Calcolo delle Probabilità e Statistica 2021/2022

Statistica descrittiva e teoremi limite

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

1 Introduzione alla Statistica

1.1 L'arte di apprendere dai dati

La raccolta di dati e la loro analisi sono strumenti indispensabili per capire a fondo la complessa realtà che ci circonda. La ***Statistica*** è l'*arte di apprendere dai dati*. Essa si occupa della loro ***raccolta***, della loro ***descrizione*** e della loro ***analisi***, guidandoci nel trarre le conclusioni. La Statistica si suddivide quindi in tre ambiti:

- il ***Campionamento statistico***, che riguarda la *raccolta* dei dati e ha lo scopo di selezionare un *campione* "sufficientemente rappresentativo" dell'insieme completo che viene detto *popolazione*;
- la ***Statistica descrittiva***, che si occupa di *descrivere* e, in particolare, di *sintetizzare* tramite *indici*, *tabelle* e *grafici* i dati in esame;
- la ***Statistica inferenziale***[1], che interviene quando i dati a nostra disposizione provengono da un campione della popolazione (anziché dall'intera popolazione); in particolare, la Statistica inferenziale si occupa di inferire/indurre[2], quindi di trarre delle conclusioni sull'intera popolazione; tali conclusioni sono in primo luogo tratte in modo ingenuo dall'osservazione di indici, tabelle e grafici forniti dalla Statistica descrittiva; in secondo luogo, l'obiettivo è *validare* (basandosi su nozioni di *Calcolo delle probabilità*) le conclusioni a cui siamo giunti a partire dai dati a nostra disposizione.

Per quanto riguarda gli aspetti matematici delle ultime due discipline, la Statistica descrittiva utilizza nozioni matematiche elementari, mentre la Statistica inferenziale necessita di strumenti avanzati di Calcolo delle probabilità, come ad esempio i *Teoremi limite* (in particolare, la *Legge dei grandi numeri* e il *Teorema centrale del limite*).

1.2 Popolazione e campioni

La Statistica è interessata ad ottenere informazioni su un insieme completo di persone o cose che viene detto ***popolazione***, la quale può essere *finita* o *infinita*.

- Esempi di *popolazioni finite* sono gli abitanti di una certa regione, i televisori prodotti da un'azienda, gli studenti di un determinato corso di laurea.
- Un esempio di *popolazione infinita* è data dagli esiti di tutti i possibili lanci di uno stesso dado.

Ogni elemento di questo insieme chiamato popolazione viene detto *unità statistica*. Un sottoinsieme della popolazione è detto ***campione***.

[1] Anche detta *Statistica induttiva* o *Statistica matematica*.

[2] *Indurre*, come anche *inferire*, significa trarre delle conclusioni di carattere generale dall'osservazione dei dati a nostra disposizione.

In Statistica si suppone che le unità statistiche possiedano una o più caratteristiche, dette anche ***caratteri*** o ***variabili***. Ogni carattere/variable corrisponde ad un determinato tipo di dato. Quindi con riferimento ad una stessa popolazione si possono studiare uno o più dati differenti.

La popolazione, anche se finita, è spesso troppo numerosa perché sia possibile esaminarla interamente, facendo quindi un *censimento*. Un'indagine censuaria[3] è infatti generalmente molto costosa e necessita di tempi lunghi, anche se ha il vantaggio di non lasciar spazio a dubbi o incertezze. Per questo motivo, anche nel caso di popolazioni finite ma numerose, si ritiene più vantaggioso esaminare un campione e generalizzare quanto osservato a tutta la popolazione. Risulta perciò fondamentale scegliere un campione che sia il più "rappresentativo" possibile dell'intera popolazione. Questo è il compito del *Campionamento statistico*. La scelta più opportuna del campione viene indicata dalla Statistica inferenziale e, in particolare, dalle ipotesi dei teoremi utilizzati in questo ambito. Una proprietà importante è ad esempio che le unità statistiche del campione siano scelte in maniera *casuale*. La proprietà di *casualità* garantisce infatti l'ipotesi di *indipendenza* richiesta dai teoremi limite.

*1.3 Una breve storia della Statistica**

L'etimologia del termine "statistica" deriva dalla parola "stato" e fa riferimento al fatto che i primi dati ad esser raccolti ed organizzati erano di interesse degli stati. Oggi questo compito è svolto in maniera sistematica dagli istituti centrali di statistica, come ad esempio l'ISTAT in Italia.

I primi censimenti sono stati effettuati sin dall'antichità. Ad esempio, nell'antico Egitto si realizzarono censimenti per motivi fiscali e militari; gli antichi egizi ebbero addirittura una dea *dei libri e dei conti*, la dea Seshat[4]. Tuttavia, la raccolta sistematica di dati riguardanti la popolazione e l'economia avvenne solo a partire dal Rinascimento, in particolare a Venezia e a Firenze. L'idea di raccogliere dati si diffuse in seguito dall'Italia a tutta l'Europa occidentale, ed entro la prima metà del sedicesimo secolo era generalmente diffusa la consuetudine, presso i governi europei, di richiedere alle parrocchie di registrare nascite, matrimoni e morti.

Nel 1662 il commerciante inglese John Graunt pubblicò un libro dal titolo *Natural and Political Observations made upon the Bills of Mortality*, che riscosse notevole successo ed è oggi considerato la prima opera di *Demografia*[5]. Graunt pensò di utilizzare i dati riguardanti i decessi per stimare la popolazione di Londra e, più in generale, l'intera popolazione dell'Inghilterra. Egli può essere dunque considerato l'iniziatore di un nuovo modo di fare statistica, non accontentandosi più di raccogliere ed organizzare i dati, ma cercando di estrapolare nuove informazioni e nuovi numeri da questi dati.

Il lavoro di Graunt sulle tabelle di mortalità ispirò nel 1693 le ricerche dello scienziato Edmond Halley. Halley, lo scopritore dell'omonima cometa, usò le tabelle di mortalità per stabilire con che probabilità una persona di una data età sarebbe vissuta fino ad un determinato numero di anni. Egli utilizzò questi studi per mettere in evidenza che

[3]Esempi di indagini censuarie sono le elezioni politiche o le elezioni amministrative.

[4]Una statua di questa dea sormonta l'ingresso dell'ISTAT a Roma.

[5]In tale opera compare la prima *tavola di mortalità*, che fornisce il numero di decessi per fasce d'età.

il premio di un'assicurazione sulla vita deve dipendere dall'età dell'assicurato. A questi studi si fa risalire l'inizio della *Scienza attuariale*.

Dopo Graunt e Halley, la raccolta di dati si accrebbe stabilmente e a partire dalla fine del diciottesimo secolo una vasta disponibilità di registrazioni censuarie ed altri dati vennero raccolti sistematicamente dai governi dell'Europa occidentale e dagli Stati Uniti.

Durante il diciannovesimo secolo, nonostante il Calcolo delle probabilità fosse già stato sviluppato in varie direzioni, la sua applicazione alla Statistica era praticamente inesistente, dato che molti statistici di quel tempo sostenevano la completa chiarezza ed evidenza dei dati. In particolare, essi non erano tanto interessati a fare inferenza su un campione di dati, cercavano invece di ottenere dati sempre più completi dell'intera popolazione. L'inferenza da un campione alla popolazione era dunque quasi del tutto ignota a quel tempo.

Solo alla fine del diciannovesimo secolo si iniziò ad occuparsi di inferenza. Tra i fautori di questo nuovo indirizzo vanno ricordati Francis Galton, i cui studi sull'ereditarietà introdussero ciò che ora chiamiamo regressione e analisi della correlazione, e Karl Pearson. Pearson fu il primo direttore del Laboratorio Galton, fondato per donazione di Francis Galton nel 1904. Qui Pearson organizzò un programma di ricerca mirato allo sviluppo di nuovi metodi di inferenza. Vi si accoglievano studenti di materie scientifiche ed industriali che venivano ad imparare le tecniche statistiche per poterle applicare nei loro campi. Uno dei primi ricercatori dell'istituto fu William Gosset, un chimico di formazione, che dimostrò la sua devozione a Pearson pubblicando i propri lavori sotto lo pseudonimo[6] di "Student" (a lui si deve l'introduzione della distribuzione t di Student).

I due campi di applicazione di maggiore importanza della Statistica all'inizio del ventesimo secolo erano l'Agricoltura e la Biologia, e ciò era dovuto al personale interesse dello stesso Pearson e di altri nel laboratorio, come pure ai notevoli risultati dello scienziato Ronald Fisher. Tuttavia, la teoria dell'inferenza sviluppata da questi pionieri (tra i quali citiamo anche il figlio di Karl Pearson, Egon, ed il matematico di origini polacche Jerzy Neyman) era abbastanza generale da adattarsi ad un gran numero di problemi applicati. Attualmente i campi di applicazione della Statistica sono innumerevoli. In tutti i quotidiani e le riviste vi sono esempi di Statistica descrittiva. La Statistica inferenziale è divenuta indispensabile anche in Medicina, Fisica, Ingegneria, Contabilità, Scienze aziendali e Marketing, Economia, Meteorologia e in varie altre discipline.

2 Statistica descrittiva

2.1 Introduzione

In questo sezione presentiamo e sviluppiamo la Statistica descrittiva, la branca della Statistica che si occupa di descrivere e sintetizzare tramite indici, tabelle e grafici i dati in esame. Non ci importa in questo momento se i dati provengono dalla popolazione intera o da un campione estratto da essa, né ci poniamo il problema di come sia stato scelto il campione.

[6] Altri sostengono che Gosset non volesse pubblicare con il suo vero nome per timore che i suoi datori di lavoro alla fabbrica di birra Guinness non avrebbero approvato che uno dei loro chimici facesse ricerche di Statistica.

Una prima distinzione che è necessario fare è tra dati (o *caratteri* o *variabili*) *quantitativi* e *qualitativi:*

- i ***dati quantitativi*** sono dati numerici; essi si suddividono a loro volta in:
  - ***discreti*** se possono assumere solo un numero finito o al più un'infinità numerabile di valori;
  - ***continui*** se possono assumere un'infinità continua di valori;
- i ***dati qualitativi*** sono tutti gli altri dati.

Ci sono meno strumenti a disposizione per trattare dati qualitativi. Inoltre, per riuscire a studiare dati qualitativi in maniera efficace è utile che tali dati assumano un numero relativamente basso di valori tra loro distinti (come accade ad esempio nel caso in cui si studia il gruppo sanguigno di un insieme di persone oppure lo stato occupazionale dei laureati di un determinato corso di laurea).

Vediamo infine tre esempi, uno per ogni tipo di dato (o carattere o variabile), a cui faremo riferimento anche in seguito.

Esempio 2.1 (***Dati quantitativi discreti***)**.** *Sono riportati qui di seguito i voti finali in Inglese dei quindici alunni di una classe:*

7 6 8 6 9 7 8 7 8 6 7 6 7 7 6

Esempio 2.2 (***Dati quantitativi continui***)**.** *I risultati (espressi in metri) ottenuti da 22 studenti nella prova di salto in lungo da fermo sono i seguenti:*

1.36 1.46 1.62 1.54 1.94 1.85 1.75 1.88 1.61 1.90 1.65
1.53 1.36 1.67 1.46 1.60 1.50 1.67 1.65 1.78 2.12 1.86

Esempio 2.3 (***Dati qualitativi***)**.** *È stato chiesto in un questionario ai 26 studenti di una classe di indicare con una delle seguenti lettere l'attività a cui dedicano la maggior parte del tempo libero:*

A: amici *S: sport* *C: cinema, TV* *H: hobby* *N: altre attività*

Sono stati ottenuti i seguenti risultati:

A, A, C, H, N, S, S, S, N, A, C, C, H, N, H, S, A, H, A, S, S, A, A, C, A, C.

2.2 Le frequenze: tabelle e grafici

La nozione forse più importante della Statistica descrittiva è quella di ***frequenza***, nelle sue varie forme. I valori di tali frequenze sono riportati innanzitutto in tabelle e poi rappresentati anche tramite opportuni grafici.

2.2.1 Tabelle delle frequenze

La ***frequenza*** (anche detta ***frequenza assoluta***) di un determinato valore è il numero di volte che quel valore è stato osservato. Ecco la tabella delle frequenze (assolute) nel caso degli Esempi 2.1 e 2.3:

Voto	Freq. ass.
6	5
7	6
8	3
9	1

Attività	Freq. ass.
A	8
S	6
C	5
H	4
N	3

A partire dalla frequenza assoluta si ricavano altri tipi di frequenze, in particolare:

- la ***frequenza relativa***, data dal rapporto tra frequenza assoluta e numero totale di dati in esame;
- la ***frequenza percentuale***, che è la frequenza relativa espressa in termini percentuali;
- la ***frequenza*** (***percentuale***) ***cumulata***, definita solo nel caso di *dati quantitativi*, la quale è la somma delle frequenze percentuali dei valori minori o uguali a quello per cui la si sta calcolando. In modo analogo si definiscono la ***frequenza assoluta cumulata*** e la ***frequenza relativa cumulata***.

Sempre con riferimento agli Esempi 2.1 e 2.3, possiamo completare le tabelle precedenti con i valori delle frequenze appena introdotte come riportato qui di seguito (nel caso dell'Esempio 2.3, dato che si tratta di dati qualitativi, non ha senso parlare di frequenza cumulata):

Voto	Freq. ass.	Freq. rel.	Freq. perc.	Freq. cum.
6	5	0.33	33%	33%
7	6	0.4	40%	73%
8	3	0.2	20%	93%
9	1	0.07	7%	100%

Attività	Freq. ass.	Freq. rel.	Freq. perc.
A	8	0.31	31%
S	6	0.23	23%
C	5	0.19	19%
H	4	0.15	15%
N	3	0.12	12%

Si noti che la somma delle frequenze assolute è pari al numero totale di dati osservati, la somma delle frequenze relative è 1, la somma delle frequenze percentuali è 100%.

2.2.2 Classi di frequenza

Nel caso di dati *quantitativi* giocano un ruolo importante le cosiddette ***classi di frequenza***. Quest'ultime sono infatti necessarie nel caso di dati *continui* e utili nel caso di dati *discreti* che assumono un numero *elevato* di valori distinti.

Le *classi di frequenza* sono ***intervalli*** *contigui* della retta reale. Stabilire quali debbano essere le classi da adottare non è una scelta univoca, infatti il *numero di classi* e la loro *ampiezza* possono essere scelti in infiniti modi. Troppe classi rendono la tabella poco leggibile; troppo poche classi la rendono poco significativa: il numero giusto va scelto con buon senso. L'importante è che *ogni dato appartenga ad una e una sola classe.*

Come si costruiscono le classi? Per semplicità noi considereremo solo il caso di classi aventi la ***stessa ampiezza*** (un'ampiezza differente può essere utile nel caso di dati con concentrazione molto variabile). Vediamo come si procede per costruire un insieme di classi di frequenze nel caso dell'Esempio 2.2.

1. Si fissa un intervallo di estremi a e b che *contenga* i dati in esame (in molti casi, a è il più piccolo dato osservato e b il più grande). Nell'Esempio 2.2 scegliamo ad esempio $a = 1.20$ e $b = 2.20$.

2. *Tipologia di intervalli.* È pratica comune scegliere come classi intervalli del tipo $(x, y]$ oppure $[x, y)$, anziché (x, y) o $[x, y]$, in modo tale da avere intervalli tutto dello stesso tipo. L'unica eccezione vale per le classi agli estremi, ad esempio anziché $(a, y]$ si può utilizzare $[a, y]$ se si vuole includere l'estremo a.

3. *Numero di classi.* Come abbiamo già detto, il numero giusto va scelto con buon senso. In genere si consiglia di scegliere non più di dieci[7] classi. Con riferimento all'Esempio 2.2, consideriamo ad esempio *cinque* classi.

4. *Ampiezza delle classi.* L'ampiezza, dato che stiamo supponendo che sia la stessa per tutte le classi, è necessariamente data dal rapporto tra $b - a$ e il *numero di classi.* Nel caso dell'Esempio 2.2, dato che $b - a = 2.20 - 1.20 = 1$ e il numero di classi che abbiamo scelto è 5, l'ampiezza è pari a 0.20.

