Corso di Laurea Specialistica in Biotecnologie per l'Industria e per l'Ambiente

Informatica Applicata ai Processi Biotecnologici

a.a. 2003-04

Note

Stefano Ferrari

14 aprile 2004

Riferimenti

- – sito: http://www.dti.unimi.it/∼fscotti (didattica)
  http://www.dti.unimi.it/∼ferrari (didattica)
  – email: ferrari@dti.unimi.it
- Mood, Graybill e Boes, "Introduzione alla statistica".

1 Introduzione

Incertezza

I vari tipi di incertezza.

Stimare grandezze.

Stimare relazioni tra grandezze.

2 Richiami di calcolo delle probabilità

2.1 Definizione di probabilità secondo la concezione classica (P. S. Laplace, 1749-1827)

Viene chiamata *probabilità a priori*. Corrisponde al concetto intuitivo di probabilità.

La probabilità $P(E)$ di un evento E è il rapporto fra il numero F dei casi favorevoli (al verificarsi di E) e il numero N dei casi possibili, giudicati egualmente possibili.

$$P(E) = \frac{F}{N}$$

$0 \leq P(E) \leq 1$

- $P(E) = 0$, se E è impossibile
- $P(E) = 1$, se E è certo

Note:

- Uno dei punti deboli della concezione classica è la condizione, pressoché impossibile da verificare, che tutti i casi in cui può manifestarsi il fenomeno siano egualmente possibili.
- La definizione si può applicare quando l'insieme dei casi è un insieme finito.

2.2 Definizione di probabilità secondo la concezione frequentista

Viene chiamata *probabilità a posteriori*. Corrisponde al concetto di probabilità sperimentale (nel senso di "basato sull'esperienza").

La concezione frequentista è basata sulla definizione di *frequenza relativa* di un evento.

Si definisce *frequenza relativa* di un evento in n prove effettuate nelle stesse condizioni, il rapporto fra il numero v delle prove nelle quali l'evento si è verificato e il numero n delle prove effettuate:

$$f = \frac{v}{n}$$

- se $f = 0$ l'evento non si è mai verificato in quelle n prove;
- se $f = 1$ ($v = n$) l'evento si è sempre verificato in quelle n prove.

Note:

- La frequenza dipende dal numero n delle prove fatte, per uno stesso n, la frequenza, può variare al variare del gruppo delle prove: se si lancia 100 volte una moneta e si presenta testa 54 volte, effettuando altri 100 lanci si può presentare 48 volte.
- Se il numero di prove è sufficientemente alto, il rapporto $\frac{v}{n}$ tende a stabilizzarsi.

2.3 Definizione di assiomatica di probabilità

Per superare le difficoltà legate alle definizioni precedenti di probabilità, si definiscono alcuni assiomi che descrivono le caratteristiche che la probabilità dovrebbe avere, e poi se ne deduce la teoria.

Alcune definizioni:

- Spazio dei campioni, Ω: insieme dei risultati di un esperimento.
- Spazio degli eventi $\boldsymbol{A}$: insieme dei sottoinsiemi di Ω.

Questi gli assiomi:

- $P(A) \geq 0 \ \ \forall A \in \boldsymbol{A}$
- $P(\Omega) = 1$
- se $A_1, A_2, \ldots$ è una sequenza di eventi che si mutuamente esclusivi, allora $P(\cup_i A_i) = \sum_i P(A_i)$

2.4 Regole di addizione e moltiplicazione

La probabilità che si verifichi almeno uno di due eventi è pari alla somma delle probabilità dei singoli eventi, meno la probabilità che si verifichino contemporaneamente:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Se A e B sono mutuamente esclusivi (se, cioè, il verificarsi di uno esclude il verificarsi dell'altro e viceversa):

