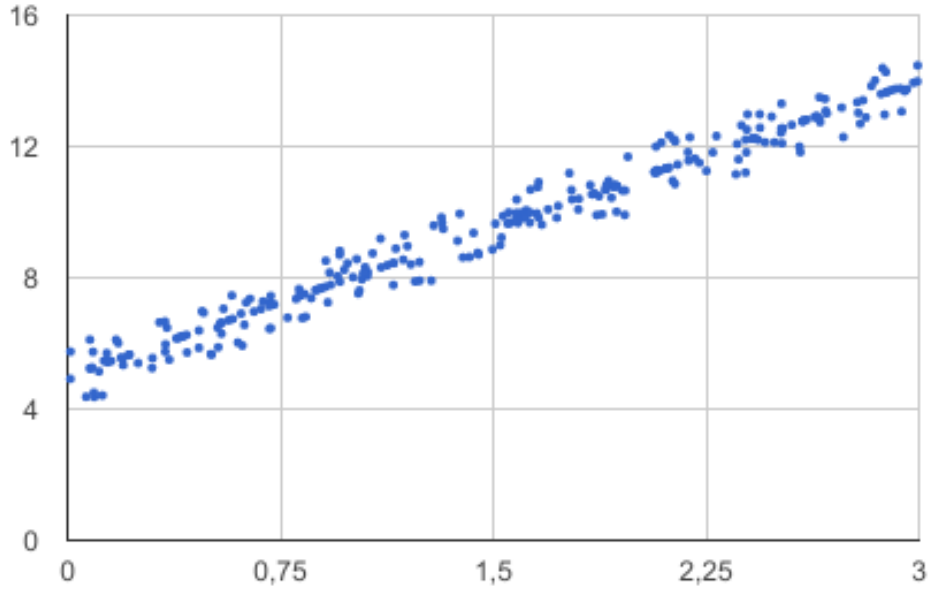


Figura 4.1 Nuvola di punti con disposizione approssimativamente lineare

Quando usare l'interpolazione

Per colmare le lacune dovute ai dati mancanti nello studio dei legami tra due variabili statistiche, si ricorre alla interpolazione, ossia si usa una funzione $Y = f(X)$ che rappresenti il fenomeno. La curva illustrata non necessariamente passerà per i punti trovati ma approssimerà la distribuzione. Infatti, in linea teorica è possibile trovare una funzione che passi realmente per tutte le coppie di valori, ed è la funzione di interpolazione matematica. Quando però il numero di coppie è elevato, essa risulta poco chiara, di scarsa utilità e matematicamente complicata. Si utilizza dunque, una funzione che approssima la statistica bivariata, che non passa necessariamente per tutti i punti del grafico. Questa viene detta funzione di interpolazione statistica.

Osservando la disposizione dei punti nel piano cartesiano è possibile dedurre di che tipologia è la funzione interpolatrice. Ad esempio, se i punti sono circa allineati la funzione sarà $Y = kX$, se invece i due caratteri sono inversamente proporzionali la funzione sarà del tipo $Y = \frac{k}{X}$.

Esistono vari metodi di interpolazione.

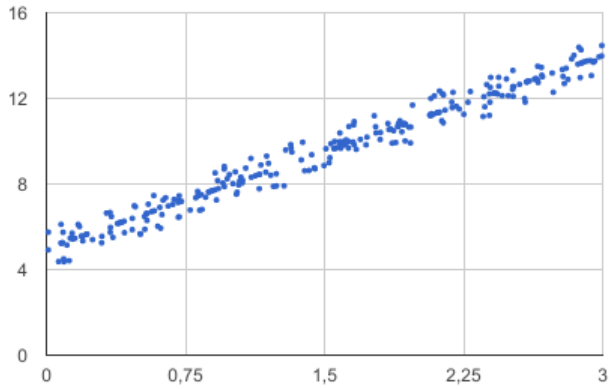
Retta di regressione lineare

Questo metodo approssima la nuvola dei punti ad una retta. Il metodo è approfondito nel paragrafo "Retta di regressione lineare".