Con riferimento all'Esempio 2.2, otteniamo dunque la seguente tabella:

Classe	Freq. ass.	Freq. rel.	Freq. perc.	Freq. cum.
$1.20 - 1.40$	2	0.09	9%	9%
$1.40 - 1.60$	6	0.27	27%	36%
$1.60 - 1.80$	8	0.36	36%	72%
$1.80 - 2.00$	5	0.23	23%	95%
$2.00 - 2.20$	1	0.05	5%	100%

In tal caso abbiamo scelto come classi intervalli del tipo $(x, y]$, che è la tipologia che utilizzeremo usualmente. Per indicare che ad esempio la classe $1.40 - 1.60$ corrisponde

[7] A meno che i dati non siano particolarmente sparpagliati.

all'intervallo $(1.40, 1.60]$ si usa a volte la notazione

$$1.40 \text{ ——}| \ 1.60$$

Analogamente, la notazione $1.40 \ |\text{——} \ 1.60$ sta per $[1.40, 1.60)$.

2.2.3 Principali rappresentazione grafiche

Vediamo ora alcune rappresentazioni grafiche che si possono costruire a partire dalle tabelle delle frequenze. Tutto ciò che diremo a questo riguardo richiede due precisazioni.

1) In primo luogo, quelle che vedremo sono solo le *principali* rappresentazione grafiche, che sono dunque da intendersi più come esempi o punti di riferimento utili per quando si dovrà costruire un grafico per un certo insieme di dati. Sottolineiamo che l'obiettivo di ogni rappresentazione grafica è sempre e solo far capire con un colpo d'occhio la distribuzione dei dati in esame. Più precisamente, si noterà due aspetti di ogni grafico: da una parte l'insieme delle "istruzioni", essenziali per sapere come costruirlo, che è necessario siano il più possibile dettagliate e quindi spesso anche molto articolate; d'altra parte, il grafico è pensato per "raccontare" l'insieme dei dati anche ad un profano o comunque ad una persona che non sia a conoscenza delle "istruzioni", quindi deve essere di facile lettura.

2) In secondo luogo, oggi queste rappresentazioni grafiche si realizzano utilizzando opportuni software (ad esempio Excel), che consentono, una volta immessi i dati, di ottenere rapidamente grafici (e indici, che introdurremo più avanti), anche per insiemi numerosi di dati.

Imparare a costruire grafici manualmente ha quindi soprattutto uno scopo didattico, che ha tra gli obiettivi:

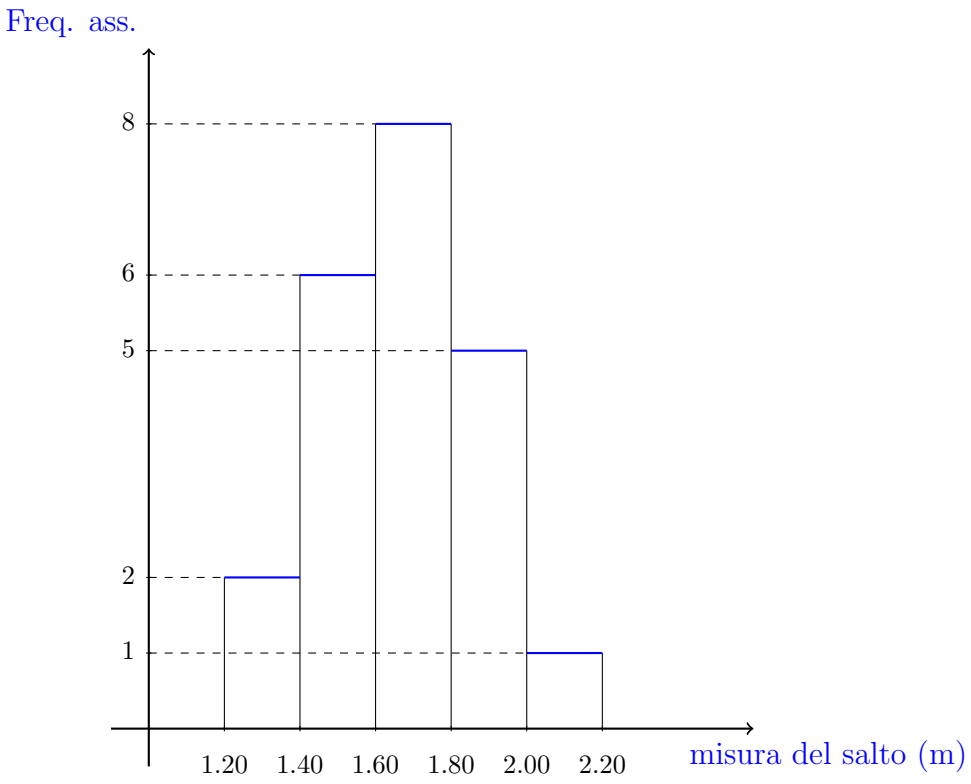
- imparare a leggere e interpretare un grafico;
- capire che tipo di grafico tra quelli noti serve (o, eventualmente, ideare un nuovo grafico) per descrivere nel modo migliore un determinato insieme di dati.

Istogramma. Nel caso di dati *quantitativi* per cui sono state adottate *classi di frequenza*, è possibile costruire il relativo ***istogramma*** (per le frequenze *assolute* oppure *relative* oppure *percentuali*). L'istogramma è costruito mediante rettangoli adiacenti,

1) le cui basi sono gli intervalli che definiscono le classi,

2) le cui altezze, nel caso di classi aventi la *stessa ampiezza*, sono uguali alle corrispondenti frequenze (assolute o relative o percentuali).

Con riferimento all'Esempio 2.2, otteniamo il seguente istogramma (delle frequenze *assolute*):

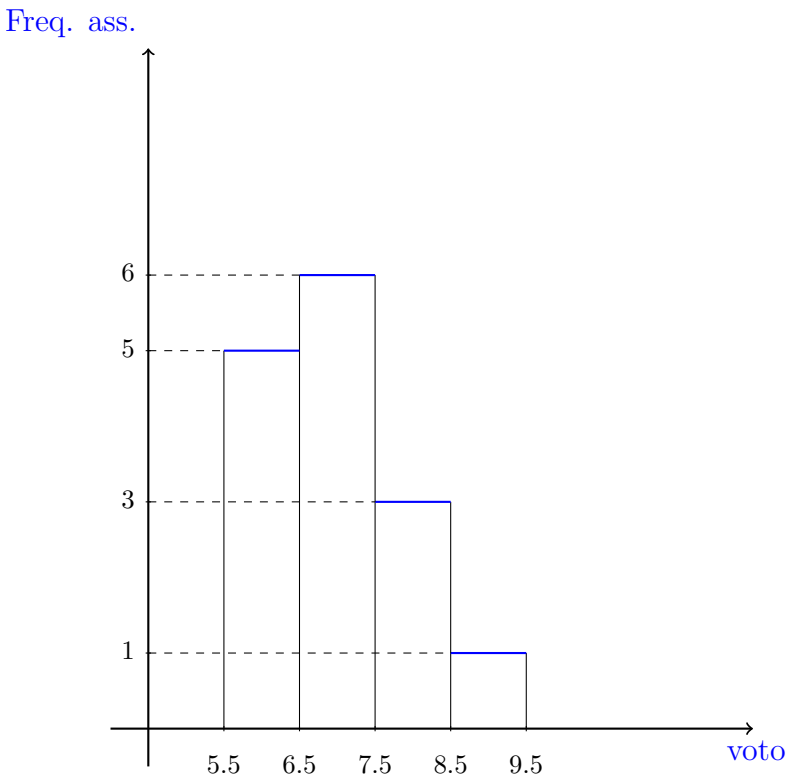


Ortogramma. Nel caso di dati *qualitativi*, o anche di dati quantitativi e discreti in cui non si utilizzano le classi di frequenze (si pensi all'Esempio 2.1), non è possibile costruire l'istogramma.

OSSERVAZIONE. *Chiaramente nel caso di dati quantitativi e discreti è comunque sempre possibile introdurre delle classi di frequenze e, quindi, costruire l'istogramma. Vediamo un esempio a questo proposito. Nell'Esempio 2.1 è possibile scegliere come classi gli intervalli di ampiezza unitaria che sono centrati nei valori assunti dai dati, quindi:* $(5.5, 6.5]$, $(6.5, 7.5]$, $(7.5, 8.5]$, $(8.5, 9.5]$. *Otteniamo dunque la seguente tabella:*

Classe	*Freq. ass.*	*Freq. rel.*	*Freq. perc.*	*Freq. cum.*
$5.5 - 6.5$	5	0.33	33%	33%
$6.5 - 7.5$	6	0.4	40%	73%
$7.5 - 8.5$	3	0.2	20%	93%
$8.5 - 9.5$	1	0.07	7%	100%

L'istogramma è allora il seguente:

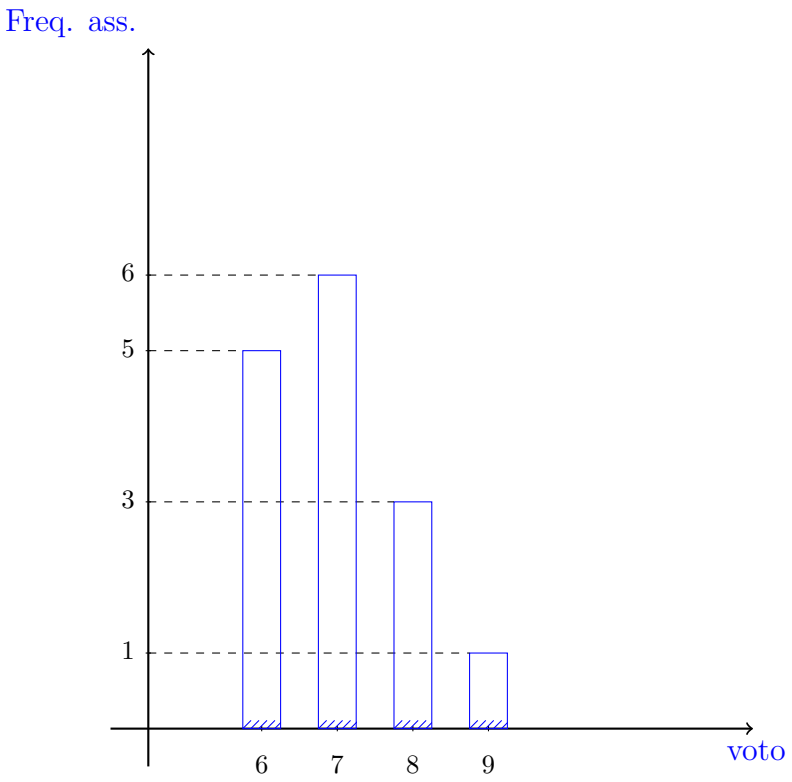


Come spiegato qui di seguito, in questo caso si può costruire anche il cosiddetto ortogramma senza bisogno di introdurre le classi di frequenza.

Consideriamo dunque il caso di dati qualitativi oppure discreti senza classi di frequenza. In tal caso si può visualizzare la distribuzione di frequenza (assoluta o relativa o percentuale) con un ***ortogramma***, che ha le seguenti caratteristiche:

- sull'asse delle ascisse si riportano i valori assunti dai dati, che nel caso di dati qualitativi non sono nemmeno numeri (a volte si preferisce riportare tali valori sull'asse delle ordinate, specialmente nel caso di dati qualitativi); chiaramente, nel caso di dati qualitativi, l'ordine in cui sono disposti i valori non ha alcuna rilevanza;
- ad ogni classe corrisponde un rettangolo, la cui base (uguale per tutte le classi) non ha alcun significato;
- l'altezza di ogni rettangolo indica la frequenza (assoluta o relativa o percentuale) del valore corrispondente;
- i rettangoli non si disegnano adiacenti, per ricordare che ogni rettangolo si riferisce non all'intervallo che costituisce la base ma ad un singolo valore (ciò rende anche più facile identificare a quale valore corrisponde ciascun rettangolo, quindi migliora la leggibilità).

Nell'Esempio 2.1 l'ortogramma è il seguente:



Come abbiamo già detto, certe volte si preferisce riportare i valori assunti dai dati sull'asse delle ordinate (specialmente nel caso di dati qualitativi), come nel seguente caso in cui è riportato l'ortogramma relativo all'Esempio 2.3:

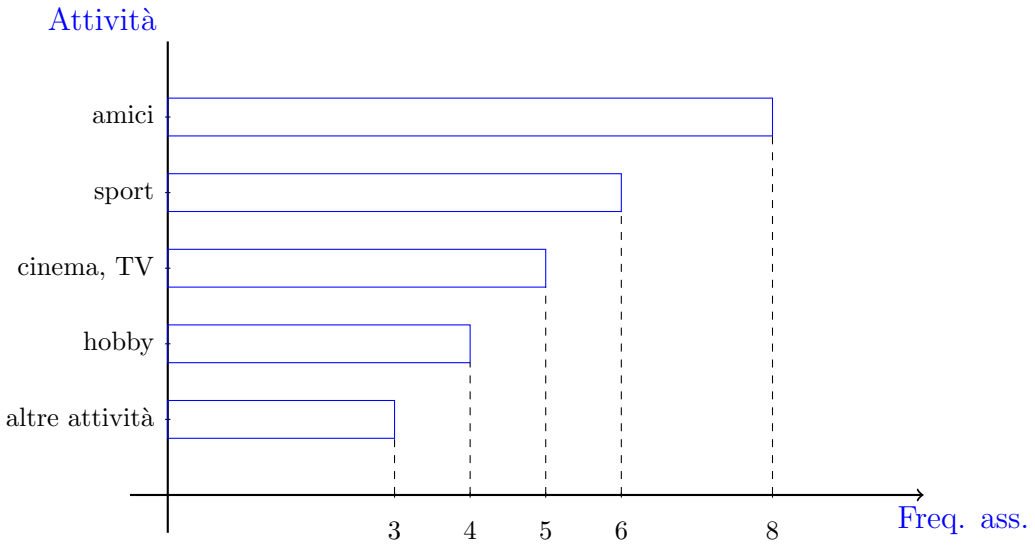


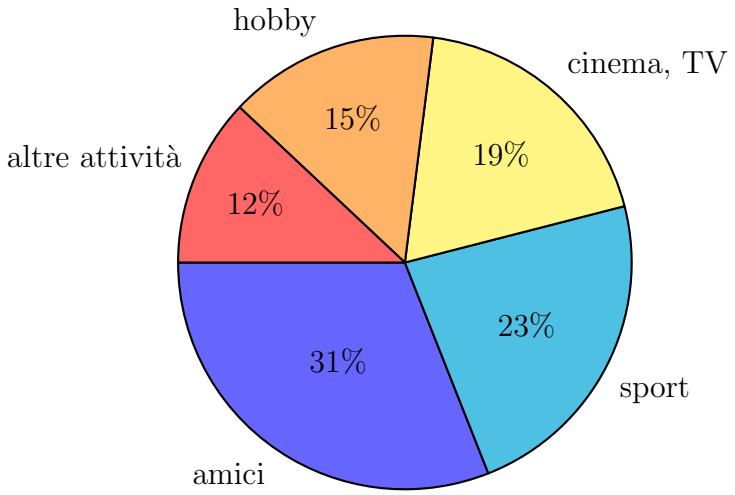
Diagramma a torta. Il ***diagramma a torta***, detto anche ***diagramma circolare*** o ***aerogramma***, è particolarmente utile per rappresentare le *frequenze percentuali*. Si può usare in tutti i casi, sia per dati qualitativi che quantitativi, in quest'ultimo caso sia nel caso in cui ci siano classi di frequenze o anche in loro assenza. Queste sono le caratteristiche del diagramma a torta:

- un cerchio viene suddiviso in tanti settori circolari, ognuno dei quali corrisponde a un valore (o ad una classe, nel caso in cui si considerino classi di frequenze);
- gli angoli al centro dei diversi settori hanno ampiezza proporzionale alle frequenze percentuali.