$$P(A \cup B) = P(A) + P(B)$$

Esempio Consideriamo un dado non truccato: ogni faccia ha probabilità di uscire pari a $\frac{1}{6}$ ($P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$). La probabilità dell'evento $A = \{1, 2, 3\}$, cioè la probabilità che con un lancio esca una delle prime tre facce, è $P(A) = P(1)+P(2)+P(3) = \frac{1}{2}$, in quanto gli eventi 1, 2 e 3 sono mutuamente esclusivi. Analogamente, se $B = \{2, 4, 6\}$, cioè la probabilità che esca un numero pari, è $P(B) = P(2) + P(4) + P(6) = \frac{1}{2}$. La probabilità dell'evento $P(A \cup B)$, cioè la probabilità che esca un numero pari oppure fra i primi tre numeri, è $P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - P(2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6}$. Allo stesso risultato si giunge anche considerando che $A \cup B = \{1, 2, 3\} \cup \{2, 4, 6\} = \{1, 2, 3, 4, 6\}$ e quindi $P(A \cup B) = P(1) + P(2) + P(3) + P(4) + P(6) = \frac{5}{6}$.

Indipendenza A e B sono *indipendenti* se $P(A \cap B) = P(A)P(B)$.

Nota: l'indipendenza stocastica non significa che due eventi si escludono l'un l'altro o cose del genere. Il significato verrà chiarito a breve.

2.5 Probabilità Condizionata e Teorema di Bayes

La probabilita condizionata di A, dato che l'evento B è accaduto, è denotata con $P(A|B)$. Vale la relazione:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Ciò significa che se A e B sono eventi indipendenti, $P(A|B) = P(A)$: se due eventi sono indipendenti, il verificarsi di uno non modifica la probabilità del verificarsi dell'altro. In questo senso, per esempio, gli esiti di estrazioni del lotto successive sono indipendenti tra loro. Ne deriva che, sebbene sia altamente improbabile una lunga serie di assenze (o presenze) di uno stesso numero, il fatto che tale numero non esca (o sia uscito) in una estrazione, non cambia la probabilità che esso esca (o non esca) nella estrazione seguente.

Formula di Bayes Se $B_1, \ldots, B_n$ formano una partizione dello spazio dei campioni, e $P(B_i) > 0$ per $i = 1, \ldots, n$:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Regola moltiplicativa Se $B_1, \ldots, B_n$ sono eventi per cui $P(B_1 \cap \cdots \cap B_n) > 0$:

$$P(B_1 \cap \cdots \cap B_n) = P(B_1)P(B_2|B_1)P(B_3|B_1 \cap B_2) \cdots P(B_n|B_1 \cap \cdots \cap B_{n-1})$$

2.6 Variabile aleatoria

Il nome è fuorviante: non è una variabile e non è aleatoria. In realtà è una funzione: ha per dominio lo spazio degli eventi e come codominio la retta dei reali.

$X : \Omega \to \mathbb{R}$

Serve per poter associare un valore numerico ad ogni evento: con l'insieme $\Omega = \{testa, croce\}$ non facciamo i calcoli, con $\{0, 1\} \subseteq \mathbb{R}$ sì.

La variabile aleatoria induce una relazione di ordinamento tra gli eventi.

2.7 Funzione di distribuzione cumulativa

La funzione di distribuzione (cumulativa) di una variabile aleatoria X è definita come:

$$F_X(x) = P(X \leq x) = P(\omega : X(\omega) \leq x) \ \ \forall x \in \mathbb{R}$$

- è una funzione monotona non decrescente

- $\lim_{x \to -\infty} F_X(x) = 0$ (probabilità del verificarsi di nessun evento)
- $\lim_{x \to \infty} F_X(x) = 1$ (probabilità che si verifichi un qualsiasi evento)

Non abbiamo definito la probabilità di un singolo elemento, ma la probabilità di un insieme di elementi. Per esempio, $F_X(3)$ è pari alla probabilità del verificarsi un evento qualsiasi tra gli eventi per cui la variabile aleatoria X è minore di 3. Questo risolve parzialmente il problema di definire la probabilità di eventi di numerosità infinita.