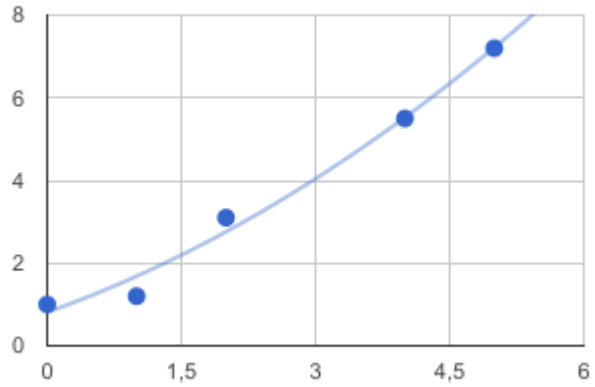
Metodo dei minimi quadrati

Questo metodo permette di determinare una funzione il cui grafico si avvicina ai punti dati nel modo migliore possibile. Questo significa che la curva che rappresenta la funzione scelta si troverà il più vicino possibile ai punti della nuvola di dati. Il grafico, detto curva di regressione, si ottiene imponendo che sia minima la somma dei quadrati degli scarti, cioè la somma dei quadrati delle distanze dei punti della funzione "modello" dai punti "reali" della statistica.
Si possono distinguere diverse tipologie di grafici (lineare, quadratico, polinomiale, esponenziale, logaritmico, logistico, ...). Il grafico da utilizzare come modello per rappresentare la nuvola di punti andrà scelto da chi analizza i dati, tenendo conto di tutte le informazioni sulla statistica e i caratteri in esame.
Nei seguenti esempi si possono vedere i grafici relativi ad alcuni casi particolari.

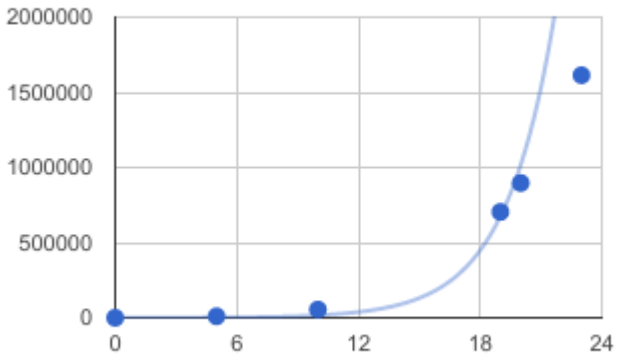
Esempio 4.2.1



Esempio 4.2.2



Esempio 4.2.3



4.3 Retta di regressione lineare

Retta di regressione lineare

La retta di regressione lineare è uno strumento utile per studiare il comportamento di una certa famiglia statistica di elementi: questo significa ricercare una funzione che approssima l'andamento dei dati, avendo scelto come tipologia di funzione quella lineare, il cui grafico è una retta I dati in questione sono rappresentati dai valori assunti dai due caratteri presi in esame, come visto nei paragrafi precedenti,

uno dei caratteri rappresenta l'ascissa e l'altro l'ordinata dei punti.
Dallo studio della geometria analitica sappiamo che una retta è identificata da un punto e dal coefficiente angolare. Quindi per trovare l'espressione analitica della retta da utilizzare come modello dei dati, sarà necessario determinare un punto e un coefficiente angolare adatti.
Le grandezze statistiche utilizzate a questo scopo sono il baricentro e il coefficiente di regressione.

