

Capitolo 6

Metodi Statistici

6.1 Introduzione

Come abbiamo visto nei capitoli precedenti i valori misurati delle grandezze fisiche sono soggetti ad incertezza e quindi possono essere considerati come variabili aleatorie generate da opportune distribuzioni di probabilità. Lo studio delle proprietà di un numero finito di variabili estratte da distribuzioni di probabilità è l'oggetto delle *discipline statistiche*. Per questo motivo l'analisi dei dati sperimentali trova nelle discipline statistiche le giustificazioni razionali, le metodologie e gli strumenti matematici necessari a stimare i valori delle grandezze che governano i fenomeni aleatori. Infatti la statistica è la disciplina che studia in modo quantitativo e qualitativo i fenomeni governati da parametri i cui valori sono soggetti ad incertezza. Come vedremo nei prossimi paragrafi, le discipline statistiche forniscono una giustificazione razionale a formule e relazioni già precedentemente introdotte e giustificate in modo intuitivo, come ad esempio l'uso della media aritmetica nelle misurazioni ripetute per la stima del valore medio e l'aumento della precisione della misura all'aumentare del numero delle misurazioni.

6.2 Statistica di base

Per l'applicazione delle metodologie statistiche ai dati sperimentali iniziamo introducendo la terminologia e i concetti di base. Consideriamo un esperimento in cui si eseguono N misurazioni ripetute di una grandezza X affetta da incertezza di tipo A e indichiamo con (x_1, \dots, x_N) i risultati di tali misurazioni. In statistica l'insieme di queste misure è detto un "campione" (in inglese *sample*) di dimensione N della variabile causale X . La distribuzione di probabilità che descrive i risultati delle N misurazioni e di cui a priori molte volte si ignorano i parametri, è detta "popolazione di riferimento" (in inglese *parent distribution*). Per chiarire questi concetti facciamo alcuni esempi.

Esempio 1. Lanciamo una moneta reale N volte e supponiamo che k volte sia uscita testa e $N - k$ volte sia uscita croce. Il "campione" in questo caso è formato dagli N risultati (k croci e $N - k$ teste). Se siamo sicuri che la moneta non sia truccata allora la popolazione di riferimento è la distribuzione binomiale $\mathcal{B}_N(k, 1/2)$.

Esempio 2. Lanciamo una puntina da disegno N volte in modo casuale su di un piano. La puntina può cadere in due modi: con la punta verso l'alto oppure con la punta che tocca il piano. E' molto difficile stabilire a priori la probabilità dei due modi. Il campione in questo

caso è ancora formato dagli N risultati (k eventi in cui la punta è verso l'alto e $N - k$ nell'altro modo). La popolazione di riferimento è la distribuzione binomiale $\mathcal{B}_N(k, p)$ con il parametro p incognito da stimare.

Esempio 3. Un contatore di raggi cosmici è acceso per un tempo di 5 minuti. Supponiamo che in questo intervallo di tempo conti 12 eventi (passaggi di raggi cosmici). In questo caso il campione consiste nei 12 eventi in 5 minuti e la popolazione di riferimento è una distribuzione di Poisson il cui parametro, il valore medio, è incognito e deve essere stimato dai dati misurati.

Lo scopo dell'analisi dei dati sperimentali è quello di *stimare* i parametri della popolazione di riferimento utilizzando i dati del campione (l'esperimento). Questo procedimento è noto anche con il nome di *inferenza statistica* o brevemente *inferenza*. Le modalità e le varie tecniche utilizzate per la stima dei parametri delle distribuzioni di probabilità saranno trattate in dettaglio nel capitolo 8.