Esempio Consideriamo l'esperimento del lancio della moneta, con la variabile aleatoria X che indica il numero di *testa*: $X(croce) = 0$, $X(testa) = 1$. La sua funzione di distribuzione di probabilità è:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

2.8 Funzione di densità

La funzione di densità di probabilità serve per definire più semplicemente la probabilità di un evento (ove possibile) o di un qualsiasi insieme di eventi (anche non contigui).

Se X è discreta ed assume valori $\{x_j\}$:

$$f_X(x) = \begin{cases} P(X = x_j), & x = x_j,\ j = 1,\ 2,\ \ldots,\ n \\ 0, & \text{altrimenti} \end{cases}$$

Se X è discreta, $F_X(\cdot)$ e $f_X(\cdot)$ ci danno la stessa informazione (da una ricaviamo l'altra e viceversa):

$$F_X(x) = \sum_{j : x \leq x_j} f_X x_j$$

$$f_X(x_j) = F_X x_j - F_X x_{j-1}$$

Se X è continua, esiste una funzione $f_X(\cdot)$ tale che $F_X(x) = \int_{-\infty}^{x} f_X(u) \mathrm{d}u$. $f_X(\cdot)$ è la funzione di densità di X. Anche per X continua, $F_X(\cdot)$ può essere ricavata da $f_X(\cdot)$ e viceversa:

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

per i punti x dove $F_X(x)$ è differenziabile (zero altrimenti).

Essa gode di due proprietà:

- $f_X(x) \geq 0$, per ogni valore di X
- $\int_{-\infty}^{\infty} = 1$

La funzione densità di probabilità serve per trattare la probabilità nei casi in cui gli eventi sono infiniti:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)\, \mathrm{d}x$$

2.9 Gli indicatori principali

Spesso si ha la necessità di riassumere con quantità numeriche le caratteristiche delle funzioni di probabilità. Vi sono diversi indicatori, i più notevoli saranno trattati nel seguito.