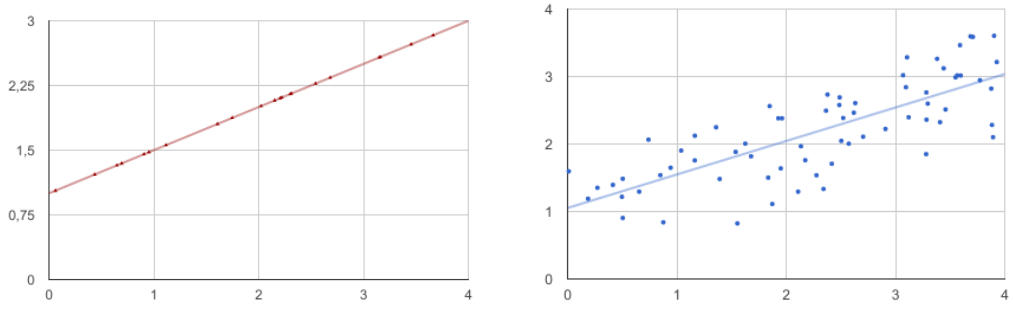
Siano quindi X e Y i due caratteri in esame, supposti legati dalla relazione:

$$Y = mX + q.$$

L'espressione indica che Y dipende da X in modo lineare. In alcuni casi X e Y sono veramente grandezze legate da una relazione di dipendenza lineare, in altri si tratta solo di un modello approssimato e semplificato.

Le coordinate della nuvola di punti

Si parte dai dati rilevati, i valori $x_1, x_2, x_3, \ldots, x_n$ e i corrispondenti valori $y_1, y_2, y_3, \ldots, y_n$. I valori corrispondenti così determinati costituiscono i coppie ordinate $(x_i; y_i)$. Nel momento in cui si rappresentano nel piano le coppie, si genera la "nuvola" di punti di cui si è parlato in precedenza.
Tale nuvola di punti appartiene alla retta $Y = mX + q$, solo se Y dipende veramente da X o se Y è calcolato a partire da X in modo matematico sfruttando una dipendenza lineare.

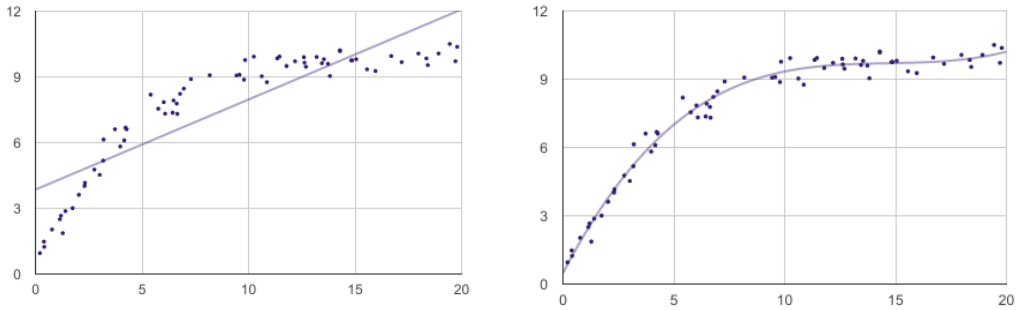


(a) *Interpolazione matematica* **(b)** *Interpolazione statistica*

Figura 4.2 In statistica non si utilizza l'interpolazione matematica, i punti spesso non appartengono alla funzione di regressione calcolata

In realtà quindi gli n punti della nuvola non appartengono tutti alla retta, ma si sceglie la retta che meglio approssima e che meglio identifica la nuvola, avvicinandosi in modo quasi uniforme a tutti i punti.

In linea di principio il procedimento si può applicare in tutti i casi, anche quando la relazione fra X e Y non è lineare, naturalmente in tale modo non si ottiene una funzione rappresentativa della nuvola di dati.



(a) *Retta di regressione* **(b)** *Polinomio interpolante*

Figura 4.3 Quando X e Y non sono legati da una relazione lineare, la retta di regressione può rivelarsi non rappresentativa, in certi casi si utilizzano altre curve, come ad esempio polinomi o esponenziali.

I valori m e q della retta calcolata con questo procedimento risulteranno quelli che adotteremo nell'equazione lineare precedente che lega Y a X. Tale retta prende il nome di retta di regressione lineare.