Come accennato nell'introduzione a questo capitolo, per ottenere una dimostrazione rigorosa di alcune regole semiempiriche come quella che indica nella media aritmetica la stima migliore del valore medio di una distribuzione di probabilità e quella sul miglioramento della precisione di una misura all'aumentare del numero delle misurazioni, seguiremo il percorso indicato nei seguenti punti:

1. Inizieremo con il dimostrare la **disuguaglianza di Chebyshev** che, per quanto di scarsa utilità pratica, è la base per la dimostrazione della:
2. **legge dei grandi numeri** mediante la quale si dimostra che la media aritmetica campionaria di n realizzazioni di un esperimento converge, all'aumentare di n , verso il valore medio della popolazione di riferimento, qualsiasi sia la distribuzione di probabilità .
3. Enunceremo il **teorema del limite centrale** che afferma, sotto ipotesi molto generali, che la somma di n variabili casuali, qualsiasi siano le distribuzioni di probabilità delle singole variabili, converge verso una *distribuzione gaussiana* al crescere di n
4. Il teorema del limite centrale applicato agli errori di misura ha come corollario che in molti casi, anche se non in tutti, le misure e le loro medie aritmetiche si distribuiscono in modo gaussiano

6.2.1 Disuguaglianza Chebyshev

Cominciamo con il dimostrare la disuguaglianza di Chebyshev nota anche come disuguaglianza di Bienaymé. Qualitativamente la disuguaglianza di Chebyshev dice che in qualsiasi distribuzione di probabilità, con una deviazione standard finita, i valori della variabile aleatoria si addensano attorno al suo valore medio. Sia $f(x)$ la *pdf* della variabile aleatoria x , di valore medio μ e varianza σ^2 . La disuguaglianza di Chebyshev dice che per ogni $t > 0$ vale la relazione:

$$P(|x - \mu| > t\sigma) \leq \frac{1}{t^2}$$

Per la dimostrazione ricordiamo che per definizione si ha:

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Essendo l'integrando sempre positivo, vale la seguente disuguaglianza:

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \geq \int_{-\infty}^{\mu - t\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + t\sigma}^{+\infty} (x - \mu)^2 f(x) dx$$

Inoltre, poiché nei due intervalli di integrazione risulta $(x - \mu)^2 \geq t^2 \sigma^2$, potremo scrivere a maggior ragione:

$$\sigma^2 \geq t^2 \sigma^2 \left(\int_{-\infty}^{\mu - t\sigma} f(x) dx + \int_{\mu + t\sigma}^{+\infty} f(x) dx \right) = t^2 \sigma^2 P(|x - \mu| > t\sigma)$$

da cui si ottiene infine

$$P(|x - \mu| > t\sigma) \leq \frac{1}{t^2} \quad (6.1)$$

La relazione (6.1) è la **disuguaglianza di Chebyshev** e vale per tutte le *pdf* con deviazione standard finita. Quindi la probabilità di osservare valori della variabile aleatoria che distino dal valore medio più di t volte la deviazione standard è sempre minore dell'inverso del quadrato di t .

I limiti posti dalla disuguaglianza (6.1) non sono molti stringenti. In realtà per le distribuzioni più comuni il primo membro della (6.1) è molto minore del secondo come è mostrato nella tabella 6.1 in cui si confronta il limite posto dalla disuguaglianza di Chebyshev con il contenuto di probabilità della distribuzione normale $\mathcal{N}(0, 1)$.

Tabella 6.1: Confronto tra i limiti della disuguaglianza di Chebyshev (6.1) e la normale. Nella colonna "Normale" è riportato il valore: $1 - (1/\sqrt{2\pi}) \int_{-t}^{+t} e^{-z^2/2} dz$, che corrisponde a $P(|x - \mu| > t\sigma)$

t	Chebyshev	Normale
1	1.00	0.317
2	0.25	0.045
3	0.11	0.003

6.2.2 La legge dei grandi numeri

La *legge dei grandi numeri* formulata inizialmente dal matematico svizzero Jakob Bernoulli nel 1713 (nota per questo anche come teorema di Bernoulli), afferma che in un campione di dimensione sufficientemente grande la media aritmetica dei valori è *rappresentativa del valore medio della popolazione di riferimento*.