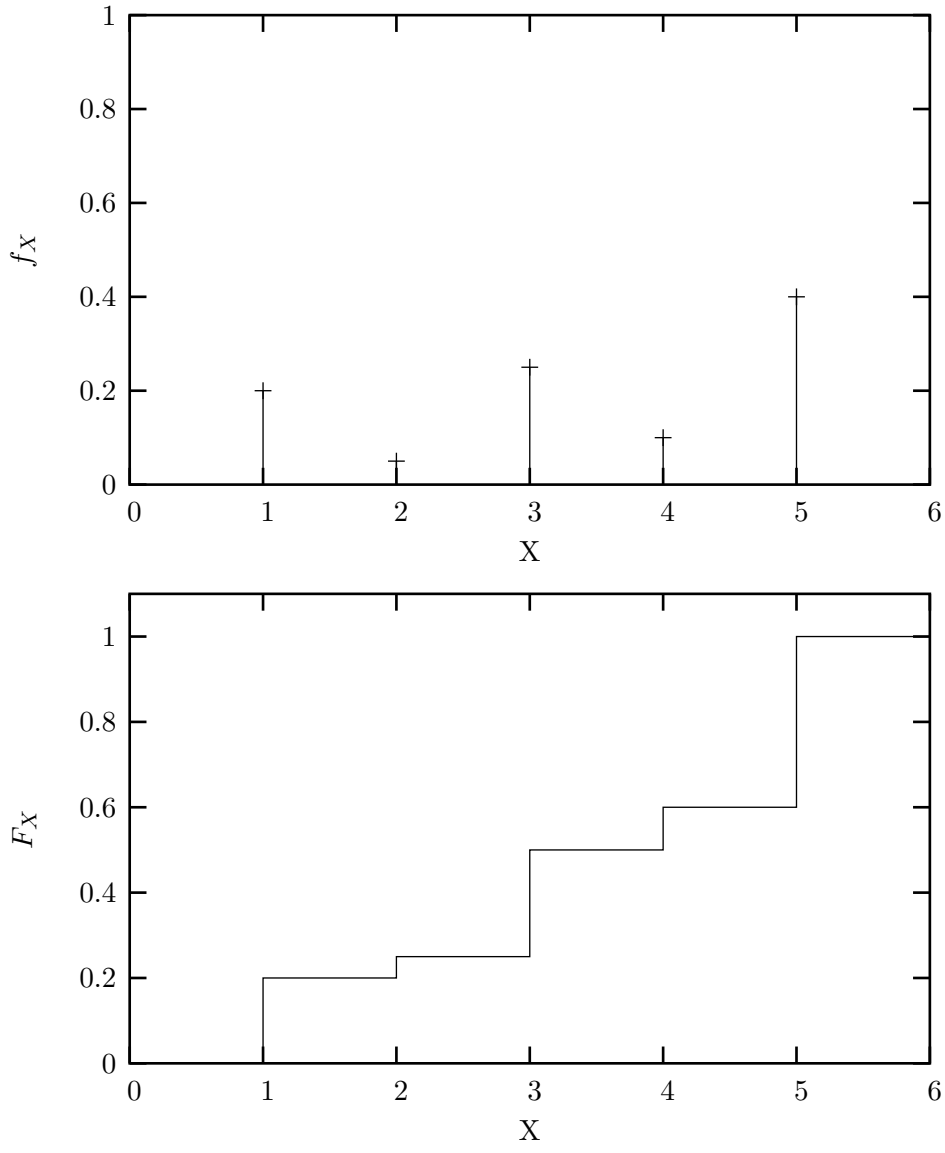


Figura 1: Distribuzione discreta

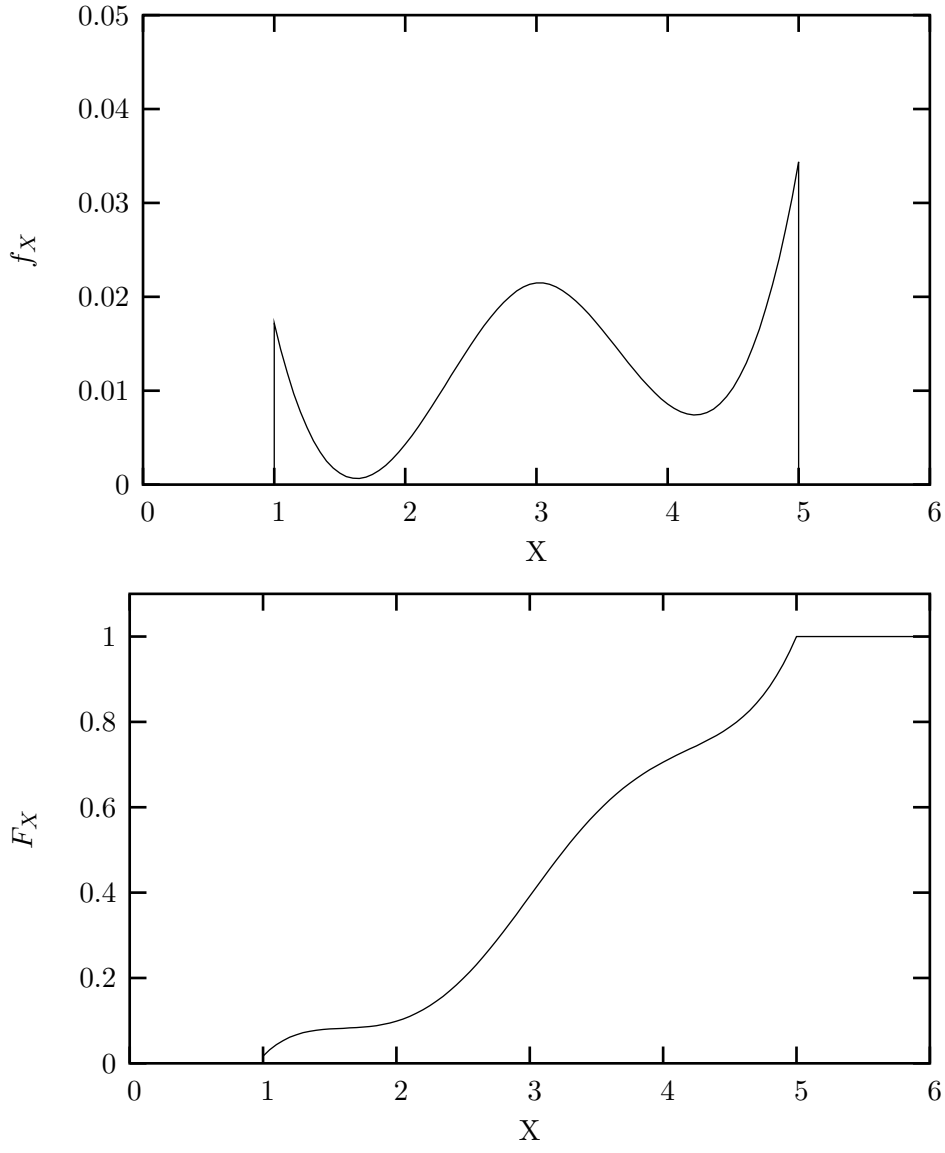


Figura 2: Distribuzione continua

2.9.1 Media

- $\boldsymbol{E}(X) = \mu_X = \bar{X} = \sum_j x_j f_X(x_j)$, se X è una variabile aleatoria discreta
- $\boldsymbol{E}(X) = \mu_X = \bar{X} = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x$, se X è una variabile aleatoria continua

Indica il baricentro della variabile aleatoria, con la funzione densità a fare il ruolo della massa. E' una misura della posizione della variabile aleatoria all'interno del dominio.

2.9.2 Varianza

- $\mathrm{var}(X) = \sigma_X^2 = \sum_j (x_j - \mu_X)^2 f_X(x_j)$ se X è una variabile aleatoria discreta
- $\mathrm{var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\mathrm{d}x$ se X è una variabile aleatoria continua

E' una misura dello scostamento della variabile aleatoria dalla media, una misura della sua dispersione lungo il dominio.

2.9.3 Deviazione standard (scarto quadratico medio)

E' definita come $\sqrt{\mathrm{var}(X)}$ e viene indicata come σ_X (o anche con $\mathrm{std}(X)$)

Anche la deviazione standard misura la dispersione della variabile aleatoria, ma utilizzando la stessa unità di misura della media. La varianza, invece, ha un'unità di misura al quadrato rispetto alla media.

Come si può notare nella figura 3, gli indicatori non sempre hanno una semplice interpretazione. La distribuzione nella sottofigura (a) è simmetrica e gli eventi con maggiore probabilità sono limitati in un intervallo: media e deviazione standard descrivono bene il l'evento e la regione più probabile. La distribuzione della figura (b), invece, non è simmetrica: media e deviazione standard non sono di molto aiuto per descrivere le caratteristiche di questa distribuzione. Nella sottofigura (c), infine, sono mostrate due distribuzioni che hanno gli stessi valori di media e varianza, ma andamenti molto differenti.

2.9.4 Momenti

Il momento di ordine r è il valore atteso della potenza r-esima della variabile aleatoria.

Il momento centrale di ordine r rispetto ad a è definito come $\boldsymbol{E}((X-a)^r)$. Se una funzione di densità di una data variabile aleatoria è simmetrica rispetto alla media, i momenti dispari rispetto alla media della variabile stessa sono 0: i momenti dispari rispetto alla media possono essere usati per misurare la simmetria della funzione di densità.