Consideriamo l'Esempio 2.3 riguardante le attività svolte nel tempo libero. Per determinare l'ampiezza x del settore corrispondente alla frequenza 31% scriviamo la proporzione:

$$x : 360^\circ = 31 : 100$$

da cui otteniamo che il settore ha angolo al centro $x = 111.6^\circ$. Allo stesso modo si ricavano le ampiezze degli altri settori, da cui otteniamo il seguente diagramma a torta:



2.3 Indici

D'ora in avanti consideriamo solo dati (o variabili o caratteri) quantitativi. Ricordiamo che la Statistica descrittiva cerca di sintetizzare i dati in esame non solo attraverso tabelle e grafici, ma anche tramite singole quantità numeriche, gli ***indici***, chiamati anche ***statistiche***. Queste sono grandezze calcolate a partire dai dati ed essenzialmente si suddividono in due grandi famiglie:

- gli ***indici di posizione*** (***centrale***), che hanno come obiettivo individuare il *"centro"* dell'insieme dei dati, a cui è dedicata la presente sezione;
- gli ***indici di dispersione*** o ***variabilità***, che hanno come obiettivo *quantificare* la dispersione dei dati attorno ad un determinato "centro", a cui è dedicata la prossima sezione.

Per poter definire tali indici è utile introdurre qualche notazione, ovvero denotare con n il numero totale di dati in esame e indicare con $x_1, x_2, \ldots, x_n$ i dati stessi. Ricordiamo che tali dati potrebbero provenire dalla popolazione intera o da un campione estratto da essa. Come abbiamo già detto, ciò non è rilevante. Notiamo solamente che se i dati provengono da un campione allora spesso si specifica che gli indici sono *campionari*, cioè riferiti al campione in esame.

2.3.1 Indici di posizione: media

Gli indici di posizione forniscono il "centro" dell'insieme dei dati in esame. Tuttavia non vi è una definizione univoca di cosa si intenda per "centro", per questa ragione numerosi indici di posizione sono stati introdotti. Vediamo il più importante, ovvero la ***media***, che corrisponde alla *media aritmetica* dei dati.

Definizione 2.1. *Si dice* ***media*** *(o anche* ***media campionaria*** *quando i dati provengono da un campione) di n dati $x_1, x_2, \ldots, x_n$ e si denota con $\bar{x}$ (o anche $\bar{x}_n$) la quantità*

$$\bar{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Osservazione. *Nel caso di dati* discreti *è possibile esprimere la media tramite una formula che fa intervenire le frequenze. Siano $v_1, v_2, \ldots, v_k$ i k valori assunti dai dati (qui stiamo usando il fatto che i dati sono discreti) e siano $n_1, n_2, \ldots, n_k$ le corrispondenti frequenze assolute. Poiché il numero complessivo di dati è $n = \sum_{j=i}^{k} n_j$ e per $j = 1, \ldots, k$ il valore v_j compare n_j volte nell'insieme di dati, segue che*

$$\bar{x} \;=\; \frac{1}{n}\sum_{j=1}^{k} v_j\, n_j.$$

Se si riscrive l'ultima formula come (indichiamo con $f_1, f_2, \ldots, f_k$ le frequenze relative)

$$\bar{x} \;=\; v_1\,\frac{n_1}{n} + v_2\,\frac{n_2}{n} + \cdots + v_k\,\frac{n_k}{n} \;=\; v_1\,f_1 + v_2\,f_2 + \cdots + v_k\,f_k,$$

si può notare come la media non sia altro che una media pesata *dei valori assunti dai dati. Ogni valore ha come peso la sua* frequenza relativa*, che è pari al rapporto tra la sua frequenza assoluta ed n.*

2.3.2 Indici di dispersione

Una questione di particolare importanza è quanto i dati siano concentrati o viceversa dispersi attorno al valore centrale. Gli ***indici di dispersione*** o ***variabilità*** hanno come obiettivo proprio *quantificare* tale concentrazione/dispersione dei dati.

Consideriamo ad esempio le due sequenze di dati:

$$\begin{array}{rrrrrrrr} 8, & 16, & 21, & 29, & 37, & 49, & 57; \\ 27, & 28, & 28, & 30, & 32, & 34, & 38. \end{array}$$

Entrambe hanno lo stesso numero di dati e la stessa media pari a 31. Tuttavia, la distribuzione dei dati intorno al valor medio 31 è diversa per le due sequenze: i dati della seconda sequenza sono maggiormente concentrati attorno alla media, mentre quelli della prima sono più sparsi. Per misurare questa differente ***dispersione*** o ***variabilità*** sono stati introdotti diversi indici, di cui ora vediamo il più importante.

Varianza. Tra gli indici di variabilità, i più importanti sono quelli che usano come valore centrale la *media*. Fra questi indici il più noto è la ***varianza***. Prima di darne la definizione è utile introdurre alcune notazioni. Si chiama *scarto* (o *deviazione*) di un dato dalla media la quantità

$$x_i - \bar{x}.$$

Si chiamano rispettivamente *scarto assoluto* (o *deviazione assoluta*) e *scarto quadratico* (o *deviazione quadratica*) le quantità

$$|x_i - \bar{x}|, \qquad\qquad (x_i - \bar{x})^2.$$

Gli indici che misurano lo scarto dalla media si ottengono *sommando* gli scarti (assoluti o quadratici) dalla media stessa.

Definizione 2.2. *Si dice* ***varianza*** *(o anche* ***varianza campionaria*** *quando i dati provengono da un campione) di n dati $x_1, x_2, \ldots, x_n$ e si denota con*[a] *s^2 (o anche s_n^2) la quantità*

$$s^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

La radice quadrata della varianza si chiama ***deviazione standard*** *(o* ***scostamento quadratico medio****):*

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

[a] Alcuni autori preferiscono indicare la varianza di un insieme di dati con σ^2 anziché s^2. Qui viene invece preferita la notazione s^2 per distinguere la varianza di un insieme di dati dalla varianza di una variabile aleatoria (che è stata indicata con σ^2).

Esempio 2.4. *Consideriamo la seguente sequenza di dati:*

$$5, \quad 6, \quad 14, \quad 15, \quad 17, \quad 20, \quad 31, \quad 36.$$

La media è 18*. Per ogni dato calcoliamo lo scarto quadratico:*

$$\begin{array}{llll} (5-18)^2 = 169, & (6-18)^2 = 144, & (14-18)^2 = 16, & (15-18)^2 = 9, \\ (17-18)^2 = 1, & (20-18)^2 = 4, & (31-18)^2 = 169, & (36-18)^2 = 324. \end{array}$$

La varianza è dunque data da

$$s^2 \;=\; \frac{169+144+16+9+1+4+169+324}{8} \;=\; 104.5.$$

Calcolando la radice quadrata otteniamo la deviazione standard

$$s \;=\; \sqrt{104.5} \;\approx\; 10.22.$$

OSSERVAZIONE. *Come per la media, nel caso di dati* discreti *è possibile esprimere la varianza in termini delle frequenze. Siano $v_1, v_2, \ldots, v_k$ i k valori assunti dai dati e siano $n_1, n_2, \ldots, n_k$ le corrispondenti frequenze assolute. Poiché il numero complessivo di dati è $n = \sum_{j=i}^{k} n_j$ e per $j = 1, \ldots, k$ il valore v_j compare n_j volte nell'insieme di dati, segue che la varianza degli n dati è anche data da*

$$s^2 \;=\; \frac{1}{n}\sum_{j=1}^{k}(v_j - \bar{x})^2\, n_j \;=\; \sum_{j=1}^{k}(v_j - \bar{x})^2\, f_j,$$

dove $f_1, f_2, \ldots, f_k$ sono le frequenze relative (in particolare, vale che $f_j = n_j/n$).

Esempio 2.5. *Consideriamo la seguente tabella relativa a un insieme di* 10 *dati che assumono i valori* 3, 5, 7 *e* 12*:*

valori	*freq. ass.*	*freq. rel.*
3	2	0.2
5	1	0.1
7	3	0.3
12	4	0.4

La media è data da (in questo caso $k = 4$)

$$\bar{x} \;=\; \sum_{j=1}^{4} v_j\, f_j \;=\; 8.$$

Quindi la varianza vale

$$s^2 \;=\; \sum_{j=1}^{4}(v_j - 8)^2\, f_j \;=\; 12.6.$$

Infine, la deviazione standard è $s = \sqrt{12.6} \approx 3.55$.

Per calcolare la varianza è spesso preferibile utilizzare la formula seguente.

Teorema 2.1. *La varianza di un insieme di n dati $x_1, x_2, \ldots, x_n$ è data dalla seguente formula alternativa:*

$$s^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Dimostrazione. Infatti, vale che

$$\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) \;=\; \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \;=\; \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2.$$

□

3 Dati bivariati: correlazione e regressione

In questa sezione affrontiamo lo studio della Statistica descrittiva nel caso di dati *multidimensionali* e, in particolare, *bidimensionali* (anche detti *bivariati*). In molte situazioni siamo infatti interessati a *due* o *più* dati (o caratteri o variabili) riguardanti la stessa unità statistica, ovvero abbiamo a che fare con un insieme di dati *vettoriali* (o *multidimensionali* o *multivariati*). Noi ci concentreremo sul caso *bidimensionale* (o *bivariato*), per il quale è possibile fornire rappresentazioni sia tabellari che grafiche. Tuttavia, molti dei concetti introdotti in questo capitolo possono essere facilmente estesi al caso multidimensionale.

3.1 Insiemi di dati bivariati e frequenza congiunta

Come nel caso unidimensionale, la nozione fondamentale è quella di ***frequenza***, che ora definiamo nel caso di dati bivariati. Per semplicità, nella presente sezione, supponiamo che i dati siano *quantitativi* e *discreti* e che assumano un numero *finito* di valori distinti. Vediamo innanzitutto un esempio di dati bivariati.

Esempio 3.1. *Nella seguente tabella sono riportati i voti in italiano e in matematica di tredici studenti.*

Numero d'ordine	*Voto in italiano*	*Voto in matematica*
1	6	8
2	6	6
3	7	6
4	6	6
5	8	7
6	7	7
7	7	6
8	8	8
9	9	8
10	6	7
11	6	6
12	6	6
13	7	6

Consideriamo in generale un insieme di dati bivariati di numerosità n, ovvero un insieme di n coppie di numeri reali, e indichiamo con X e Y i due *tipi di dati* (o *caratteri* o *variabili*). Nell'Esempio 3.1 ci sono $n = 13$ coppie date da $(6, 8), (6, 6), (7, 6), (6, 6), (8, 7), \ldots,$ $(7, 6)$, una per ogni studente. Inoltre

$$\begin{aligned} X &= \text{“voto in italiano”}, \\ Y &= \text{“voto in matematica”}. \end{aligned}$$

Siano h ed k uguali al numero di valori *distinti* assunti rispettivamente dal primo e dal secondo carattere. Nell'Esempio 3.1 i voti in italiano assumono i valori $6, 7, 8, 9$, mentre i voti in matematica assumono solo i valori $6, 7, 8$, quindi $h = 4$ e $k = 3$.

Infine, siano $v_1, \ldots, v_h$ e $w_1, \ldots, w_k$ i valori *distinti* assunti rispettivamente dal primo e dal secondo carattere. Nell'Esempio 3.1 abbiamo che

$$\begin{array}{llllllll} v_1 & = & 6, & v_2 & = & 7, & v_3 & = & 8, & v_4 & = & 9, \\ w_1 & = & 6, & w_2 & = & 7, & w_3 & = & 8. \end{array}$$

Allora, per ogni coppia (v_i, w_j) definiamo:

- la ***frequenza assoluta congiunta*** della coppia (v_i, w_j), indicata con n_{ij} e pari al numero di volte in cui la coppia (x_i, y_j) compare nell'insieme di dati in esame;
- la ***frequenza relativa congiunta*** della coppia (v_i, w_j), indicata con f_{ij} e data da

$$f_{ij} \;=\; \frac{n_{ij}}{n}.$$

- la ***frequenza percentuale congiunta*** della coppia (v_i, w_j), indicata con p_{ij} e pari alla frequenza relativa espressa in termini percentuali.

Tali frequenze possono essere riportate in forma tabellare tramite una ***tabella a doppia entrata***, il cui corpo centrale è costituito dalla *matrice* $h \times k$ delle frequenze congiunte. Ad esempio, la tabella a doppia entrata delle frequenze *assolute* è la seguente:

X \ Y	w_1	w_2	$\cdots$	w_k	n^X
v_1	n_{11}	n_{12}	$\cdots$	n_{1k}	n_1^X
v_2	n_{21}	n_{22}	$\cdots$	n_{2k}	n_2^X
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
v_h	n_{h1}	n_{h2}	$\cdots$	n_{hk}	n_h^X
n^Y	n_1^Y	n_2^Y	$\cdots$	n_k^Y	n

Ai margini della tabella compaiono le frequenze assolute dei due caratteri X e Y, che si chiamano appunto *frequenze assolute* ***marginali***. Le marginali relative al carattere X si ottengono sommando le frequenze congiunte che compaiono sulla stessa riga, mentre le marginali relative al carattere Y si ottengono sommando i valori che compaiono sulla stessa colonna. Infine, sommando i valori dell'ultima colonna si ottiene n. Analogamente, sommando i valori dell'ultima riga si ottiene ancora n. Questo spiega la presenza del numero n nell'angolo in basso a destra della tabella.

La tabella a doppia entrata delle frequenza assolute congiunte relativa all'Esempio 3.1 è la seguente:

Voto in italiano \ Voto in matematica	6	7	8	Totale
6	4	1	1	6
7	3	1	0	4
8	0	1	1	2
9	0	0	1	1
Totale	7	3	3	13

In modo analogo si costruiscono le tabelle a doppia entrata delle frequenze *relative* congiunte e delle frequenze *percentuali* congiunte. Ad esempio, la tabella a doppia entrata delle frequenze *relative* congiunte dell'Esempio 3.1 è la seguente:

Voto in italiano \ Voto in matematica	6	7	8	Totale
6	0.307	0.077	0.077	0.461
7	0.231	0.077	0	0.307
8	0	0.077	0.077	0.154
9	0	0	0.077	0.077
Totale	0.538	0.231	0.231	1

La tabella a doppia entrata delle frequenze *percentuali* congiunte è invece la seguente:

Voto in italiano \ Voto in matematica	6	7	8	Totale
6	30.7%	7.7%	7.7%	46.1%
7	23.1%	7.7%	0%	30.7%
8	0%	7.7%	7.7%	15.4%
9	0%	0%	7.7%	7.7%
Totale	53.8%	23.1%	23.1%	100%

3.2 Dipendenza: diagramma a dispersione e correlazione

Il motivo per cui si studiano i dati in maniera congiunta, quindi come dati bivariati o più in generale multivariati, è per studiare la *dipendenza* tra i vari caratteri. L'esistenza di una *dipendenza* significa, in termini matematici, che esiste una funzione f tale che

$$y_i \;=\; f(x_i), \qquad \forall\, i = 1, \ldots, n,$$

dove $\{(x_i, y_i) \colon i = 1, \ldots, n\}$ è l'insieme di dati bivariati in esame (che supporremo quantitativi, discreti o continui). Come vedremo, nella maggior parte dei casi è sostanzialmente impossibile dall'osservazione dei dati trovare una tale funzione. Ciò che si riesce a fare è generalmente trovare una funzione $\hat{f}$ tale che

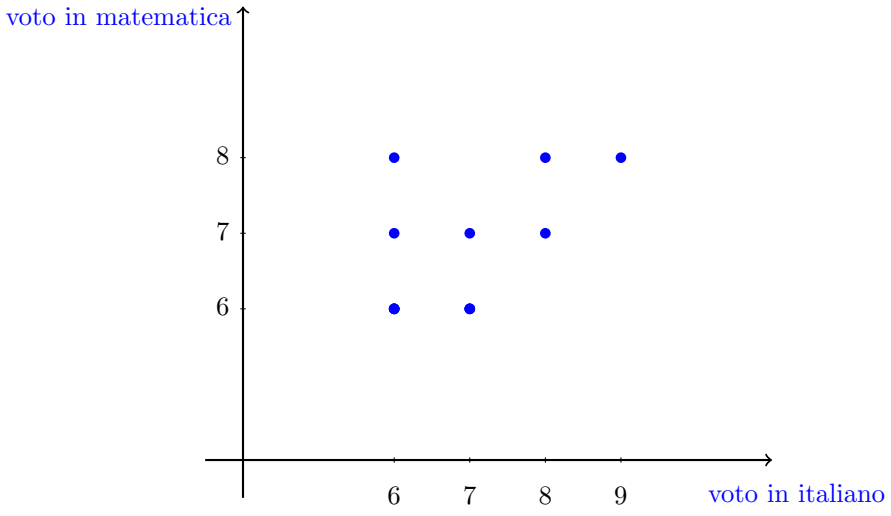
$$y_i \;\approx\; \hat{f}(x_i), \qquad \forall\, i = 1, \ldots, n,$$

ovvero in grado di descrivere la relazione tra i dati osservati in modo *approssimativo*[8] anche se comunque ***significativo***.