Per dimostrare la legge dei grandi numeri partiamo dalla disuguaglianza di Chebyshev (6.1) nel caso in cui la variabile aleatoria sia \bar{x} , *media aritmetica* di n variabili x_i indipendenti e appartenenti alla stessa distribuzione di probabilità (come n misure della stessa grandezza). Siano μ il valore medio delle x_i e σ la loro deviazione standard, allora utilizzando la (5.14) oppure la successiva (6.5), la deviazione standard di \bar{x} è: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Sostituendo questi dati nella (6.1) si ha:

$$P(|\bar{x} - \mu| \geq t \frac{\sigma}{\sqrt{n}}) \leq \frac{1}{t^2}$$

Poniamo $\epsilon = t\sigma/\sqrt{n}$ ovvero $t = \epsilon\sqrt{n}/\sigma$ per cui la relazione precedente si scrive:

$$P(|\bar{x} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}$$

la relazione precedente è equivalente a

$$P(|\bar{x} - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2 n} \quad (6.2)$$

La (6.2) rappresenta la legge dei grandi numeri e si legge:

scelto ϵ comunque piccolo, la probabilità che la media aritmetica \bar{x} scarti dal valore atteso μ meno di ϵ tende ad 1 al crescere di n .

Si noti che la (6.2) **non** dice che $\bar{x} = \mu$, ma solo che la *probabilità* che $\bar{x} = \mu$ tende a 1 al crescere di n . In certi testi questa frase è sintetizzata nell'espressione "evento praticamente certo".

La legge empirica del caso. La legge dei grandi numeri o teorema di Bernoulli è spesso confusa e associata con la legge empirica del caso che, diversamente dalla precedente, non è un teorema dimostrabile ma è, appunto, un'osservazione empirica che permette di valutare il valore ignoto della probabilità di un evento dalla sua frequenza in un grande numero di prove ripetute (nelle stesse condizioni). La legge empirica del caso si basa sull'osservazione sperimentale che nei casi in cui si ritiene di conoscere a priori la probabilità di un evento¹ la frequenza con cui l'evento si presenta si avvicina, all'aumentare del numero di prove, al valore della probabilità noto a priori.

Su questa osservazione si basa l'interpretazione frequentista della probabilità.

Come verifica di questa legge consideriamo un esperimento nel quale si lancia ripetutamente una moneta reale e si registra il numero delle volte che esce testa, considerato come evento favorevole. Per ogni lancio si aggiorna il valore della frequenza con cui è uscita testa: se dopo n lanci si sono avute k teste, la frequenza di testa è k/n . Empiricamente si osserva che il rapporto k/n , al crescere di n , tende ad un valore stabile che definisce la probabilità (in senso frequentista²) dell'evento. Nella figura 6.1 si mostrano gli esiti di tre esperimenti del lancio di una moneta ben equilibrata.

6.2.3 Il teorema del limite centrale

Il teorema limite centrale (TLC) è uno dei più importanti risultati della statistica e afferma che se x_1, x_2, \dots, x_n sono n variabili casuali indipendenti, con distribuzione di probabilità *qualsiasi* ma aventi valori attesi dello stesso ordine di grandezza e deviazioni standard finite e dello stesso ordine di grandezza allora una qualsiasi combinazione lineare y di queste variabili ha una distribuzione gaussiana. Indichiamo con y una generica combinazione lineare delle x_i :

$$y = a_1 x_1 + a_2 x_2 + \dots + a_n x_n \quad (a_i \text{ costanti}) \quad (6.3)$$

e poniamo $\mu_i = E[x_i]$ e $\sigma_i^2 = \text{Var}[x_i]$. Valore medio e varianza di y sono date da:

$$\mathbb{E}[y] \equiv \mu = \sum a_i \mu_i \quad \text{Var}[y] \equiv \sigma^2 = \sum a_i^2 \sigma_i^2$$

la prima relazione deriva dalla linearità dell'operatore valore atteso e la seconda dalle proprietà della varianza applicate alla somma di variabili indipendenti (vedi l'equazione (5.51)). Il

¹La conoscenza a priori (e a posteriori) della probabilità di un evento *reale* non è *mai* completamente nota.