2.9.5 Quantile

Il quantile q-esimo di una variabile aleatoria X è il più piccolo numero ξ tale che:

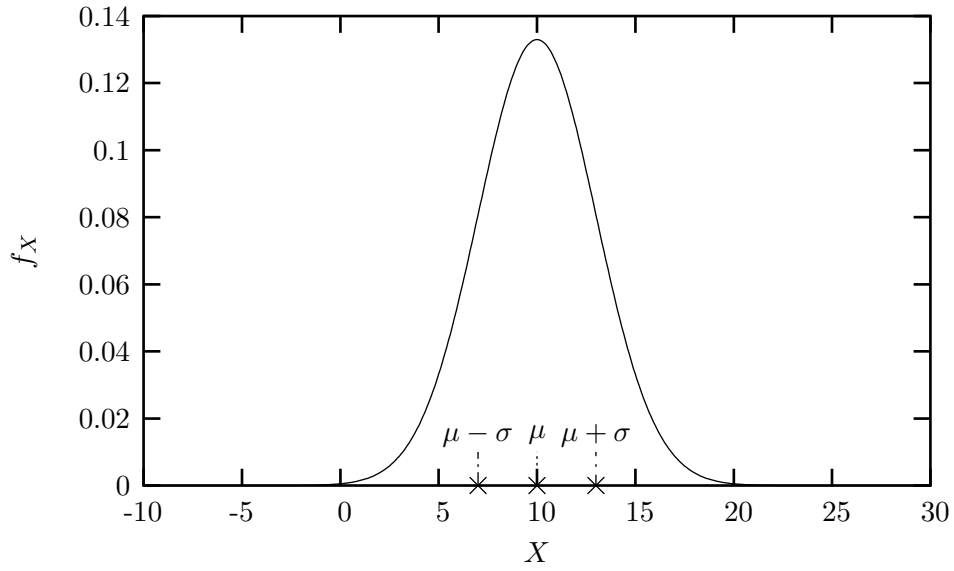
- $F_X(\xi) \geq q$ se X è discreta
- $F_X(\xi) = q$ se X è continua

2.9.6 Mediana

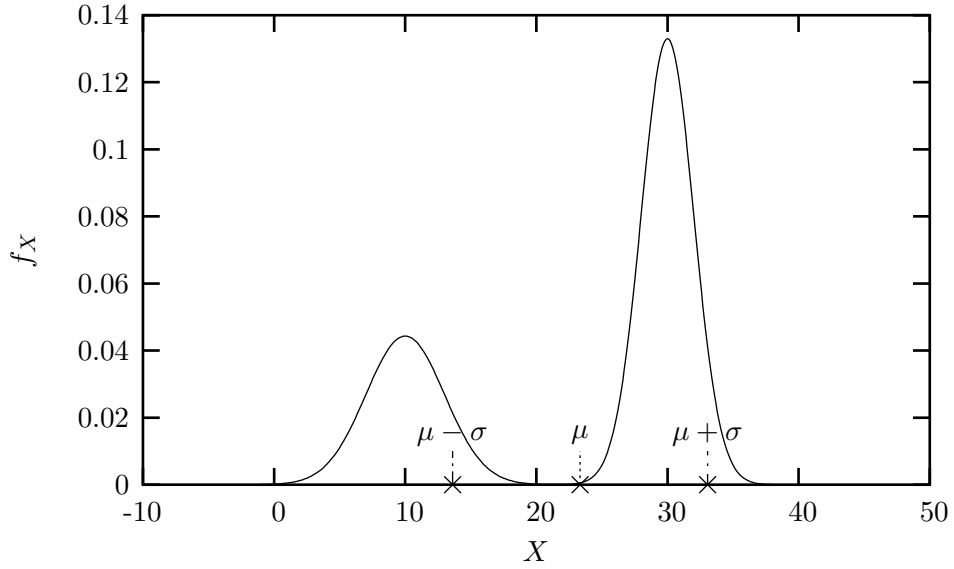
E' il quantile 0.5.