4.3.1 Il punto: il baricentro

Il punto scelto per il calcolo della retta, il baricentro, viene calcolato utilizzando la media dei valori assunti dai caratteri:

$$G(\mu_X;\mu_Y) : \mu_X = \frac{\sum x_i}{n} \quad \mu_Y = \frac{\sum y_i}{n} \tag{4.3.1}$$

Le rette passanti per G hanno equazione $Y - \mu_y = m(x - \mu_X)$.

4.3.2 Il coefficiente angolare: il coefficiente di regressione

Per calcolare il coefficiente di regressione dobbiamo imporre l'equazione che renda minima la somma delle distanze dei punti della nuvola di dati dalla retta di regressione, cioè, avendo chiamato $f(X) = m(X - \mu_X) + \mu_Y$ la funzione di regressione:

$$S(m) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - \mu_Y - m(x_i - \mu_X))^2$$

Svolgendo i calcoli del secondo membro, si osserva che l'equazione ha la forma matematica di un polinomio di secondo grado in m:

$$S(m) = am^2 + bm + c$$

dove $a = \sum_{i=1}^{n}(x_i - \mu_x)^2$, $b = -2\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$, $c = \sum_{i=1}^{n}(y_i - \mu_y)^2$.
All'equazione precedente corrisponde una parabola con concavità verso l'alto, dato che, essendo a una somma di quadrati, $a > 0$ per ciascun x. L'ascissa del vertice di una parabola (sull'asse di simmetria) si calcola con il rapporto $-\frac{b}{2a}$. Essendo $a > 0$, ossia una parabola con la concavità rivolta verso l'alto, il vertice risulta essere il valore delle ordinate minimo m che la parabola può assumere.
Quindi il problema di determinare il minimo della somma $S(m)$ si riconduce a quello di determinare l'ascissa del vertice della parabola.
Perciò :

$$m = -\frac{b}{2a} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{n}(x_i - \mu_X)^2}$$

Coefficiente di regressione di Y in X

In relazione alle definizioni di covarianza (4.1) e varianza (2.3.2), si può riscrivere la formula per calcolare m come

$$m = \frac{\mathbf{cov}(X;Y)}{\sigma_X^2} \tag{4.3.2}$$

m è quindi il coefficiente di regressione di y in x.
In conclusione, essendo l'equazione della retta di coefficiente m passante per x_0 $y - y_0 = m(x - x_0)$, la retta di regressione lineare passante per il baricentro che rende minima la somma dei quadrati degli scarti (x_0) ha equazione

Regressione di Y in X

$$y = \mu_y + \frac{\mathbf{cov}(X;Y)}{\sigma_X^2}(x - \mu_x) \tag{4.3.3}$$

Se invece cerchiamo una retta di equazione $X = \mu_X + M(Y - \mu_Y)$ che passa ancora per il baricentro della distribuzione $G(\mu_X;\mu_Y)$ e che rende minima la somma dei minimi quadrati, si ottiene:

$$S_1(M) = \sum_{i=1}^{n}\left(x_i - \mu_X - M(y_i - \mu_Y)\right)^2$$

Il calcolo, che è analogo al precedente, è lasciato al lettore per esercizio:

$$M = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{n}(y_i - \mu_Y)^2} = \frac{\mathbf{cov}(X;Y)}{\sigma_Y^2}$$

M è definito come coefficiente di regressione di X in Y. In questo caso la retta di regressione lineare ha equazione:

$$X = \mu_X + \frac{\mathbf{cov}(X;Y)}{\sigma_Y^2}(Y - \mu_Y) \tag{4.3.4}$$

Regressione di X in Y

Se i punti fossero allineati le due rette coinciderebbero passando esattamente per tutti gli n punti e si avrebbe

$$M = \frac{1}{m} \quad \text{cioè} \quad \frac{\mathbf{cov}(X,Y)^2}{\sigma_X^2 \cdot \sigma_Y^2} = 1$$