Per analizzare la dipendenza introduciamo uno strumento grafico, il *diagramma a dispersione*, e un indice, il *coefficiente di correlazione lineare*.

3.2.1 Diagramma a dispersione

Uno strumento particolarmente utile per studiare la *dipendenza* tra due caratteri, almeno in un primo momento, è il ***diagramma a dispersione*** (o ***diagramma di dispersione*** o ***nuvola di punti***). Tale diagramma è la rappresentazione sul piano cartesiano tramite punti di tutte le coppie che costituiscono l'insieme dei dati. Ad ogni unità statistica appartenente all'insieme di dati in esame corrisponde dunque un punto nel piano cartesiano. Ad esempio, con riferimento all'Esempio 3.1 abbiamo il seguente diagramma a dispersione:



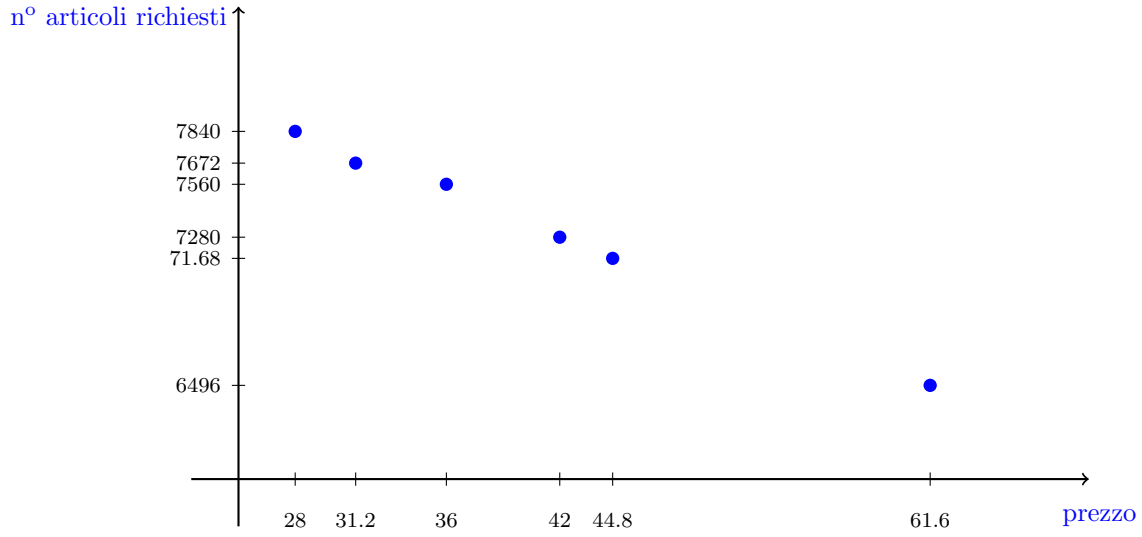
Come si vede dal grafico, è impossibile che esista una funzione f tale che $y_i = f(x_i)$, per $i = 1, \ldots, n$, infatti in corrispondenza della stessa ascissa vengono assunti valori differenti. Tuttavia, come abbiamo già detto, l'obiettivo è determinare una funzione $\hat{f}$ tale che $y_i \approx \hat{f}(x_i)$, per ogni $i = 1, \ldots, n$. In questo caso sembra comunque molto difficile che esista una dipendenza tra i due caratteri, anche se si intuisce una lieve *correlazione positiva* (di cui parleremo nella prossima sezione). Consideriamo invece il seguente esempio.

[8] In certi casi, il fatto che la funzione $\hat{f}$ non verifichi esattamente l'uguaglianza può essere semplicemente dovuto a errori di misurazione dei dati nella fase sperimentale.

Esempio 3.2. *Nella seguente tabella sono riportati i prezzi di sei prodotti e la loro richiesta sul mercato.*

Numero d'ordine	*Prezzo*	*n° articoli richiesti*
1	28	7840
2	31.2	7672
3	36	7560
4	42	7280
5	44.8	7168
6	61.6	6496

Il diagramma a dispersione relativo all'Esempio 3.2 è il seguente:



In questo caso, il grafico suggerisce l'esistenza di una relazione tra i due caratteri, che sembra essere approssimativamente lineare decrescente.

3.2.2 Coefficiente di correlazione lineare

Per ottenere una misura quantitativa della *dipendenza* tra due caratteri, introduciamo una nuova statistica: il ***coefficiente di correlazione lineare***. Tale indice è stato proposto per la prima volta da Francis Galton e formalizzato da Karl Pearson (per tale ragione a volte viene anche chiamato ***coefficiente di correlazione di Pearson***). Tale indice misura se esista una dipendenza tra i due caratteri e se tale dipendenza possa essere descritta in maniera significativa da una relazione di tipo *lineare*. Per definire tale indice è utile introdurre le nozioni di *correlazione positiva* e *negativa:*

- si parla di ***correlazione positiva*** o ***diretta*** quando se uno dei due caratteri assume un valore grande (risp. piccolo) allora anche l'altro carattere assume un valore grande (risp. piccolo);

- si parla di ***correlazione negativa*** o ***inversa*** quando se uno dei due caratteri assume un valore grande (risp. piccolo) allora l'altro carattere assume un valore piccolo (risp. grande).

La definizione di coefficiente di correlazione lineare necessita di alcune notazioni. Consideriamo un insieme di dati bivariati $\{(x_i, y_i)\colon i = 1, \dots, n\}$ quantitativi (discreti o continui). Siano $\bar{x}$ e $\bar{y}$ le media aritmetiche relative ai due caratteri. Allo scopo di misurare la correlazione positiva/negativa, è naturale considerare per ciascuna coppia (x_i, y_i) il prodotto degli *scarti* dalle rispettive medie:

$$(x_i - \bar{x})\,(y_i - \bar{y}).$$

Infatti, se x_i è un valore grande rispetto a quelli tipici, significa che $x_i - \bar{x}$ è positivo. Al contrario, se x_i è piccolo allora $x_i - \bar{x}$ è negativo. Analoghe considerazioni valgono per lo scarto $y_i - \bar{y}$. Quindi, se consideriamo il prodotto $(x_i - \bar{x})(y_i - \bar{y})$, esso sarà:

- maggiore di zero per le coppie (x_i, y_i) in cui x_i e y_i sono correlati positivamente;
- minore di zero per quelle coppie in cui vi è correlazione negativa.

In conclusione, se tra i due caratteri esiste una correlazione positiva/negativa c'è da aspettarsi che la media aritmetica dei prodotti degli scarti di tutte le coppie

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) \tag{3.1}$$

lo percepisca, assumendo di conseguenza un valore positivo/negativo, tanto più grande in valore assoluto quanto maggiore è tale correlazione. La media aritmetica (3.1) prende il nome di *covarianza*, mentre la sua normalizzazione si chiama *coefficiente di correlazione lineare*.

Definizione 3.1. *Sia dato un insieme di dati bivariati $\{(x_i, y_i)\colon i = 1, \dots, n\}$, con medie aritmetiche $\bar{x}, \bar{y}$ e deviazioni standard s_x, s_y, rispettivamente.*
Si dice ***covarianza*** *(o anche* ***covarianza campionaria*** *quando i dati provengono da un campione) e si denota con s_{xy} la quantità*

$$s_{xy} \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}).$$

Supponiamo ora che s_x e s_y siano entrambi diverse da zero.
Si dice ***coefficiente di correlazione lineare*** *(o anche* ***coefficiente di correlazione lineare campionario*** *quando i dati provengono da un campione) e si denota con r_{xy} la quantità*

$$r_{xy} \;=\; \frac{s_{xy}}{s_x s_y}.$$

Osservazione 1. *Dalla definizione segue direttamente la proprietà di* simmetria: $r_{xy} = r_{yx}$. *Inoltre r_{xx} vale* 1, *infatti $s_{xx} = s_x^2$.*

OSSERVAZIONE 2. *Due caratteri si dicono* ***incorrelati*** *se* $r_{xy} = 0$.

Una formula utile per calcolare la covarianza è la seguente, che generalizza la formula analoga ottenuta per la varianza.

Teorema 3.1. *La covarianza di un insieme di* n *dati bivariati* (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ *è data dalla seguente formula alternativa:*

$$s_{xy} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i - \bar{x}\,\bar{y}.$$

Dimostrazione. Infatti, vale che

$$\begin{aligned}\sum_{i=1}^{n}(x_i y_i - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y}) \;&=\; \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y} \\ &=\; \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}.\end{aligned}$$

□

Il coefficiente di correlazione lineare possiede due importanti proprietà che sono enunciate nel seguente risultato.

Teorema 3.2. *Il coefficiente di correlazione lineare verifica le seguenti proprietà.*

1) $-1 \le r_{xy} \le 1$.

2) $r_{xy} = \pm 1$ *se e solo se esistono due costanti reali* $m \neq 0$ *e* q *tali che*

$$y_i \;=\; m\,x_i + q, \qquad \forall\, i = 1, \ldots, n.$$

Inoltre, il segno di r_{xy} *è uguale al segno di* m.

Dimostrazione*. Per quanto riguarda 1), notiamo innanzitutto che le due disuguaglianze $-1 \le r_{xy} \le 1$ sono equivalenti a dire che

$$s_{xy}^2 \;\le\; s_x^2\, s_y^2,$$

ovvero che il quadrato della covarianza è minore o uguale al prodotto delle varianze. Questo segue da un'applicazione diretta della seguente disuguaglianza (nota come disuguaglianza di Cauchy-Schwarz):

$$\left(\sum_{i=1}^{n} a_i\, b_i\right)^2 \;\le\; \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right). \tag{3.2}$$

Nel nostro caso è sufficiente applicare (3.2) con $a_i := \frac{x_i - \bar{x}}{\sqrt{n}}$ e $b_i := \frac{y_i - \bar{y}}{\sqrt{n}}$.
Una dimostrazione della disuguaglianza (3.2) è la seguente. Consideriamo il polinomio di

secondo grado

$$f(x) \;=\; \Big(\sum_{i=1}^{n} a_i^2\Big)x^2 - 2\Big(\sum_{i=1}^{n} a_i b_i\Big)x + \sum_{i=1}^{n} b_i^2 \;=\; \sum_{i=1}^{n}(a_i x - b_i)^2. \tag{3.3}$$

Poiché $f(x) \geq 0$ per ogni $x \in \mathbb{R}$, segue che il discriminante è negativo, ossia

$$\Big(\sum_{i=1}^{n} a_i\, b_i\Big)^2 \;\leq\; \Big(\sum_{i=1}^{n} a_i^2\Big)\Big(\sum_{i=1}^{n} b_i^2\Big),$$

che corrisponde alla disuguaglianza (3.2).

Resta da dimostrare 2). Consideriamo ancora la disuguaglianza di Cauchy-Schwarz (3.2). Notiamo che tale disuguaglianza è un'uguaglianza se e solo se

a) $a_i = 0$, $\forall\, i = 1, \ldots, n$, oppure

b) il polinomio (3.3) ha discriminante uguale a zero, ovvero ha una radice reale (di molteplicità due); ricordando che

$$f(x) \;=\; \sum_{i=1}^{n}(a_i x - b_i)^2,$$

segue che $m \in \mathbb{R}$ è una radice di tale polinomio se e solo se

$$b_i \;=\; m\, a_i, \qquad \forall\, i = 1, \ldots, n. \tag{3.4}$$

Poiché, nel nostro caso, $a_i = \frac{x_i - \bar{x}}{\sqrt{n}}$ e $b_i = \frac{y_i - \bar{y}}{\sqrt{n}}$, possiamo riscrivere a) e b) come segue:

a) $x_i = \bar{x}$, $\forall\, i = 1, \ldots, n$,

b) esiste $m \in \mathbb{R}$ tale che

$$\frac{y_i - \bar{y}}{\sqrt{n}} \;=\; m\,\frac{x_i - \bar{x}}{\sqrt{n}}, \qquad \forall\, i = 1, \ldots, n.$$

Si noti che il punto a) è impossibile dato che $s_x \neq 0$. Al contrario, il punto b) corrisponde al punto 2) nell'enunciato del teorema, infatti vale che

$$y_i \;= m\, x_i + \bar{y} - m\,\bar{x}, \qquad \forall\, i = 1, \ldots, n$$

Infine, essendo $s_y \neq 0$ segue che $m \neq 0$. □

Osservazione 1. *Il valore assoluto di r_{xy} è una misura della* ***significatività*** *di una relazione lineare (come ad esempio la* retta di regressione *di cui parleremo nella prossima sezione) nel descrivere la dipendenza tra i due caratteri (il segno di r_{xy} fornisce semplicemente la pendenza della retta). Infatti:*

- *quando $|r_{xy}| = 1$ vi è una relazione lineare perfetta (con riferimento al campione in esame, non necessariamente alla popolazione intera), dato che i punti del diagramma a dispersione sono disposti lungo una retta, come affermato nel Teorema 3.2;*

- *se ad esempio* $|r_{xy}| = 0.8$ *allora, nonostante i punti del diagramma non siano tutti disposti lungo una stessa retta, è possibile trovare una relazione lineare (ad esempio la* retta di regressione*) che non passa troppo distante dai vari punti;*

- *se invece ad esempio* $|r_{xy}| = 0.3$ *allora non esiste alcuna relazione lineare in grado di descrivere in maniera significativa la dipendenza tra i due caratteri.*

Si noti infine che $|r_{xy}|$ *fa riferimento solo a relazioni di tipo* ***lineare*** *(da cui coefficiente di correlazione* lineare*). Anche nel caso estremo* $r_{xy} = 0$ *significa solamente che non è assolutamente opportuno utilizzare una relazione lineare per descrivere la dipendenza tra i due caratteri. Ciò non esclude che si possa trovare una relazione* non lineare *che risulti particolarmente significativa, come ad esempio* $y = \sin(x)$ *oppure* $y = x^3$*.*

Osservazione 2. *Per quanto detto nell'osservazione precedente, potrebbe sembrare di poca utilità il coefficiente* r_{xy} *dato che in molte situazioni reali si ha che fare con relazioni* non lineari*. In realtà, supponiamo ad esempio di sapere, da precedenti analisi scientifiche, che questa è la relazione che ci aspettiamo:*

$$y \;=\; \log(1 + \sin(x)).$$

Tale relazione chiaramente è non lineare. Tuttavia, ponendo

$$z_i \;:=\; \log(1 + \sin(x_i)), \qquad \forall\, i = 1, \ldots, n,$$

notiamo che la relazione che ci aspettiamo tra le variabili y *e* z *è invece di tipo lineare. Scegliendo quindi i dati da esaminare in modo opportuno, possiamo ricondurci ad una relazione lineare e quindi utilizzare il coefficiente di correlazione lineare per verificare la validità di tale relazione confrontandoci con dati reali.*

Esempio 3.3. *Consideriamo i dati riportati nell'Esempio 3.2, relativi ai prezzi e alla richiesta sul mercato di sei prodotti, e calcoliamone il coefficiente di correlazione lineare. Abbiamo che* $\bar{x} = 40.6$*,* $\bar{y} = 7336$*,* $s_{xy} = -4827.2$*,* $s_x^2 = 121.48$ *e* $s_y^2 \approx 192341$*. Quindi*

$$r_{xy} \;=\; \frac{s_{xy}}{s_x s_y} \;\approx\; -0.9986.$$

Notiamo che r_{xy} *è molto vicino a* -1*, dunque è possibile trovare una relazione lineare che descriva in modo significativo la dipendenza tra questi due caratteri, in accordo con l'esame visivo del diagramma a dispersione (anche se, come si vede dal digramma a dispersione, tale relazione sarà comunque valida approssimativamente per i dati in esame).*

Esempio 3.4. *Consideriamo i dati riportati nell'Esempio 3.1, relativi ai voti in italiano e in matematica di tredici studenti, e calcoliamone il coefficiente di correlazione lineare. Abbiamo che* $\bar{x} = \frac{89}{13} \approx 6.85$, $\bar{y} = \frac{87}{13} \approx 6.69$, $s_{xy} = \frac{70}{169} \approx 0.41$, $s_x^2 = \frac{532}{13} \approx 40.92$ *e* $s_y^2 = \frac{504}{13} \approx 38.77$. *Quindi*

$$r_{xy} \;=\; \frac{s_{xy}}{s_x s_y} \;\approx\; 0.01.$$

Notiamo che r_{xy} *è molto vicino a zero, concludiamo dunque che non è possibile trovare alcuna relazione lineare in grado di descrivere in maniera significativa la dipendenza tra i due caratteri (si noti inoltre che il solo valore della covarianza non ci permette di arrivare a questa conclusione, dato che la covarianza non è normalizzata).*

3.3 Metodo dei minimi quadrati e regressione lineare

In questa sezione ci occupiamo del problema di determinare concretamente una relazione lineare in grado di descrivere, seppur approssimativamente, la relazione tra due caratteri. Faremo ciò indipendentemente dal valore assunto da r_{xy}, tuttavia tale relazione lineare sarà chiaramente tanto più significativa quanto più $|r_{xy}|$ assumerà un valore vicino ad 1.