²La definizione frequentista di probabilità prevederebbe un numero infinito di lanci.

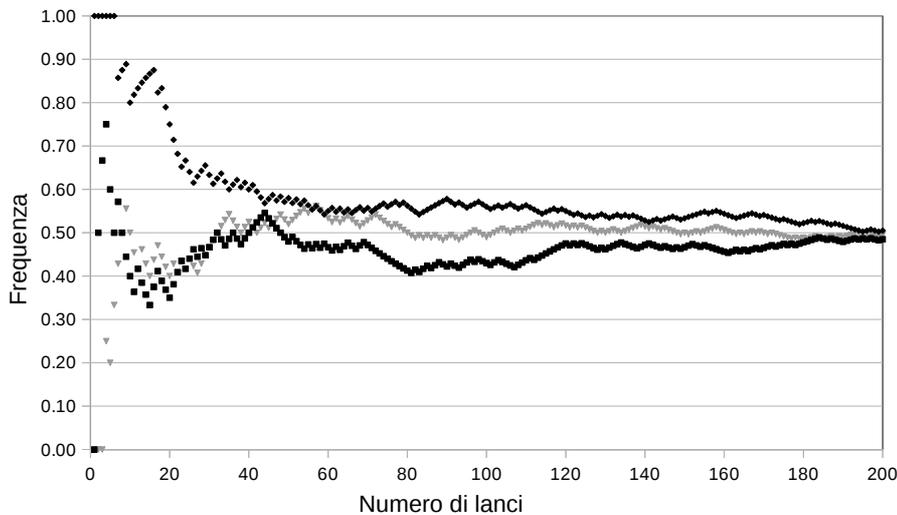


Figura 6.1: Legge empirica del caso. Frequenza dell'uscita di testa in tre serie di 200 lanci di una moneta non truccata

teorema del limite centrale afferma che la funzione di distribuzione della densità di probabilità di y è una gaussiana della forma:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$$

La dimostrazione matematica di questo teorema è piuttosto complessa e richiede strumenti che verranno appresi più avanti nel corso di studi.

Come regola approssimata si verifica che la somma di circa 30 variabili casuali, con distribuzioni di probabilità che soddisfano le condizioni sul valore medio e sulla varianza prima dette, ha una distribuzione molto prossima a quella di una gaussiana. In proposito si veda l'esempio in figura 6.2 che dimostra in modo empirico l'enunciato del teorema del limite centrale. Infatti aumentando il numero delle variabili aleatorie sommate, la distribuzione della loro somma tende ad assumere la forma gaussiana. Nella didascalia della figura sono dati i dettagli delle distribuzioni mostrate.

Un altro esempio³ che dimostra quanto rapidamente la somma di distribuzioni diverse può convergere ad una distribuzione normale consiste nel considerare le seguenti quattro distribuzioni indipendenti:

1. Distribuzione uniforme: $f(x) = 1/4$ per $0 < x < 4$, 0 altrove
2. Distribuzione esponenziale: $f(x) = e^{-x}$ per $x > 0$, 0 altrove
3. Distribuzione t-Student⁴ con $\nu = 5$: $f(x) \propto (1 + x^2/\nu)^{-(\nu+1)/2}$ con $-\infty < x < +\infty$
4. Distribuzione esponenziale doppia: $f(x) = (1/2)e^{-|x|}$ con $-\infty < x < +\infty$

³Esempio tratto dal testo di Robin Willink *Measurement Uncertainty and Probability* - Cambridge University Press

⁴La distribuzione t di Student è descritta nel paragrafo 5.9.4.