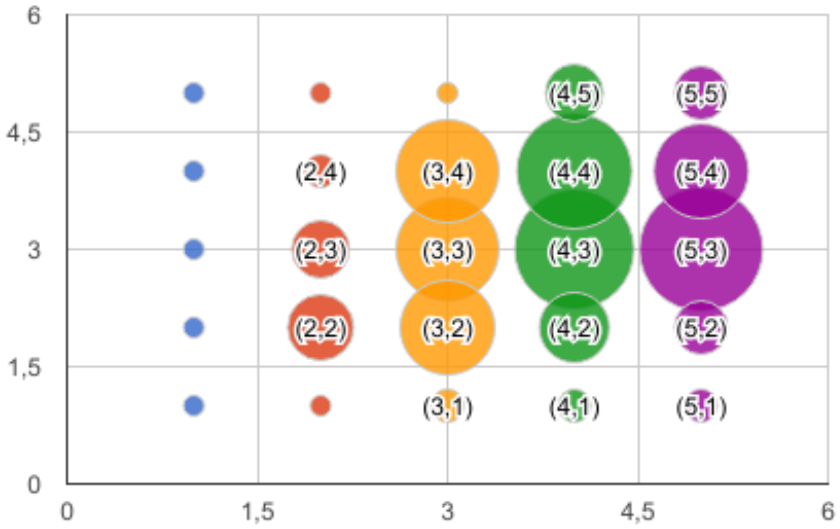
Da un punto di vista algebrico, questa condizione equivale a richiedere che il quadrato del coefficiente di correlazione sia 1, si veda l'equazione (4.1.3).
Solitamente però i punti non sono allineati perciò:

$$\frac{\mathbf{cov}(X,Y)^2}{\sigma_X^2 \cdot \sigma_Y^2} \neq 1$$

Si può quindi dedurre che più il coefficiente di correlazione è vicino a 1, più i punti del grafico sono allineati.

Esempio 4.3.1

Il fatto che i docenti facciano uso degli strumenti tecnologici durante le lezioni e il fatto che gli studenti si sentano incentivati a loro volta ad utilizzarli sono fra loro correlati? Per stabilirlo si analizza la correlazione e l'eventuale dipendenza o indipendenza di due caratteri X e Y che si riferiscono alle possibili risposte ai due quesiti sotto riportati.
Quesito 1: Quanto gli insegnanti della tua classe fanno uso degli strumenti tecnologici durante le lezioni?
A questo quesito si associa il carattere X (rappresentato sull'asse x) con le sue modalità da 1 a 5. Quesito 2: Credi che la scuola ti incentivi ad utilizzare questi devices (computer, cellulare, tablet...)?
Anche a questo quesito si associa un carattere, Y (rappresentato sull'asse y), avente sempre come modalità i valori da 1 a 5.



Per stabilire la dipendenza dei due caratteri X e Y utilizziamo il coefficiente di correlazione (4.1.3).

$$\mathbf{cov}(X;Y) = \frac{(x_1-\mu_X)(y_1-\mu_Y)+\cdots+(x_n-\mu_X)(y_n-\mu_Y)}{n} = 0.183\,668\,639$$

$$\sigma_X = 0.930\,148\,327 \quad \text{e} \quad \sigma_Y = 0.887\,332\,169$$

Sostituendo nella formula si ottiene

$$\rho = \frac{\mathbf{cov}(X;Y)}{\sigma_X\sigma_Y} = 0.222\,534\,101\,5$$

Essendo $\rho \neq 0$ si può affermare che i due caratteri sono dipendenti. Si nota però che il valore che descrive la correlazione è lontano da 1, pertanto la correlazione è bassa.
Come si poteva immaginare c'è un legame fra questi due caratteri, che si influenzano a vicenda, ma il risultato ottenuto non permette di stabilire un rapporto di causalità.