2.9.7 Moda

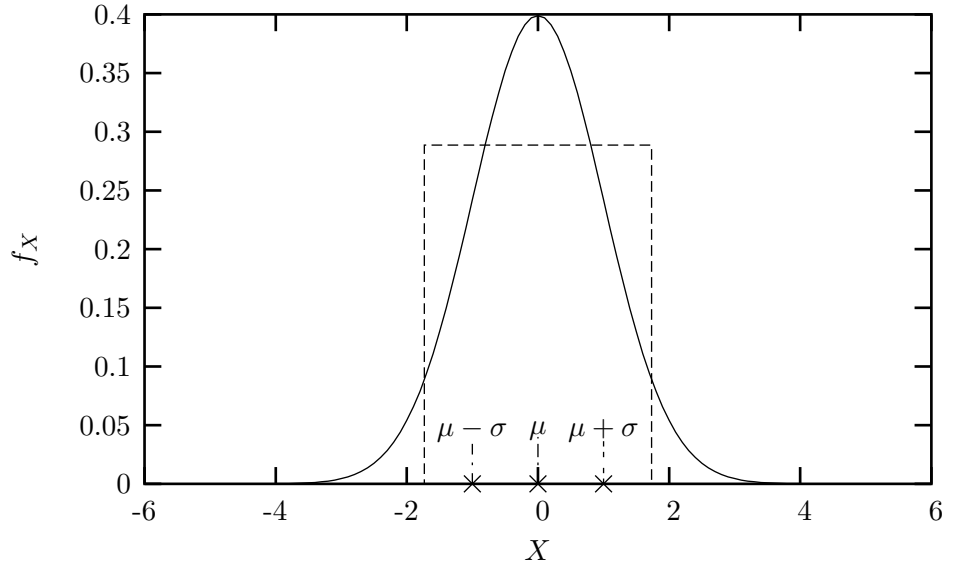
E' il punto in cui $f_X(\cdot)$ è massima.



(a)



(b)



(c)