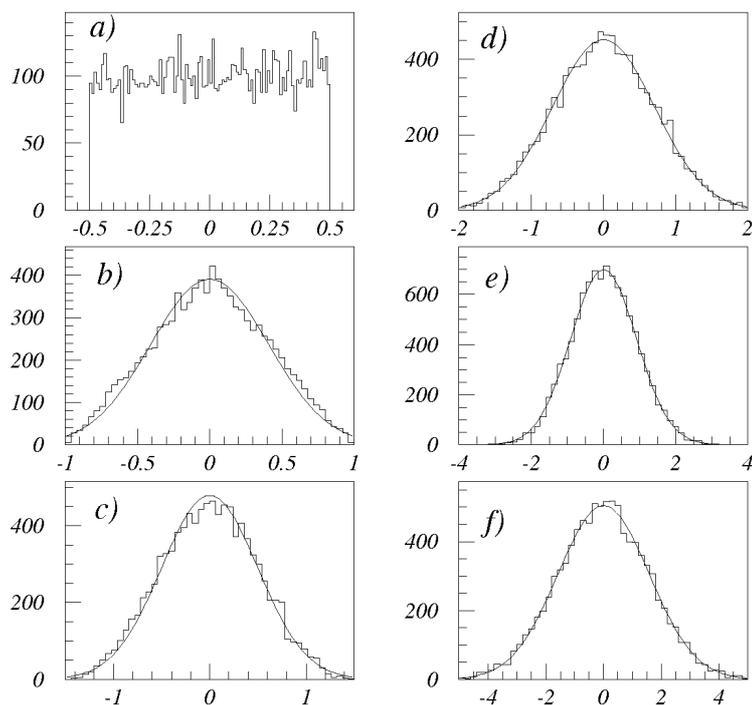


Figura 6.2: Il teorema del limite centrale. Istogramma a) distribuzione uniforme di nell'intervallo $(-0.5, +0.5)$; da b) a f) distribuzioni rispettivamente della somma di 2, 3, 6, 10 e 30 variabili con la distribuzione uniforme mostrata in a); ad ogni somma è stato sottratto il valore medio in modo che risultino centrate a zero. Ad ogni istogramma è sovrapposta una gaussiana, opportunamente normalizzata, con la stessa deviazione standard dell'istogramma. Tutti gli istogrammi hanno 10·000 eventi. La figura mostra come all'aumentare del numero delle variabili aleatorie sommate la distribuzione tende ad assumere una forma che si avvicina a quella della gaussiana.

Indicando con E_i , ($i = 1, 2, 3, 4$) i numeri estratti dalle quattro distribuzioni descritte, definiamo la variabile aleatoria E come somma delle variabili delle sopra definite:

$$E = E_1 + E_2 + E_3 + E_4$$

Estraendo molte quaterne di variabili E_i , ($i = 1, 2, 3, 4$) possiamo generare un campione della variabile E la cui distribuzione di probabilità approssima molto bene quella normale come dimostrato nella figura 6.3.

6.2.4 Teorema del limite centrale e incertezze di misura

Nello studio delle incertezze di misura il teorema del limite centrale ha due importanti applicazioni. La prima riguarda l'osservazione sperimentale che le misure di alcune grandezze fisiche hanno una distribuzione a campana bene approssimata da una curva gaussiana. In questi casi si ipotizza che le incertezze siano dovute ad una somma di molteplici fenomeni aleatori, ciascuno con la propria distribuzione, che si sommano in modo incoerente. In ogni misurazione ciascuno dei detti fenomeni aleatori genererà un *errore* di misura E_k , e l'errore totale E sulla misura sarà:

$$E = E_1 + E_2 + E_3 \dots$$

Se il numero totale degli errori è sufficientemente elevato, allora E , per il TLC, tenderà ad avere una distribuzione gaussiana.