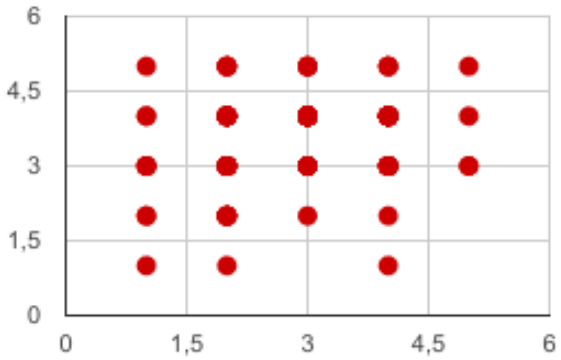
Esempio 4.3.2

Consideriamo i due caratteri, ricavabili dai seguenti quesiti:

"Quanto la scuola e le attività svolte a scuola con i device hanno influenzato o modificato il tuo metodo di lavoro?" (carattere X);

"Quanto ritieni importante l'utilizzo dei devices in ambito scolastico in una scala da 1 a 5 ?" (carattere Y).

Per studiare il modo in cui sono legati questi due caratteri, ne calcoliamo la retta di regressione. Prima di tutto però rappresentiamo il grafico cartesiano con la nuvola di punti (x_i, y_i).



Dall'osservazione del grafico si può notare che i punti sono molto "sparpagliati" e non sembrano situarsi lungo una linea. Ci si aspetta perciò che la retta di regressione sia in questo caso poco rappresentativa dei dati. E' importante anche considerare che i punti si sovrappongono, infatti possono solo assumere valori interi per X e Y, compresi fra 1 e 5, quindi nel grafico non saranno visibili tutti i 131 punti ed è più difficile farsi un'idea "ad occhio" di dove si trovi il baricentro.

Ricaviamo la retta di regressione lineare, ossia la retta che rende minima la somma dei quadrati degli scarti. Per prima cosa ricaviamo il baricentro $G(X;Y)$ dei 131 punti:

$$\mu_X = \frac{\sum x_i}{n} = \frac{20}{7} \quad \text{e} \quad \mu_Y = \frac{\sum y_i}{n} = \frac{32}{9},$$

$$G\left(\frac{20}{7}; \frac{32}{9}\right).$$

Imponiamo, poi, che la retta passi per G:

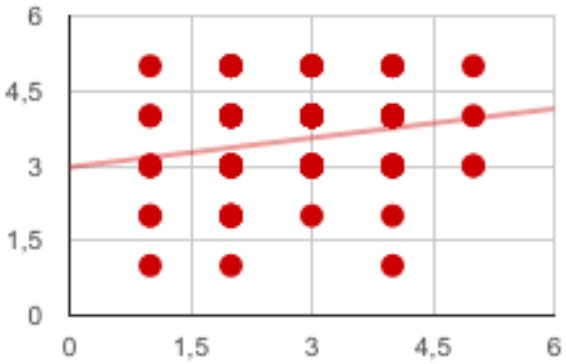
$$Y - \frac{32}{9} = m\left(X - \frac{20}{7}\right).$$

Ricaviamo infine il coefficiente angolare della retta di regressione con la formula (4.3.2)

$$\mathbf{cov}(X;Y) = \cdots = \frac{1}{6} \quad \text{e} \quad \sigma_X^2 = \cdots = \frac{8}{9} \quad \longrightarrow \quad m = \frac{3}{16}.$$

Tutti i calcoli sono lasciati al lettore. Sostituendo e semplificando si ottiene l'equazione della retta.

$$Y = \frac{3}{16}X + \frac{761}{252},$$



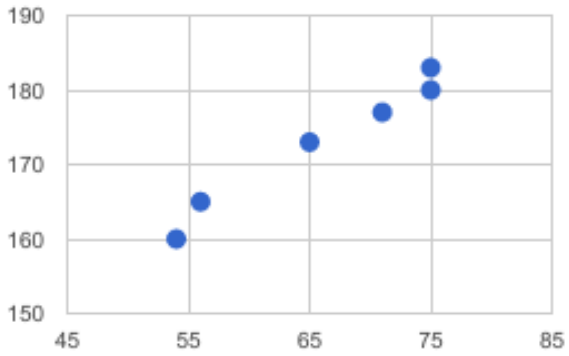
come si evince dal grafico la retta di regressione trovata non rappresenta in modo ottimale la nuvola di punti, questo è dovuto al fatto che tra i due caratteri non vi è una dipendenza lineare.