L'idea per determinare tale retta è la seguente: abbiamo una *nuvola di punti* nel piano cartesiano e cerchiamo due numeri reali m e q tali che la retta

$$y \;=\; m\,x + q$$

passi il più vicino possibile a tutti i punti $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Per ogni punto consideriamo l'*errore relativo* a tale punto (anche detto *errore parziale*), ovvero la differenza tra il valore reale e quello lungo la retta

$$e_i \;:=\; y_i - m\,x_i - q.$$

Un procedimento che permette di determinare i numeri m e q è il ***metodo dei minimi quadrati***, in base al quale si cercano m e q tali per cui la somma degli errori parziali elevati al quadrato, ossia

$$\sum_{i=1}^{n} (y_i - m\,x_i - q)^2,$$

sia minima. La retta che si ottiene procedendo in questo modo si chiama ***retta di regressione*** (o ***retta dei minimi quadrati***).

Teorema 3.3. *Sia dato un insieme di dati bivariati* $\{(x_i, y_i)\colon i = 1, \dots, n\}$*, con medie aritmetiche* $\bar{x}, \bar{y}$*, deviazioni standard* s_x, s_y *e covarianza* s_{xy}*. Supponiamo inoltre che* $s_x \neq 0$*.*
La ***retta di regressione*** *(o* ***retta dei minimi quadrati****) è data da*

$$y \;=\; m^* x + q^*$$

con

$$m^* \;=\; \frac{s_{xy}}{s_x^2}, \qquad\qquad q^* \;=\; \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}.$$

Quindi

$$y \;=\; \frac{s_{xy}}{s_x^2}(x - \bar{x}) + \bar{y}.$$

Dimostrazione*. Poniamo

$$S(m, q) \;:=\; \sum_{i=1}^{n} (y_i - m\, x_i - q)^2.$$

Un modo per determinare m e q consiste nel calcolare il gradiente della funzione S e porlo uguale a zero, ottenendo così un sistema di due equazioni nelle due incognite m e q. Vediamo invece una dimostrazione alternativa che non usa il calcolo differenziale in più variabili.
Riscriviamo innanzitutto $S(m, q)$ come segue:

$$\begin{aligned} S(m, q) \;&=\; \sum_{i=1}^{n} \big(y_i - \bar{y} - m\,(x_i - \bar{x}) + \bar{y} - \bar{x}\, m - q\big)^2 \\ &=\; \sum_{i=1}^{n} (y_i - \bar{y})^2 + m^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\,(\bar{y} - \bar{x}\, m - q)^2 - 2\, m \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \\ &\quad + 2\,(\bar{y} - \bar{x}\, m - q) \sum_{i=1}^{n} (y_i - \bar{y}) - 2\, m\,(\bar{y} - \bar{x}\, m - q) \sum_{i=1}^{n} (x_i - \bar{x}). \end{aligned}$$

Ricordando che

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x}) \;&=\; 0, & \sum_{i=1}^{n} (x_i - \bar{x})^2 \;&=\; n\, s_x^2, \\ \sum_{i=1}^{n} (y_i - \bar{y}) \;&=\; 0, & \sum_{i=1}^{n} (y_i - \bar{y})^2 \;&=\; n\, s_y^2, \\ & & \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \;&=\; n\, s_{xy}, \end{aligned}$$

otteniamo la seguente espressione quadratica di S nelle variabili m e q:

$$\begin{aligned} S(m, q) \;&=\; n\, s_y^2 + n\, s_x^2\, m^2 + n\,(\bar{y} - \bar{x}\, m - q)^2 - 2\, n\, s_{xy}\, m \\ &=\; n\left(\bar{x}^2 + s_x^2\right) m^2 + n\, q^2 - 2\, n\left(\bar{x}\,\bar{y} + s_{xy}\right) m - 2\, n\,\bar{y}\, q \end{aligned}$$

$$+ 2\,n\,\bar{x}\,m\,q + n\,\bar{y}^2 + n\,s_y^2.$$

Per poter determinare i valori di m e q che minimizzano S, è utile eliminare il termine misto $2\,n\,\bar{x}\,m\,q$ eseguendo il seguente cambio di variabili:

$$\begin{cases} m &= a + b, \\ q &= (a - b)\,\sqrt{\bar{x}^2 + s_x^2}. \end{cases} \tag{3.5}$$

Nelle nuove variabili a e b, otteniamo la seguente espressione di S:

$$\begin{aligned} &2\,n\Big(\big(\bar{x}^2 + s_x^2 + \bar{x}\,\sqrt{\bar{x}^2 + s_x^2}\big)\,a^2 - \big(\bar{x}\,\bar{y} + s_{xy} + \bar{y}\,\sqrt{\bar{x}^2 + s_x^2}\big)\,a\Big) \\ &+ 2\,n\Big(\big(\bar{x}^2 + s_x^2 - \bar{x}\,\sqrt{\bar{x}^2 + s_x^2}\big)\,b^2 - \big(\bar{x}\,\bar{y} + s_{xy} - \bar{y}\,\sqrt{\bar{x}^2 + s_x^2}\big)\,b\Big) + n\,\bar{y}^2 + n\,s_y^2. \end{aligned}$$

In questa nuova espressione, essendo le variabili a e b separate, i valori di a e b che realizzano il minimo si ottengono minimizzando separatamente l'espressione nella variabile a, ossia

$$\big(\bar{x}^2 + s_x^2 + \bar{x}\,\sqrt{\bar{x}^2 + s_x^2}\big)\,a^2 - \big(\bar{x}\,\bar{y} + s_{xy} + \bar{y}\,\sqrt{\bar{x}^2 + s_x^2}\big)\,a, \tag{3.6}$$

e l'espressione nella variabile b, ossia

$$\big(\bar{x}^2 + s_x^2 - \bar{x}\,\sqrt{\bar{x}^2 + s_x^2}\big)\,b^2 - \big(\bar{x}\,\bar{y} + s_{xy} - \bar{y}\,\sqrt{\bar{x}^2 + s_x^2}\big)\,b. \tag{3.7}$$

È facile mostrare che il coefficiente $\bar{x}^2 + s_x^2 + \bar{x}\sqrt{\bar{x}^2 + s_x^2}$ è non negativo, come anche il coefficiente $\bar{x}^2 + s_x^2 - \bar{x}\sqrt{\bar{x}^2 + s_x^2}$. Quindi (3.6) e (3.7) descrivono due parabole con concavità rivolta verso l'alto, perciò il minimo è realizzato in corrispondenza del vertice:

$$\begin{aligned} a^* &= \frac{\bar{x}\,\bar{y} + s_{xy} + \bar{y}\,\sqrt{\bar{x}^2 + s_x^2}}{2\big(\bar{x}^2 + s_x^2 + \bar{x}\,\sqrt{\bar{x}^2 + s_x^2}\big)}, \\ b^* &= \frac{\bar{x}\,\bar{y} + s_{xy} - \bar{y}\,\sqrt{\bar{x}^2 + s_x^2}}{2\big(\bar{x}^2 + s_x^2 - \bar{x}\,\sqrt{\bar{x}^2 + s_x^2}\big)}. \end{aligned}$$

Ricordando (3.5), otteniamo

$$\begin{aligned} m^* &= a^* + b^* = \frac{s_{xy}}{s_x^2}, \\ q^* &= (a^* - b^*)\,\sqrt{\bar{x}^2 + s_x^2} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}, \end{aligned}$$

come volevasi dimostrare. □

Osservazione 1. *Aver determinato la retta di regressione non significa affatto che tra i due caratteri esista effettivamente una relazione lineare, anche nel caso in cui* $|r_{xy}| = 1$. *Infatti, se* $|r_{xy}| = 1$ *è vero che i* dati osservati *sono disposti lungo una retta, ma se tali dati costituiscono solo un* campione *dell'intera popolazione, non possiamo affermare che esista davvero una relazione lineare in generale.*

Osservazione 2. *Come affermato nell'osservazione precedente, se* $|r_{xy}|$ *è vicino*[9] *ad uno, la* retta di regressione *può essere considerata un buon modello per descrivere la relazione*

[9] Oltre a verificare che $|r_{xy}|$ sia vicino ad uno, è utile anche l'esame visivo del diagramma a dispersione.

tra i due caratteri. Possiamo dunque utilizzare la retta di regressione per fare delle previsioni: *dato un valore x_0 diverso dai valori $x_1, \dots, x_n$ osservati, il* ***valore previsto*** *del secondo carattere sarà*

$$y_0 \;=\; \frac{s_{xy}}{s_x^2}(x_0 - \bar{x}) + \bar{y}.$$

Come è naturale aspettarsi, la previsione sarà tanto più affidabile quanto più x_0 è vicino ai valori $x_1, \dots, x_n$ già osservati.
Se i due caratteri sono incorrelati, *quindi $r_{xy} = 0$ (e di conseguenza $s_{xy} = 0$), allora la retta di regressione è la retta orizzontale*

$$y \;=\; \bar{y}.$$

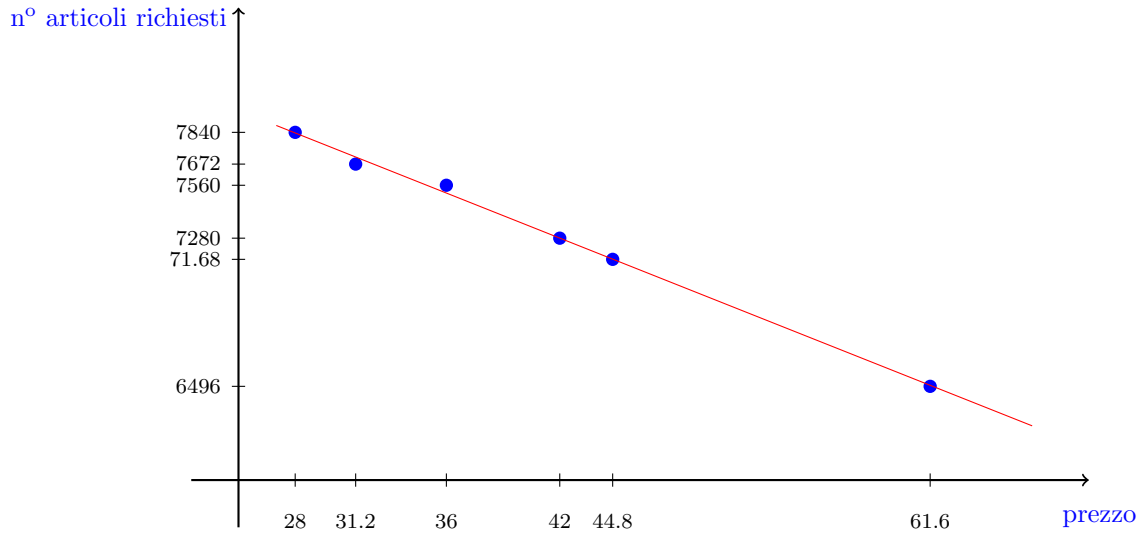
Questo significa che nessuna previsione può essere fatta su y a partire dall'osservazione di x, perciò la migliore previsione di y è la media stessa $\bar{y}$.

Esempio 3.5. *Consideriamo nuovamente i dati riportati nell'Esempio 3.2, relativi ai prezzi e alla richiesta sul mercato di sei prodotti, e determiniamo la retta di regressione. Dall'Esempio 3.3 sappiamo che $\bar{x} = 40.6$, $\bar{y} = 7336$, $s_{xy} = -4827.2$ e $s_x^2 = 121.48$. Quindi la* retta di regressione *è data da*

$$y \;=\; -39.74\,(x - 40.6) + 7336.$$

In questo caso $r_{xy} \approx -0.9986$, quindi tale retta è effettivamente un buon modello per descrivere la relazione tra i due caratteri.

Ecco il grafico della retta di regressione relativa all'Esempio 3.5, riportato insieme al diagramma a dispersione:



Esempio 3.6. *Consideriamo nuovamente i dati riportati nell'Esempio 3.1, relativi ai voti in italiano e in matematica di tredici studenti, e determiniamo la retta di regressione. Dall'Esempio 3.4 sappiamo che* $\bar{x} \approx 6.85$, $\bar{y} \approx 6.69$, $s_{xy} \approx 0.41$ *e* $s_x^2 \approx 40.92$. *Quindi la* retta di regressione *è data da*

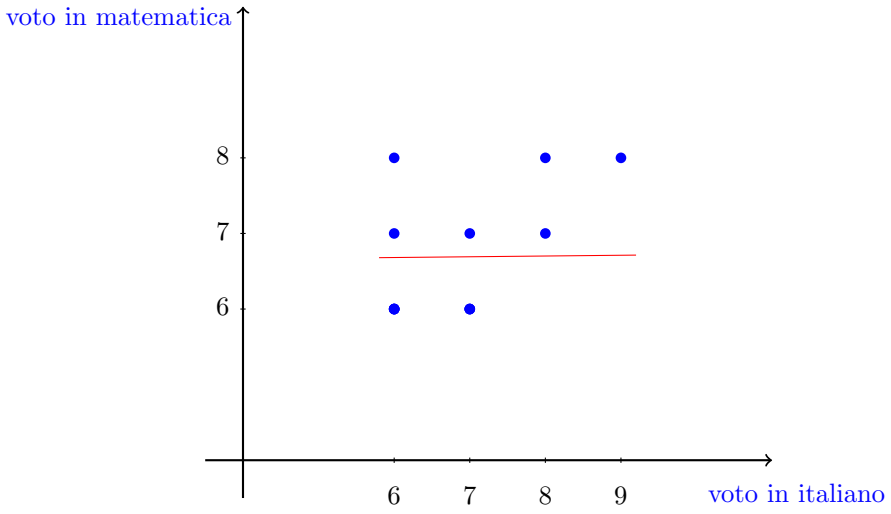
$$y \;=\; 0.01\,(x - 6.85) + 6.69.$$

In questo caso $r_{xy} \approx 0.01$, *quindi, come già osservato nell'Esempio 3.4, non esiste alcuna retta in grado di descrivere in maniera significativa la dipendenza tra i due caratteri. Infatti, abbiamo che*

$$y \;=\; 0.01\,(x - 6.85) + 6.69 \;\approx\; 6.69 \;=\; \bar{y},$$

dove l'approssimazione va intesa nell'intervallo di variazione dei dati, quindi per x *tra* 6 *e* 9.