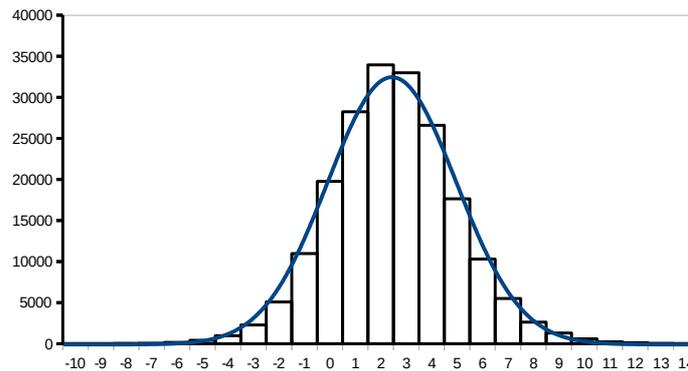


Figura 6.3: Il teorema del limite centrale. L'istogramma è quello della variabile E definita nel testo. Si dimostra che la variabile E ha media 3 e varianza 6 (la dimostrazione è lasciata come utile esercizio). La curva continua è una gaussiana con media e varianza uguali a quelle della variabile aleatoria E . Si noti che, in questo caso, la somma di sole 4 variabili aleatorie è già bene approssimata da una gaussiana.

Tuttavia poiché nella realtà il numero degli errori è sempre limitato, la distribuzione di E devia da quella gaussiana tipicamente nelle *code* corrispondenti a zone con una distanza maggiore di circa due deviazioni standard dal valore medio della gaussiana.

La seconda applicazione riguarda la distribuzione di probabilità delle medie campionarie. Consideriamo un esperimento in cui si eseguano N misurazioni ripetute di una grandezza X affetta da incertezza di tipo A e indichiamo con (x_1, \dots, x_N) i risultati di tali misurazioni. La media campionaria di queste misure è:

$$\bar{x} = \frac{\sum x_i}{N}$$

poiché \bar{x} è una combinazione lineare delle grandezze x_i , se il numero di misure N è sufficientemente elevato la distribuzione di probabilità di \bar{x} tende ad essere gaussiana, qualsiasi sia la distribuzione delle x_i .

6.3 Media e varianza campionaria

Consideriamo un esperimento nel quale si ripete n volte la misurazione di una grandezza nelle stesse condizioni sperimentali. Supponiamo inoltre che alla misurazione della grandezza sia associata un'incertezza di tipo A (vedi il paragrafo 3.6). Da quanto detto, ciascuna delle misure, che indichiamo con x_1, x_2, \dots, x_n , può essere considerata generata dalla stessa popolazione di riferimento di valore medio μ e varianza σ^2 . Riferendoci alle definizioni del paragrafo 6.2 diremo che l'insieme delle n misure è il "campione statistico" e la distribuzione di probabilità con valore atteso μ e varianza σ^2 è la "popolazione di riferimento". Anche la media campionaria \bar{x} delle n misure ha valore atteso μ , infatti:

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{n} \sum_i^n x_i\right] = \frac{\mathbb{E}[x_1 + x_2 + \dots + x_n]}{n} = \frac{1}{n} \sum_i^n \mathbb{E}[x_i] = \mu \quad (6.4)$$

Si noti come questa relazione, per altro molto intuitiva, trova una sua giustificazione teorica anche nella legge dei grandi numeri (6.2). Quindi:

Il valore atteso della media campionaria di n misure, affette da incertezze di tipo A , coincide con il valore atteso della popolazione di riferimento

Supponiamo ora di conoscere la deviazione standard σ delle x_i (vedremo nel seguito come valutarla) e calcoliamo la deviazione standard di \bar{x} . Dalle proprietà della varianza si ha:

$$\text{Var}[\bar{x}] = \text{Var}\left[\frac{1}{n} \sum_i^n x_i\right] = \frac{1}{n^2} \text{Var}\left[\sum_i^n x_i\right] = \frac{\sigma^2}{n} \quad (6.5)$$

quindi la deviazione standard della media campionaria di n misure è σ/\sqrt{n} ; concludiamo che:

la deviazione standard della media campionaria di n misure fornisce una valutazione del valore medio μ piú precisa di un fattore $1/\sqrt{n}$ rispetto alla singola misura.