Esempio 4.3.3

In quest'ultimo esempio, invece, consideriamo un caso estraneo alla statistica sui device, ma nel quale si può osservare veramente bene cosa accade quanto i caratteri sono lineari, anche sei punti non sono matematicamente allineati sulla stessa retta.
Consideriamo i due caratteri "peso" e "altezza" in un campione costituito da 6

studenti scelti in una classe della scuola.

Peso in kg (x)	**Altezza in** cm (y)
54	160
56	165
65	173
71	177
75	180
75	183

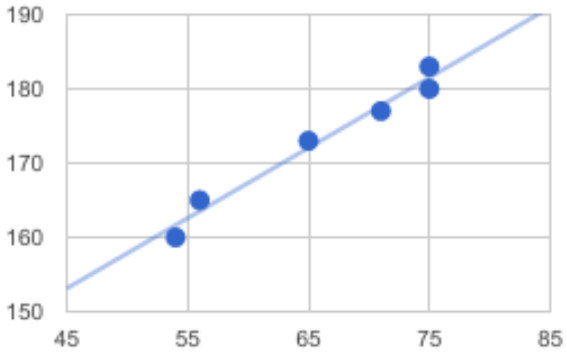


Calcoliamo il baricentro G dei sei punti: $G(66;173)$. Imponendo poi il passaggio della retta per il punto G si ottiene: $y-173=m(x-66)$. Si ricava infine il coefficiente angolare della retta di regressione utilizzando lo stesso procedimento usato nell'esempio precedente. Ora si ottiene

$$\mathbf{cov}(X;Y)=\cdots=68.166\,666\,67 \quad \text{e} \quad \sigma_X^2=\cdots=86.4 \quad \longrightarrow \quad m=0.79.$$

Tutti i calcoli sono lasciati al lettore. Sostituendo e semplificando si ottiene l'equazione della retta.

$$Y=0.79X+121.52$$



Come si evince dal grafico in questo caso la retta di regressione approssima in modo corretto i dati raccolti.

Conclusioni

Per un lavoro il cui fine era un'indagine sull'utilizzo di device e dispositivi elettronici, proprio un device (Google Drive) si è rivelato fondamentale per la cooperazione tra studenti di classi diverse e i rispettivi professori.

La realizzazione di questa attività è stata possibile grazie all'unione delle conoscenze (sia matematiche che tecnologiche) dell'intero team, che ha permesso di gestire al meglio le tempistiche e di dividersi equamente il lavoro. Inoltre l'utilizzo dei Moduli di Google per l'indagine all'interno del liceo è stato di fondamentale importanza, tanto quanto l'impiego del linguaggio LaTeXper risolvere il problema della scrittura delle formule.

Bibliografia e sitografia

Massimo Bergamini, Anna Trifone, Graziella Barozzi, *Matematica.blu 2.0, capitolo 1 ("La statistica") Vol.3*, Zanichelli editore, 2012.

L. Weinstein, J.A. Adam, *Più o meno quanto?*, Zanichelli, 2009.

L. Lamberti, L. Mereu, A. Nanni, *Nuovo Lezioni di Matematica, vol. C*, ETAS, 2012, Milano.

Oscar Chisini, *Sul concetto di media*, Periodico di Matematiche 4, 1929.

Ilaria Ottino, *Statistica e Probabilità*, Ghisetti e Corvi, 2016

`http://old.sis-statistica.org`, *Di medie non ce n'è una sola*.

Liceo Attilio Bertolucci Editore

Via Toscana 10/a - 43122 Parma
prps05000e@istruzione.it - 0521 798459

© Liceo Attilio Bertolucci Editore
ISBN 978-XXXXXXXXXXX
Editato in Parma, aprile 2018