Ecco il grafico della retta di regressione relativa all'Esempio 3.6, riportato insieme al diagramma a dispersione:



4 Teoremi limite

L'interesse nei confronti dei teoremi limite del Calcolo delle probabilità nasce proprio dalle applicazioni statistiche. Più precisamente, supponiamo di essere interessati ad un esperimento aleatorio e, in particolare, ad una variabile aleatoria ad esso collegata. Indichiamo tale variabile aleatoria con X. Come possiamo determinare, o meglio "stimare", la distribuzione di X? Per stimare la distribuzione di X, un buon punto di partenza consiste nello stimare la media di tale distribuzione, cioè $\mathbb{E}[X]$. In questo capitolo forniremo le basi teoriche per lo studio del seguente problema: *data una generica variabile aleatoria* X, *come si stima* $\mathbb{E}[X]$*?*

Si noti che saper stimare il valore atteso di una generica variabile aleatoria X, significa non solo saper stimare $\mathbb{E}[X]$ ma anche $\mathbb{E}[f(X)]$ con $f\colon \mathbb{R} \to \mathbb{R}$ funzione arbitraria. Si può dimostrare che la conoscenza di tutti i valori attesi $\mathbb{E}[f(X)]$, con f funzione arbitraria, è equivalente alla conoscenza della distribuzione di X. In altre parole, saper stimare $\mathbb{E}[X]$, con X arbitraria, permette di risolvere, per lo meno a livello teorico, il problema che abbiamo posto all'inizio: *data una qualunque variabile aleatoria X, come si stima la distribuzione di X?*

Notiamo inoltre che saper stimare il valore atteso di una qualunque variabile aleatoria X significa anche saper stimare la probabilità di un qualunque evento A. Infatti, è sufficiente scegliere $X = 1_A$, la variabile aleatoria indicatrice relativa all'evento A, e ricordare che $\mathbb{P}(A) = \mathbb{E}[1_A]$.

Sia dunque X una generica variabile aleatoria di cui si vuole stimare il valore atteso. Per ottenere una "stima" si segue questo classico procedimento della Statistica: si ripete un numero "elevato" di volte l'esperimento aleatorio, ogni volta registrando quale valore ha assunto la variabile aleatoria X. Si ottiene dunque una sequenza di valori numerici[10]:

$$x_1 \qquad x_2 \qquad x_3 \qquad \cdots \qquad x_n$$

La sequenza così ottenuta è perciò un *campione* di dati. Come otteniamo a partire da $x_1, x_2, x_3, \ldots, x_n$ una stima di $\mathbb{E}[X]$? La scelta naturale è quella di considerare la *media campionaria*

$$\bar{x}_n \;=\; \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}. \tag{4.1}$$

La media aritmetica del campione $x_1, x_2, x_3, \ldots, x_n$ si chiama anche *media campionaria* e si indica con il simbolo $\bar{x}_n$, quindi

$$\bar{x}_n \;=\; \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

4.1 Successioni di variabili aleatorie i.i.d.

Iniziamo a formalizzare il problema presentato nella sezione precedente. Immaginiamo dunque di ripetere l'esperimento aleatorio a cui siamo interessati *infinite*[11] volte. Prima di eseguire tali esperimenti, il valore assunto dalla variabile aleatoria di interesse è da ritenersi aleatorio. Quindi è naturale considerare una *successione di variabili aleatorie* che rappresentano gli ipotetici valori assunti dalla variabile aleatoria di interesse nei vari esperimenti:

$$X_1 \qquad X_2 \qquad X_3 \qquad \cdots \qquad X_n \qquad \cdots$$

La lettera maiuscola sta dunque ad indicare che gli esperimenti devono ancora essere svolti e le quantità sono quindi aleatorie. Solo dopo aver eseguito gli esperimenti conosceremo i valori da esse assunti, che saranno indicati con le lettere minuscole $x_1, x_2, x_3, \ldots, x_n, \ldots$

La successione $X_1, X_2, X_3, \ldots, X_n, \ldots$ verrà indicata anche con il simbolo

$$(X_n)_n.$$

[10] Indichiamo i valori assunti da X nei vari esperimenti con lettere *minuscole* dato che non sono aleatori, infatti stiamo supponendo che gli esperimenti si siano già svolti e noi ne conosciamo l'esito.

[11] Chiaramente nella realtà potremo fare solo un numero *finito*, anche se "elevato", di ripetizioni.

Per quanto detto finora è chiaro che le variabili aleatorie $X_1, X_2, X_3, \ldots, X_n, \ldots$ verificano la seguente proprietà.

1) *Le variabili aleatorie $X_1, X_2, X_3, \ldots, X_n, \ldots$ hanno tutte la stessa distribuzione.*

2) *Le variabili aleatorie $X_1, X_2, X_3, \ldots, X_n, \ldots$ sono indipendenti.*

Nel seguito considereremo sempre successioni di variabili aleatorie che verificano le proprietà 1) e 2). Risulta quindi utile la seguente definizione.

Definizione 4.1. *$(X_n)_n$ è una* ***successione di variabili aleatorie i.i.d.***[a] *se valgono le seguenti due proprietà:*

1) *$X_1, X_2, \ldots, X_n, \ldots$ hanno tutte la stessa distribuzione;*

2) *$(X_n)_n$ è una successione di variabili aleatorie indipendenti.*

[a] *i.i.d.* sta per *indipendenti e identicamente distribuite.*

4.2 Legge dei grandi numeri (LGN)

Iniziamo con un risultato preliminare, particolarmente importante.

Lemma 4.1 (**Disuguaglianza di Chebyshev**). *Sia Y una variabile aleatoria con media μ. Per ogni $\varepsilon > 0$, vale che*

$$\mathbb{P}\big(|Y-\mu| > \varepsilon\big) \;\leq\; \frac{\mathrm{Var}(Y)}{\varepsilon^2}.$$

Dimostrazione*. Sia

$$Z \;=\; 1_{\{|Y-\mu|>\varepsilon\}} \;=\; \begin{cases} 1, & \text{se } |Y-\mu| > \varepsilon, \\ 0, & \text{altrimenti.} \end{cases}$$

In altre parole, Z è la variabile aleatoria indicatrice relativa all'evento $\{|Y-\mu| > \varepsilon\}$. Quindi, in particolare, $Z \sim B(p)$ con

$$p \;=\; \mathbb{P}\big(|Y-\mu| > \varepsilon\big).$$

Si noti che

$$\begin{aligned} \mathrm{Var}(Y) \;&=\; \mathbb{E}\big[(Y-\mu)^2\big] \\ &\geq\; \mathbb{E}\big[(Y-\mu)^2\, 1_{\{|Y-\mu|>\varepsilon\}}\big] \\ &\geq\; \mathbb{E}\big[\varepsilon^2\, 1_{\{|Y-\mu|>\varepsilon\}}\big] \;=\; \varepsilon^2\, \mathbb{E}\big[1_{\{|Y-\mu|>\varepsilon\}}\big] \;=\; \varepsilon^2\, \mathbb{E}[Z]. \end{aligned}$$

Dato che $Z \sim B(p)$, si ha che $\mathbb{E}[Z] = p = \mathbb{P}(|Y-\mu| > \varepsilon)$, quindi

$$\mathrm{Var}(Y) \;\geq\; \varepsilon^2\, \mathbb{P}\big(|Y-\mu| > \varepsilon\big).$$

□

Passiamo ora alla *Legge dei grandi numeri*. Consideriamo dunque una successione di variabili aleatorie i.i.d.

$$X_1 \qquad X_2 \qquad X_3 \qquad \cdots \qquad X_n \qquad \cdots$$

o, più sinteticamente, $(X_n)_n$. Indichiamo con μ e σ^2 rispettivamente la loro media e la loro varianza. Per ogni n fissato, definiamo la media campionaria delle prime n variabili aleatorie come segue:

$$\bar{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n}.$$

Si noti che $\bar{X}_n$ è anch'essa una variabile aleatoria, infatti il suo valore non è ancora noto. Sarà noto solo dopo aver svolto i primi n esperimenti. A quel punto indicheremo il suo valore con la lettera minuscola $\bar{x}_n$.

Come abbiamo sottolineato all'inizio, la media campionaria si usa in Statistica per stimare la vera media μ delle variabili aleatorie $X_1, \ldots, X_n, \ldots$. Questa è una conseguenza della *Legge dei grandi numeri*, la quale stabilisce che $\bar{X}_n$ "converge" verso μ quando n tende all'infinito.

Teorema 4.1 (**Legge dei grandi numeri**). *Sia $(X_n)_n$ una successione di variabili aleatorie i.i.d. con media μ e varianza σ^2. Allora, posto*

$$\bar{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n},$$

si ha

$$\forall\, \varepsilon > 0, \qquad \lim_{n \to +\infty} \mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \;=\; 0. \tag{4.2}$$

Inoltre, vale che

$$\mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \;\leq\; \frac{\sigma^2}{\varepsilon^2\, n}. \tag{4.3}$$

Notazione. *Se vale che*

$$\forall\, \varepsilon > 0, \qquad \lim_{n \to +\infty} \mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \;=\; 0,$$

si dice che $\bar{X}_n$ ***converge in probabilità*** *a μ quando n tende all'infinito. In tal caso, si scrive*

$$\bar{X}_n \;\xrightarrow[n \to +\infty]{\mathbb{P}}\; \mu.$$

Osservazione. *La* (4.3) *fornisce una stima della velocità di convergenza in probabilità.*

Dimostrazione della Legge dei grandi numeri*. La dimostrazione consiste nell'applicazione della disuguaglianza di Chebyshev alla variabile aleatoria $\bar{X}_n$. Per applicare tale disuguaglianza, dobbiamo prima calcolare media e varianza di $\bar{X}_n$.
La *media* di $\bar{X}_n$ si calcola usando la linearità del valore atteso:

$$\mathbb{E}[\bar{X}_n] \;=\; \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] \;=\; \frac{1}{n}\mathbb{E}[X_1 + \cdots + X_n]$$

$$\underset{\substack{\uparrow\\ \text{linearità di } \mathbb{E}[\cdot]}}{=} \frac{1}{n}\big(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\big)$$

$$\underset{\substack{\uparrow\\ \text{ident. distr.}}}{=} \frac{1}{n} n\,\mu \;=\; \mu.$$

Quindi anche $\bar{X}_n$ ha media μ.
Per quanto riguarda la *varianza*, grazie all'indipendenza di $X_1, \ldots, X_n$ si ha che

$$\mathrm{Var}(\bar{X}_n) \;=\; \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) \;=\; \frac{1}{n^2}\,\mathrm{Var}(X_1 + \cdots + X_n)$$

$$\underset{\substack{\uparrow\\ \text{indipendenza}}}{=} \frac{1}{n^2}\big(\,\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\big)$$

$$\underset{\substack{\uparrow\\ \text{ident. distr.}}}{=} \frac{1}{n^2} n\,\sigma^2 \;=\; \frac{\sigma^2}{n}.$$

Adesso, per ogni $\varepsilon > 0$ fissato, applicando la disuguaglianza di Chebyshev alla variabile aleatoria $\bar{X}_n$, otteniamo

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \;\le\; \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} \;=\; \frac{\sigma^2}{\varepsilon^2\, n},$$

che dimostra la formula (4.3). Dimostriamo infine la formula (4.2). Poiché

$$0 \;\le\; \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \;\le\; \frac{\sigma^2}{\varepsilon^2\, n} \underset{n\to+\infty}{\longrightarrow} 0,$$

concludiamo che

$$\lim_{n\to+\infty} \mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \;=\; 0.$$

□

*4.2.1 Metodo Monte Carlo**

La Legge dei grandi numeri è alla base di un metodo numerico *probabilistico* molto importante, noto come *metodo Monte Carlo*[12]. Consideriamo il seguente problema.

Come si può approssimare $\displaystyle\int_a^b f(x)\,\mathrm{d}x$*, con* f *funzione integrabile?*

[12]Il metodo Monte Carlo è stato sviluppato nell'ambito della ricerca nucleare. La sua nascita si attribuisce in particolare al matematico polacco Stanislaw Ulam, che lavorava nell'ambito del progetto Manhattan. Anche il fisico italiano Enrico Fermi e il matematico ungherese John von Neumann hanno contribuito alla nascita di questo metodo. Il nome è stato coniato successivamente dal matematico statunitense Nicholas Metropolis (anch'egli all'interno del progetto Manhattan), facendo proprio riferimento alla città di Monte Carlo e al suo casinò. Nella sua autobiografia Ulam descrive come l'idea gli sia venuta cercando di calcolare la probabilità di vincere al solitario. Più precisamente, si consideri un mazzo di 52 carte. La riuscita o meno del solitario dipende solamente da come sono ordinate le carte nel mazzo. In totale ci sono 52! ordinamenti. Quindi

$$\mathbb{P}(\text{“vincere”}) \;=\; \frac{\text{numero di solitari riusciti}}{52!}.$$

Per semplicità, consideriamo il caso $a = 0$ e $b = 1$, quindi l'integrale diventa $\int_0^1 f(x)\,\mathrm{d}x$. Possiamo riscrivere questo integrale come valore atteso:

$$\int_0^1 f(x)\,\mathrm{d}x \;=\; \mathbb{E}[f(U)],$$

dove $U \sim \mathrm{Unif}\,(0,1)$. Ci siamo dunque ricondotti al problema di stimare il valore atteso della variabile aleatoria $X = f(U)$. Il metodo Monte Carlo consiste nell'approssimare numericamente il valore atteso $E[f(U)]$ facendo uso della Legge dei grandi numeri. Più precisamente, sia $(U_n)_n$ una successione di variabili aleatorie i.i.d. con la medesima distribuzione di U, quindi uniforme su $(0,1)$. Definiamo

$$X_n \;=\; f(U_n), \qquad \forall\, n.$$

Allora $(X_n)_n$ è ancora una successione di variabili aleatorie i.i.d., con la medesima distribuzione di $f(U)$. Quindi, per la Legge dei grandi numeri,

$$\frac{f(U_1) + \cdots + f(U_n)}{n} \;\xrightarrow[n\to+\infty]{\mathbb{P}}\; \mathbb{E}[f(U)] \;=\; \int_0^1 f(x)\,\mathrm{d}x.$$

L'implementazione del metodo Monte Carlo si basa sull'utilizzo dei *generatori aleatori*. Nell'esempio qui considerato, per approssimare $\mathbb{E}[f(U)]$ si genera una sequenza di numeri casuali con distribuzione uniforme su $(0,1)$, quindi

$$u_1 \qquad u_2 \qquad \cdots \qquad u_n$$

Tali numeri sono scritti con la lettera minuscola in quanto sono noti, infatti sono i numeri forniti dal generatore aleatorio. Dopodiché, si calcola la quantità

$$\frac{f(u_1) + \cdots + f(u_n)}{n}.$$

Se n è "elevato" si ottiene una buona approssimazione dell'integrale $\int_0^1 f(x)\,\mathrm{d}x$.

I principali vantaggi rispetto ai metodi *deterministici* di integrazione numerica sono i seguenti:

- non si richiedono ipotesi di regolarità sulla funzione integranda f;
- l'ordine di convergenza del metodo, che è $\frac{1}{\sqrt{n}}$ come seguirà dal Teorema centrale del limite, è indipendente dalla dimensione e l'implementazione del metodo in dimensione maggiore di uno non comporta alcuna difficoltà aggiuntiva.

Come racconta egli stesso: *"L'idea del metodo Monte Carlo mi è venuta giocando a carte un solitario durante un periodo di convalescenza, nel 1946. Avevo sprecato un mucchio di tempo per calcolare, senza successo, con tecniche combinatorie, la probabilità di riuscita del solitario. Pensai allora che, giocando un centinaio di volte il solitario, avrei potuto stimare questa probabilità con la frequenza delle volte con cui era riuscito, aggirando così con la pratica il pensiero astratto. Questo metodo era ormai possibile, visto l'avvento dei calcolatori veloci. Era ovvio pensare anche a soluzioni simili per problemi legati alla diffusione dei neutroni o di fisica matematica e, più in generale, a come scambiare processi descritti da certe equazioni differenziali con un modello equivalente interpretabile come successione di operazioni aleatorie. In seguito descrissi l'idea a John von Neumann e cominciammo a realizzare veri e propri calcoli matematici al riguardo."*

*4.2.2 Metodo del gradiente stocastico**

In questa sezione presentiamo un altro metodo numerico, noto come *metodo del gradiente stocastico*, particolarmente importante nell'ambito delle reti neurali. Iniziamo col descrivere il *metodo del gradiente*, che è un metodo completamente deterministico.