Fino ad ora abbiamo ipotizzato di conoscere la deviazione standard σ della popolazione di riferimento che tuttavia, tranne rarissimi casi, è ignota e non può che essere stimata dagli n dati. Ricordando la definizione di varianza, vedi l'equazione (5.13), si potrebbe essere tentati di assumere come valutazione della varianza σ^2 l'espressione

$$s^{*2} = \frac{1}{n} \sum (x_i - \bar{x})^2 \quad (6.6)$$

tuttavia si dimostra che s^{*2} dà una stima sistematicamente minore⁵ della varianza σ^2 . Con un calcolo non difficile ma lungo (riportato nel paragrafo 6.4) si ottiene la stima non distorta della varianza la cui espressione è:

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \quad (6.7)$$

Una giustificazione intuitiva che s^{*2} sia una sottostima di σ^2 , varianza della popolazione di riferimento, si ottiene osservando che gli n dati x_i sono stati utilizzati anche per il calcolo di \bar{x} .

6.4 Formula della stima della varianza campionaria

Per dimostrare la (6.7) partiremo dalla seguente identità

$$x_i - \bar{x} = (x_i - \mu) - (\bar{x} - \mu)$$

quadrando e sommando sull'indice i fino a n ,

$$\begin{aligned} \sum (x_i - \bar{x})^2 &= \sum [(x_i - \mu) - (\bar{x} - \mu)]^2 = \\ &= \sum (x_i - \mu)^2 - 2(\bar{x} - \mu) \sum (x_i - \mu) + n(\bar{x} - \mu)^2 = \\ &= \sum (x_i - \mu)^2 - 2(\bar{x} - \mu)n(\bar{x} - \mu) + n(\bar{x} - \mu)^2 = \\ &= \sum (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \end{aligned} \quad (6.8)$$

⁵In teoria degli estimatori questa stima della varianza è detta distorta

dove si è utilizzata la definizione $\bar{x} = (1/n) \sum x_i$. Dividendo la (6.8) per n e prendendo il valore atteso dei due membri, otteniamo:

$$\mathbb{E} \left[\frac{1}{n} \sum (x_i - \bar{x})^2 \right] = \mathbb{E} \left[\frac{\sum (x_i - \mu)^2}{n} - (\bar{x} - \mu)^2 \right] = \mathbb{E} \left[\frac{1}{n} \sum (x_i - \mu)^2 \right] - \mathbb{E} [(\bar{x} - \mu)^2] \quad (6.9)$$

Tenendo conto che:

$$\mathbb{E} \left[\frac{1}{n} \sum (x_i - \mu)^2 \right] = \frac{1}{n} \sum \mathbb{E} [(x_i - \mu)^2] = \frac{1}{n} \sum \sigma^2 = \frac{1}{n} n \sigma^2 = \sigma^2$$

e che la varianza della media è σ^2/n , abbiamo:

$$\mathbb{E} [(\bar{x} - \mu)^2] = \frac{\sigma^2}{n}$$

la (6.9) diviene:

$$\mathbb{E} \left[\frac{1}{n} \sum (x_i - \bar{x})^2 \right] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2$$

Ricordando la (6.6) possiamo scrivere che la stima campionaria della varianza del campione è data da:

$$\sigma^2 = \frac{n}{n-1} s^{*2}$$

la deviazione standard otteniamo

$$\sigma = \sqrt{\frac{n}{n-1} s^{*2}} = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