Metodo del gradiente. Sia $f\colon \mathbb{R}^d \to \mathbb{R}$ e consideriamo il seguente problema[13] di ottimizzazione:

$$\textit{Trovare } \boldsymbol{x}^* \in \mathbb{R}^d \textit{ punto di minimo di } f\colon f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} f(\boldsymbol{x}). \tag{4.4}$$

Quando f verifica opportune ipotesi di regolarità, il metodo del gradiente permette di determinare in modo approssimato un tale punto $\boldsymbol{x}^*$. Alla base di questo metodo vi è una proprietà del gradente che ora richiamiamo. Innanzitutto, ricordiamo che il gradiente di f calcolato nel punto $\boldsymbol{x} = (x_1, \ldots, x_d)$, indicato con $\nabla f(\boldsymbol{x})$, è il vettore delle derivati parziali prime:

$$\nabla f(\boldsymbol{x}) \;=\; \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\boldsymbol{x}) \end{pmatrix}.$$

Il vettore $\nabla f(\boldsymbol{x})$ rappresenta l'incremento infinitesimo della funzione f nel punto $\boldsymbol{x}$, si deduce dunque la seguente fondamentale proprietà: se a partire da $\boldsymbol{x}$ ci muoviamo lungo il grafico di f, allora le direzioni di *massima crescita* e *massima decrescita* sono individuate rispettivamente dai vettori $\nabla f(\boldsymbol{x})$ e $-\nabla f(\boldsymbol{x})$.

Veniamo dunque alla descrizione del metodo del gradiente. Per determinare $\boldsymbol{x}^*$ si procede in modo *iterativo*:

- al passo 0, si sceglie in modo arbitrario un punto di partenza $\boldsymbol{x}_0 \in \mathbb{R}^d$;
- al generico passo $k = 1, 2, 3, \ldots$, in cui sono già stati determinati i valori $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, si determina $\boldsymbol{x}_{k+1}$ come segue:

$$\boldsymbol{x}_{k+1} \;=\; \boldsymbol{x}_k + \alpha_k\, \boldsymbol{v}_k, \tag{4.5}$$

  dove:

  - $\boldsymbol{v}_k$ è un vettore di lunghezza unitaria (detto anche *versore*), che individua la direzione in $\mathbb{R}^d$ lungo cui muoversi;
  - α_k è un numero reale strettamente positivo che rappresenta la distanza da compiere lungo la direzione $\boldsymbol{v}_k$.

La direzione ottimale lungo cui muoversi ad ogni passo k è quella che congiunge $\boldsymbol{x}_k$ a $\boldsymbol{x}^*$, che tuttavia non è ovviamente nota a priori. Dato che $\boldsymbol{x}^*$ è un *punto di minimo*, l'idea più naturale è prendere come direzione quella di *massima decrescita*, data da $-\nabla f(\boldsymbol{x}_k)$. Per tale ragione, il metodo del gradiente corrisponde al seguente schema iterativo:

$$\boldsymbol{x}_{k+1} \;=\; \boldsymbol{x}_k - \beta_k\, \nabla f(\boldsymbol{x}_k).$$

[13] Se invece che al minimo siamo interessati al massimo basta notare che un punto di massimo di f è un punto di minimo di $-f$. Si applica dunque il metodo del gradiente alla funzione $-f$.

La quantità che qui abbiamo chiamato β_k non corrisponde in generale al parametro α_k che compare in (4.5), infatti $\nabla f(\boldsymbol{x}_k)$ non ha generalmente lunghezza unitaria. Più precisamente, vale la relazione

$$\alpha_k \;=\; \beta_k \frac{\nabla f(\boldsymbol{x}_k)}{\text{lunghezza di } \nabla f(\boldsymbol{x}_k)}.$$

Quando f verifica opportune ipotesi di regolarità, vale che

$$\lim_{k\to+\infty} f(\boldsymbol{x}_k) \;=\; f(\boldsymbol{x}^*).$$

Metodo del gradiente stocastico. Nell'ambito delle reti neurali si è interessati al problema di ottimizzazione (4.4) per una particolare funzione f, avente la seguente espressione:

$$f(\boldsymbol{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}), \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^d.$$

In altre parole, f è la media aritmetica di n funzioni qui indicate con $f_1, \ldots, f_n$. Il *metodo del gradiente* applicato ad una tale funzione corrisponde al seguente schema iterativo:

$$\boldsymbol{x}_{k+1} \;=\; \boldsymbol{x}_k - \beta_k \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}_k).$$

Se n è molto elevato può essere particolarmente oneroso determinare $\boldsymbol{x}_{k+1}$. Il *metodo del gradiente stocastico* consiste dunque nell'individuare, ad ogni passo k, un sottoinsieme di addendi in modo *casuale*; sono tali addendi i soli che vengono utilizzati al passo k per determinare $\boldsymbol{x}_{k+1}$.

4.3 Teorema centrale del limite (TCL)

Come nel caso della Legge dei grandi numeri, consideriamo una successione $(X_n)_n$ di variabili aleatorie i.i.d. e indichiamo con μ e σ^2 media e varianza di ciascuna variabile aleatoria X_n. Sia inoltre

$$\bar{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n}.$$

Grazie alla Legge dei grandi numeri sappiamo che vale la convergenza

$$\bar{X}_n \;\xrightarrow[n\to+\infty]{\mathbb{P}}\; \mu.$$

Il Teorema centrale del limite[14] descrive come avviene questa convergenza, o più precisamente, ci dice qual è approssimativamente la distribuzione di $\bar{X}_n$ per n grande.
Prima di enunciare il Teorema centrale del limite, è utile introdurre la variabile aleatoria $\bar{Z}_n$ data da

$$\bar{Z}_n \;=\; \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

[14]Il nome *"Teorema centrale del limite"* (o *"Teorema limite centrale"*) è stato dato dal matematico ungherese George Pólya per sottolineare come tale teorema abbia un ruolo *centrale* in Probabilità e Statistica.

Si noti che

$$\mathbb{E}[\bar{Z}_n] \;=\; 0, \qquad\qquad \mathrm{Var}(\bar{Z}_n) \;=\; 1.$$

La variabile aleatoria $\bar{Z}_n$ si chiama ***media campionaria standardizzata***.

Teorema 4.2 (**Teorema centrale del limite**). *Sia $(X_n)_n$ una successione di variabili aleatorie i.i.d. con media μ e varianza σ^2. Supponiamo che $\sigma > 0$. Allora, posto*

$$\bar{Z}_n \;=\; \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}},$$

si ha

$$\lim_{n\to+\infty} \mathbb{P}(\bar{Z}_n \le x) \;=\; \lim_{n\to+\infty} F_{\bar{Z}_n}(x) \;=\; \Phi(x), \qquad \forall\, x \in \mathbb{R},$$

dove $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}y^2}\, \mathrm{d}y$ è la funzione di ripartizione della distribuzione normale standard.

Per una rappresentazione grafica (in termini di istogramma) della convergenza a cui si fa riferimento nel Teorema centrale del limite, si veda l'esempio **Testa e Croce** riportato qui di seguito. Per una dimostrazione del Teorema centrale del limite si veda invece l'**Appendice**.

Osservazione 1. *Se $\sigma = 0$, il Teorema centrale del limite non vale. Tuttavia, in tal caso possiamo dire molto di più sulla successione $\bar{X}_n$. Infatti, se $\sigma = 0$ allora ciascuna variabile aleatoria X_n è costante e inoltre $X_n = \mu$. Di conseguenza, anche $\bar{X}_n = \mu$, mentre $\bar{Z}_n = 0$, per ogni n.*

Osservazione 2. *Se vale che*

$$\lim_{n\to+\infty} F_{\bar{Z}_n}(x) \;=\; \Phi(x), \qquad \forall\, x \in \mathbb{R},$$

*si dice che $\bar{Z}_n$ **converge in legge** (o **in distribuzione**) ad una variabile aleatoria normale standard quando n tende all'infinito. In tal caso, si scrive*

$$\bar{Z}_n \;\xrightarrow[n\to+\infty]{\mathscr{L}}\; Z \;\sim\; \mathcal{N}(0,1).$$

Osservazione 3. *Sulla base dell'esperienza empirica, generalmente si applica il valore $n = 30$ come soglia di applicabilità del Teorema centrale del limite. Tuttavia questa soglia funziona bene solo per distribuzioni simmetriche. Se la distribuzione è particolarmente asimmetrica, bisogna considerare valori più grandi di n.*

Osservazione 4. *Se n è "elevato", dal Teorema centrale del limite si ha che*

$$F_{\bar{Z}_n}(x) \;\simeq\; \Phi(x), \qquad \textit{per ogni } x \in \mathbb{R}.$$

Questo significa che

$$\bar{Z}_n \;\approx\; Z,$$

con Z variabile aleatoria normale standard. Il simbolo $\approx$ indica che $\bar{Z}_n$ e Z hanno approssimativamente la stessa distribuzione. Dato che

$$\bar{X}_n \;=\; \mu + \frac{\sigma}{\sqrt{n}}\,\bar{Z}_n,$$

si ha

$$\bar{X}_n \;=\; \mu + \frac{\sigma}{\sqrt{n}}\,\bar{Z}_n \;\approx\; \mu + \frac{\sigma}{\sqrt{n}}\,Z. \tag{4.6}$$

Poiché $\mu + \frac{\sigma}{\sqrt{n}}\,Z \sim \mathcal{N}\big(\mu, \frac{\sigma^2}{n}\big)$, si deduce che $\bar{X}_n$ ha approssimativamente distribuzione normale di media μ e varianza σ^2/n. Inoltre, l'approssimazione

$$\bar{X}_n \;\approx\; \mu + \frac{\sigma}{\sqrt{n}}\,Z$$

precisa ed esplicita il risultato di convergenza della Legge dei grandi numeri. In particolare, fornisce l'ordine di convergenza $\frac{1}{\sqrt{n}}$. Infatti, l'errore (aleatorio) di approssimazione è dato da

$$|\bar{X}_n - \mu| \;\approx\; \frac{\sigma}{\sqrt{n}}\,|Z|.$$

Quindi l'errore medio è pari a

$$\mathbb{E}[|\bar{X}_n - \mu|] \;\simeq\; \frac{\sigma}{\sqrt{n}}\,\mathbb{E}[|Z|] \;=\; \frac{\sigma}{\sqrt{n}}\,\sqrt{\frac{2}{\pi}},$$

dove l'ultima uguaglianza segue dall'integrale

$$\mathbb{E}[|Z|] \;= \int_{-\infty}^{+\infty} |x|\,\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x \;=\; 2\int_0^{+\infty} x\,\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x \;=\; -\frac{2}{\sqrt{2\pi}}\,\Big[\mathrm{e}^{-\frac{1}{2}x^2}\Big]_0^{+\infty} \;=\; \sqrt{\frac{2}{\pi}}.$$

Osservazione 5. ***Come mai Z ha proprio distribuzione normale?*** *Non esiste una risposta completamente soddisfacente a questa domanda. Tuttavia, supponiamo di sapere solamente che $\bar{Z}_n$ converge in legge ad una qualche variabile aleatoria Z (di cui ancora non conosciamo la distribuzione). Consideriamo ora le seguenti sotto-successione della successione $(X_n)_n$:*

$$X_n^p \;:=\; X_{2n}, \qquad\qquad X_n^d \;:=\; X_{2n+1},$$

ovvero la sotto-successione delle v.a. con indice pari e quella delle v.a. con indice dispari. Siano $\bar{Z}_n^p$ e $\bar{Z}_n^d$ le corrispondenti medie campionarie standardizzate. Notiamo che $\bar{Z}_n^p$ e $\bar{Z}_n^d$ hanno la stessa legge di $\bar{Z}_n$. Quindi, dato che

$$\bar{Z}_n \;\xrightarrow[n\to+\infty]{\mathscr{L}}\; Z,$$

allora deve valere anche che

$$\bar{Z}_n^p \;\xrightarrow[n\to+\infty]{\mathscr{L}}\; Z^p, \qquad\qquad \bar{Z}_n^d \;\xrightarrow[n\to+\infty]{\mathscr{L}}\; Z^d,$$

con Z, Z^p e Z^d identicamente distribuite. Inoltre, essendo $\bar{Z}_n^p$ e $\bar{Z}_n^d$ indipendenti, segue che anche Z^p e Z^d sono indipendenti. Ora notiamo che

$$\bar{Z}_{2n} \;=\; \frac{\bar{Z}_n^p + \bar{Z}_n^d}{\sqrt{2}}.$$

Facendo quindi tendere $n \to +\infty$, si ottiene che le variabili aleatorie

$$Z \qquad\qquad e \qquad\qquad \frac{Z^p + Z^d}{\sqrt{2}}$$

hanno necessariamente la stessa legge. Allora la tesi è una conseguenza del seguente risultato. ***Siano X, X' e X'' v.a. identicamente distribuite. Siano inoltre X' e X'' indipendenti. Se le variabili aleatorie***

$$X \qquad\qquad \boldsymbol{e} \qquad\qquad \frac{X' + X''}{\sqrt{2}}$$

hanno la stessa legge, allora $X \sim \mathcal{N}(0, \sigma^2)$, per qualche $\sigma^2 \geq 0$.

Testa e Croce. Vediamo un esempio di applicazione del Teorema centrale del limite. Supponiamo di lanciare più volte una moneta e sia

$$X_n \;=\; \text{“vale 1 se esce testa all'}n\text{-esimo lancio, vale 0 altrimenti”}.$$

Allora $(X_n)_n$ è una successione di variabili aleatorie i.i.d. con distribuzione di Bernoulli di parametro $1/2$. Inoltre

$$S_n \;=\; X_1 + \cdots + X_n$$

ha distribuzione binomiale $S_n \sim B(n, 1/2)$. Quindi la media campionaria

$$\bar{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n} \;=\; \frac{S_n}{n}$$

ha una distribuzione descritta dalla seguente tabella:

$\bar{X}_n$	0	$\frac{1}{n}$	$\frac{2}{n}$	$\cdots$	$\frac{n-1}{n}$	1
$p_{\bar{X}_n}$	$\binom{n}{0}\frac{1}{2^n}$	$\binom{n}{1}\frac{1}{2^n}$	$\binom{n}{2}\frac{1}{2^n}$	$\cdots$	$\binom{n}{n-1}\frac{1}{2^n}$	$\binom{n}{n}\frac{1}{2^n}$

Dal Teorema centrale del limite (ricordando in particolare (4.6)) si ha che

$$\bar{X}_n \;\approx\; \mu + \frac{\sigma}{\sqrt{n}}\, Z,$$

con $Z \in \mathcal{N}(0,1)$. Poiché $\mu = \frac{1}{2}$ e $\sigma^2 = \frac{1}{4}$, otteniamo

$$\bar{X}_n \;\approx\; \frac{1}{2} + \frac{1}{2\sqrt{n}}\, Z,$$

quindi $\bar{X}_n$ ha approssimativamente distribuzione normale di media $\frac{1}{2}$ e varianza $\frac{1}{4n}$. Consideriamo ad esempio il caso $n = 10$. Allora $\bar{X}_{10}$ ha densità discreta

$\bar{X}_{10}$	0	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{5}{10}$	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{9}{10}$	1
$p_{\bar{X}_{10}}$	$\frac{1}{2^{10}}$	$\frac{10}{2^{10}}$	$\frac{45}{2^{10}}$	$\frac{120}{2^{10}}$	$\frac{210}{2^{10}}$	$\frac{252}{2^{10}}$	$\frac{210}{2^{10}}$	$\frac{120}{2^{10}}$	$\frac{45}{2^{10}}$	$\frac{10}{2^{10}}$	$\frac{1}{2^{10}}$

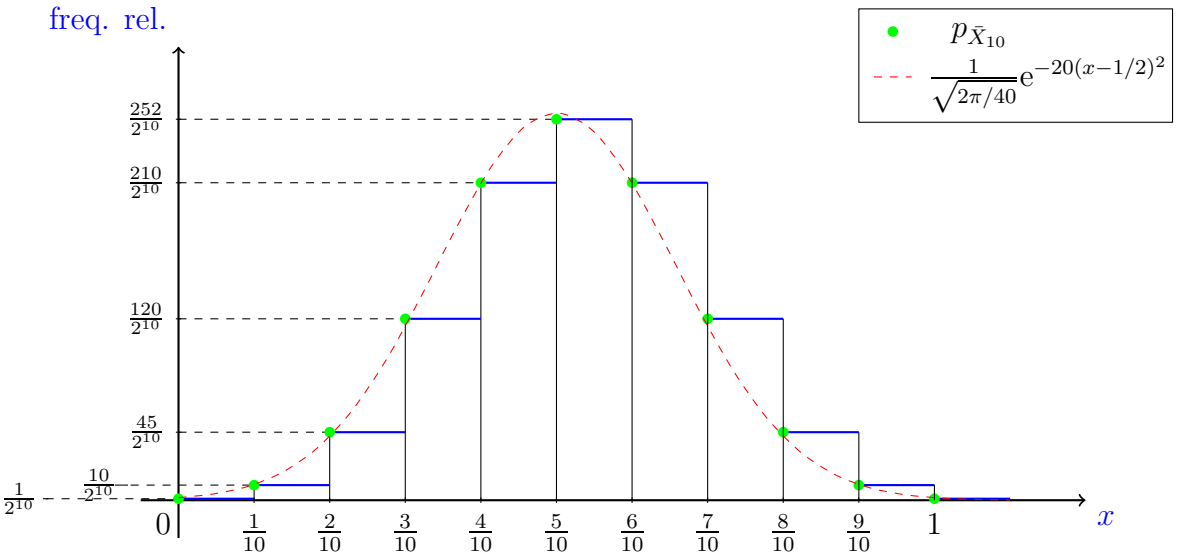
D'altra parte la variabile aleatoria

$$\frac{1}{2} + \frac{1}{2\sqrt{10}}\, Z$$

ha legge normale di media $\frac{1}{2}$ e varianza $\frac{1}{40}$, quindi ha densità (continua) data da

$$\frac{1}{\sqrt{2\pi/40}}\, \mathrm{e}^{-20\,(x-1/2)^2}.$$

Costruendo l'istogramma di $\bar{X}_{10}$, utilizzando 11 classi di uguale ampiezza a partire dal valore $a = 0$ al valore $b = 11$ (l'ampiezza è dunque 1/10), si ottiene



Notiamo che già per $n = 10$ l'istogramma è una buona approssimazione della densità gaussiana di media $\frac{1}{2}$ e varianza $\frac{1}{40}$.

*4.4 Appendice**

In questa appendice riportiamo una dimostrazione del Teorema centrale del limite utilizzando la *funzione generatrice dei momenti*, uno strumento nuovo che ora definiamo.

Momenti di una variabile aleatoria. Sia X una variabile aleatoria. Per ogni numero naturale $k = 1, 2, \ldots$, si chiama ***momento di ordine*** k la quantità

$$\mu'_k \;:=\; \mathbb{E}\big[X^k\big].$$

Si chiama invece ***momento centrato di ordine*** k la quantità

$$\mu_k \;:=\; \mathbb{E}\big[\big(X-\mathbb{E}[X]\big)^k\big].$$

Più precisamente, se X è discreta allora

$$\mu'_k \;=\; \sum_i x_i^k\, p_X(x_i), \qquad\qquad \mu_k \;=\; \sum_i \big(x_i-\mathbb{E}[X]\big)^k\, p_X(x_i). \tag{4.7}$$

Se X è continua vale invece che

$$\mu'_k \;=\; \int_{-\infty}^{+\infty} x^k\, f_X(x)\,\mathrm{d}x, \qquad\qquad \mu_k \;=\; \int_{-\infty}^{+\infty} \big(x-\mathbb{E}[X]\big)^k\, f_X(x)\,\mathrm{d}x. \tag{4.8}$$

Osservazione 1. *Si noti che μ_1 coincide con il valore atteso $\mathbb{E}[X]$, mentre μ_2 coincide con la varianza* $\mathrm{Var}(X)$.

Osservazione 2. *Quando X è una v.a. discreta e il supporto $\mathcal{S}_X$ è un insieme finito allora i momenti esistono sempre. Questo potrebbe non essere vero quando $\mathcal{S}_X$ è un insieme infinito numerabile, dato che non è assicurato che le serie che compaiono in* (4.7) *siano convergenti. La stessa osservazione vale per gli integrali in* (4.8).

Funzione generatrice dei momenti. La *funzione generatrice dei momenti* di una variabile aleatoria X è data da

$$M_X(t) \;:=\; \mathbb{E}\big[\mathrm{e}^{tX}\big], \qquad \forall\, t\in\mathbb{R}.$$

Più precisamente, se X è discreta allora

$$M_X(t) \;=\; \sum_i \mathrm{e}^{tx_i}\, p_X(x_i), \qquad \forall\, t\in\mathbb{R}.$$

Se X è continua vale invece che

$$M_X(t) \;=\; \int_{-\infty}^{+\infty} \mathrm{e}^{tx}\, f_X(x)\,\mathrm{d}x, \qquad \forall\, t\in\mathbb{R}.$$

Osservazione. *Come accade per i momenti (si veda l'*Osservazione 2 *riportata qui sopra), anche la funzione generatrice dei momenti potrebbe non essere definita per ogni $t\in\mathbb{R}$. Tuttavia se X è una v.a. discreta e il supporto $\mathcal{S}_X$ è un insieme finito, allora la funzione generatrice dei momenti è definita per ogni $t\in\mathbb{R}$.*

Proprietà della funzione generatrice dei momenti.

1) *Derivate di M_X e momenti della v.a. X.* Come suggerisce il suo nome, la funzione generatrice dei momenti permette di calcolare i momenti della v.a. X. Infatti, calcolando la derivata k-esima della funzione generatrice dei momenti, otteniamo

$$\frac{d^k M_X}{dt^k}(t) \;=\; \mathbb{E}\big[X^k\,\mathrm{e}^{tX}\big], \qquad \forall\, t\in\mathbb{R}.$$

Quindi vale che

$$\frac{d^k M_X}{dt^k}(0) \;=\; \mathbb{E}\big[X^k\big] \;=\; \mu'_k.$$

In particolare, si ha che

$$M_X(0) \;=\; 1, \qquad\qquad M'_X(0) \;=\; \mu'_1, \qquad\qquad M''_X(0) \;=\; \mu'_2. \tag{4.9}$$

Si può dimostrare che la funzione generatrice dei momenti è data dalla serie di Taylor

$$M_X(t) \;=\; \sum_{k=0}^{+\infty} \mathbb{E}\big[X^k\big]\, \frac{t^k}{k!}, \qquad \forall\, t \in \mathbb{R}.$$

2) *Funzione generatrice dei momenti di una v.a. normale standard.* Sia $Z \sim \mathcal{N}(0,1)$. Allora vale che

$$M_Z(t) \;=\; \mathrm{e}^{\frac{1}{2}t^2}, \qquad \forall\, t \in \mathbb{R}. \tag{4.10}$$

Infatti

$$\begin{aligned} M_Z(t) \;&=\; \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{tx}\, \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}x \;=\; \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}(x^2-2tx)}\, \mathrm{d}x \\ &=\; \mathrm{e}^{\frac{1}{2}t^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}(x^2-2tx+t^2)}\, \mathrm{d}x. \end{aligned}$$

Resta dunque da dimostrare che $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}(x^2-2tx+t^2)}\, \mathrm{d}x = 1$. Poiché $(x^2-2tx+t^2) = (x-t)^2$, otteniamo

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}(x^2-2tx+t^2)}\, \mathrm{d}x \;=\; \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}(x-t)^2}\, \mathrm{d}x.$$

Si noti che $\frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}(x-t)^2}$ coincide con la densità continua di una v.a. $X \sim \mathcal{N}(t,1)$. Quindi, per definizione di densità, si ha che

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}(x-t)^2}\, \mathrm{d}x \;=\; 1.$$

3) *Funzione generatrice dei momenti e legge della v.a. X.* La funzione generatrice dei momenti determina completamente la legge di X:

- se conosco $\mathbb{P}_X$ allora conosco M_X (questo segue direttamente dalla definizione di M_X),
- ma vale anche il viceversa, cioè se conosco M_X allora conosco $\mathbb{P}_X$ (omettiamo la dimostrazione di questo risultato, che segue dalla proprietà secondo cui se si conoscono tutti i momenti della v.a. X, quindi $\mathbb{E}[X^k]$ per ogni k, allora si conosce $\mathbb{P}_X$).

4) *Funzione generatrice dei momenti e convergenza in legge.* Come abbiamo visto al punto 3), la funzione generatrice dei momenti determina completamente la *legge*. Ora vedremo che la funzione generatrice dei momenti permette anche di determinare la *convergenza in legge*. Più precisamente, sia $\bar{Z}_n$ la v.a. che si considera nel Teorema centrale del limite. Sia inoltre $Z \sim \mathcal{N}(0,1)$. Sia infine $M_{\bar{Z}_n}$ la funzione generatrice dei momenti di $\bar{Z}_n$, quindi

$$M_{\bar{Z}_n}(t) \;:=\; \mathbb{E}\big[\mathrm{e}^{t\bar{Z}_n}\big], \qquad \forall\, t \in \mathbb{R}.$$

Allora vale il seguente risultato: se

$$\lim_{n\to+\infty} M_{\bar{Z}_n}(t) \;=\; M_Z(t), \qquad \forall\, t \in \mathbb{R},$$

dove ricordiamo che $M_Z(t) = \mathrm{e}^{\frac{1}{2}t^2}$ (si veda (4.10)), allora segue che

$$\bar{Z}_n \;\xrightarrow[n\to+\infty]{\mathscr{L}}\; Z.$$

Più in generale, data una successione di v.a. $(X_n)_n$ ed una v.a. Z vale la seguente proprietà: se

$$\lim_{n\to+\infty} M_{X_n}(t) \;=\; M_Z(t), \qquad \forall\, t \in \mathbb{R}$$

allora segue che

$$X_n \;\xrightarrow[n\to+\infty]{\mathscr{L}}\; Z,$$

ovvero X_n *converge in legge* a Z.

Dimostrazione del Teorema centrale del limite.[15]

Per semplicità supponiamo che $\mu = 0$ e $\sigma = 1$. Indichiamo la funzione generatrice dei momenti M_{X_1} della v.a. X_1 semplicemente con M. Si noti che M è la funzione generatrice dei momenti anche delle v.a. $X_2, \ldots, X_n, \ldots$, dato che tali v.a. sono identicamente distribuite.
Dobbiamo dimostrare che

$$\bar{Z}_n \;\xrightarrow[n\to+\infty]{\mathscr{L}}\; Z.$$

Per quanto detto nel paragrafo riportato qui sopra, è sufficiente dimostrare che

$$\lim_{n\to+\infty} M_{\bar{Z}_n}(t) \;=\; \mathrm{e}^{\frac{1}{2}t^2}, \qquad \forall\, t \in \mathbb{R}.$$

Calcoliamo dunque $M_{\bar{Z}_n}$ ed esprimiamola in termini di M, la funzione generatrice dei momenti di X_1. Ricordiamo che

$$\bar{Z}_n \;=\; \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}, \qquad\qquad \bar{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n}.$$

[15] **N.B.** Tale dimostrazione richiede la seguente ipotesi aggiuntiva: per ogni $t \in \mathbb{R}$ esiste la funzione generatrice dei momenti $M_{X_1}(t)$ della v.a. X_1. Si noti che questa ipotesi è verificata ad esempio quando X_1 è una v.a. discreta e il supporto $\mathcal{S}_{X_1}$ è un insieme finito.

Essendo $\mu = 0$ e $\sigma = 1$, si ottiene

$$\bar{Z}_n \;=\; \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

Quindi

$$M_{\bar{Z}_n}(t) \;=\; \mathbb{E}\big[\mathrm{e}^{t\bar{Z}_n}\big] \;=\; \mathbb{E}\Big[\mathrm{e}^{\frac{1}{\sqrt{n}}tX_1+\cdots+\frac{1}{\sqrt{n}}tX_n}\Big] \;=\; \mathbb{E}\Big[\mathrm{e}^{\frac{1}{\sqrt{n}}tX_1}\cdots\mathrm{e}^{\frac{1}{\sqrt{n}}tX_n}\Big].$$

Poiché $X_1, \ldots, X_n$ sono v.a. *indipendenti*, segue che anche le v.a. $\mathrm{e}^{\frac{1}{\sqrt{n}}tX_1}, \ldots, \mathrm{e}^{\frac{1}{\sqrt{n}}tX_n}$ sono indipendenti. Ricordando allora che il valore atteso del prodotto di v.a. indipendenti è uguale al prodotto dei valori attesi, otteniamo

$$M_{\bar{Z}_n}(t) \;=\; \mathbb{E}\Big[\mathrm{e}^{\frac{1}{\sqrt{n}}tX_1}\Big]\cdots\mathbb{E}\Big[\mathrm{e}^{\frac{1}{\sqrt{n}}tX_n}\Big].$$

Dato che $X_1, \ldots, X_n$ sono v.a. *identicamente distribuite*, i valori attesi $\mathbb{E}[\mathrm{e}^{\frac{1}{\sqrt{n}}tX_1}], \ldots, \mathbb{E}[\mathrm{e}^{\frac{1}{\sqrt{n}}tX_n}]$ sono identici e pari a $M\big(\frac{1}{\sqrt{n}}t\big)$. Perciò

$$M_{\bar{Z}_n}(t) \;=\; M\left(\frac{1}{\sqrt{n}}\,t\right)^n.$$

Resta ora da calcolare il limite per $n \to +\infty$. Ricordando che $M(0) = 1$ (si veda (4.9)), il limite qui sopra è nella forma indeterminata 1^∞. Procediamo come si fa normalmente per studiare questa forma indeterminata, ovvero passando ai logaritmi. Più precisamente, mostriamo che

$$\lim_{n\to+\infty} \log\big(M_{\bar{Z}_n}(t)\big) \;=\; \frac{1}{2}t^2, \qquad \forall\, t \in \mathbb{R}.$$

Si ha che

$$\lim_{n\to+\infty} \log\big(M_{\bar{Z}_n}(t)\big) \;=\; \lim_{n\to+\infty} \log\left(M\left(\frac{1}{\sqrt{n}}\,t\right)^n\right) \;=\; \lim_{n\to+\infty} \frac{\log M\big(\frac{1}{\sqrt{n}}\,t\big)}{\frac{1}{n}}.$$

Quest'ultimo limite è nella forma indeterminata $\frac{0}{0}$. Applicando la regola di De L'Hôpital, si ottiene (per semplificare i conti, poniamo $y = 1/\sqrt{n}$)

$$\lim_{n\to+\infty} \frac{\log M\big(\frac{1}{\sqrt{n}}\,t\big)}{\frac{1}{n}} \;=\; \lim_{y\to0} \frac{\log M(y\,t)}{y^2} \;\overset{\text{(H)}}{=}\; \lim_{y\to0} \frac{t\,M'(y\,t)}{2y\,M(y\,t)} \;=\; \frac{1}{2}\,t\lim_{y\to0} \frac{M'(y\,t)}{y},$$

dove nell'ultima uguaglianza abbiamo eliminato il termine $M(y\,t)$, dato che $M(y\,t)$ tende a $M(0) = 1$ quando $y \to 0$. Ricordando che $\mu = 0$, quindi $M'(0) = 0$ (si veda (4.9)), abbiamo di nuovo una forma indeterminata $\frac{0}{0}$. Applicando ancora una volta la regola di De L'Hôpital, si ottiene infine

$$\frac{1}{2}\,t\lim_{n\to+\infty} \frac{M'(y\,t)}{y} \;\overset{\text{(H)}}{=}\; \frac{1}{2}\,t^2\lim_{n\to+\infty} M''(y\,t) \;=\; \frac{1}{2}\,t^2\,M''(0) \;=\; \frac{1}{2}\,t^2.$$

□