Figura 3: Media e varianza

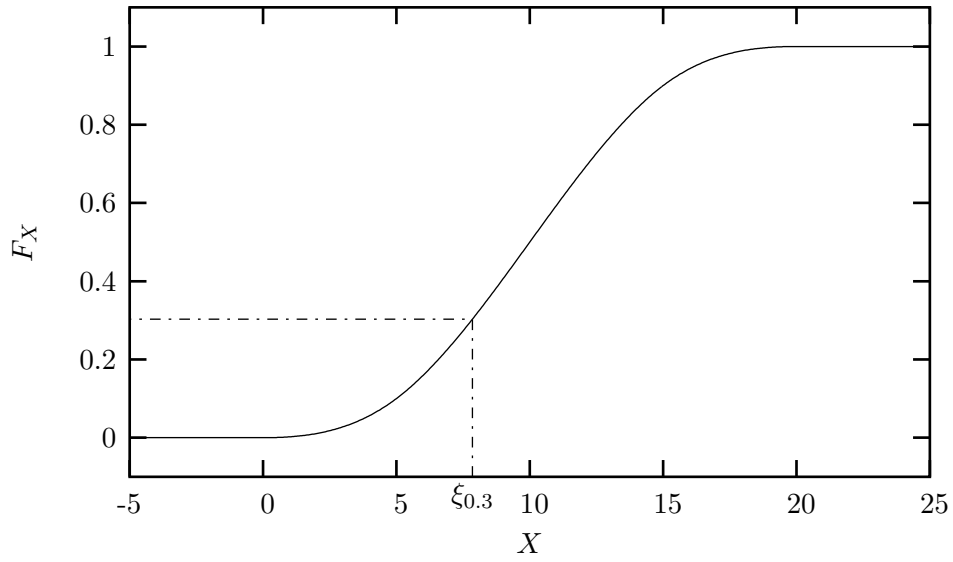


Figura 4: Quantile

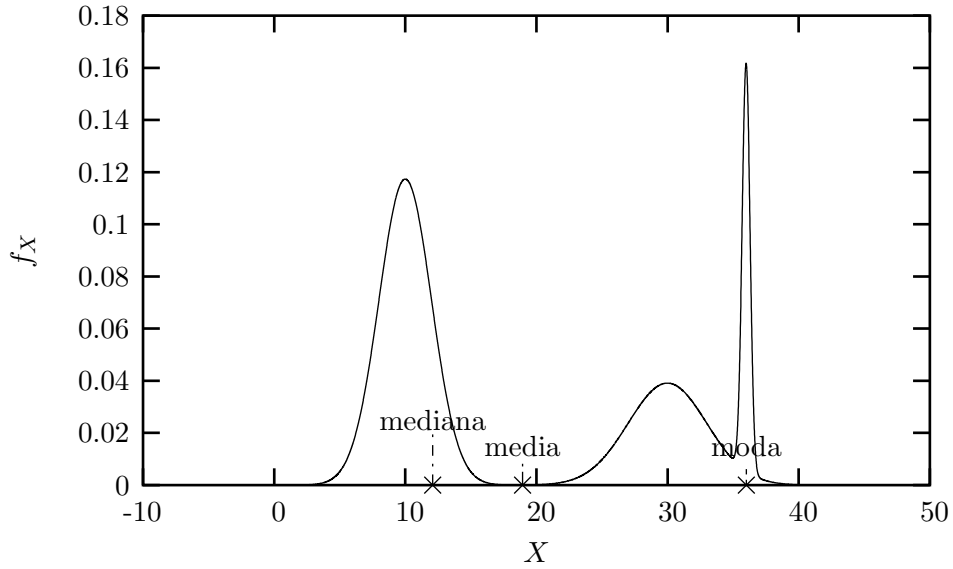


Figura 5: Media, mediana, moda

2.10 Tipi di funzione di distribuzione di probabilità

Fenomeni diversi sono descritte da funzioni di probabilità diverse.

2.10.1 Distribuzione uniforme discreta

Modella i fenomeni in cui gli eventi sono equiprobabili.

Per esempio, gli esiti del lancio di un dado o di una moneta (non truccati, ovviamente).

In generale, se gli eventi possibili sono N, ad essi si può associare la variabile aleatoria X con valori di probabilità non nulla in $\{1, 2, \ldots, N\}$. In tal caso, la distribuzione avrà le seguenti caratteristiche:

- $f_X(x) = \begin{cases} \frac{1}{N}, & x = 1, 2, \ldots, N \\ 0, & \text{altrimenti} \end{cases}$
- $F_X(x) = \begin{cases} 0, & x < 1 \\ \frac{\lfloor x \rfloor}{N}, & 1 \leq x \leq N \\ 1, & \text{altrimenti} \end{cases}$
- $\boldsymbol{E}(X) = \frac{(N+1)}{2}$
- $\text{var}(X) = \frac{(N^2-1)}{12}$

NB : Talvolta può essere più opportuno associare, tramite la variabile aleatoria X, agli N eventi anziché i valori tra 1 e N, altri valori, per esempio i valori tra 0 e $N-1$, . In tal caso, i valori di media e varianza sopra riportati non possono più essere utilizzati.

2.10.2 Distribuzione di Bernoulli

Modella fenomeni in cui un evento può accadere, oppure no: esperimenti che possono risolversi con un successo o un fallimento. I due esito possono anche avere probabilità diverse. Per esempio, immaginiamo di avere un tavolo da gioco quadrato con sponde e con un cerchio inscritto disegnato sul fondo. Se l'esperimento consiste nel lanciare una biglia ed osservare dove si ferma, l'esperimento ha due esiti (A = la biglia si ferma nel cerchio, B = la biglia si ferma all'esterno del cerchio), ed essi non sono equiprobabili ($P(A) = \frac{\pi}{4}$, $P(B) = 1 - \frac{\pi}{4}$).

Una distribuzione di Bernoulli ha le seguenti caratteristiche:

- $f_X(x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0o1 \\ 0, & \text{altrimenti} \end{cases}$
- $F_X(x) = \begin{cases} 0, & x < 0 \\ (1-p), & 0 = x < 1 \\ 1, & x \geq 1 \end{cases}$
- $\boldsymbol{E}(X) = p$
- $\text{var}(X) = p(1-p)$

La variabile aleatoria X viene usata per indicare il numero di successi (0 o 1), i quali possono avvenire con probabilità p. La probabilità di fallimento è $(1-p)$, spesso indicata con q.

2.10.3 Distribuzione binomiale

Modella fenomeni che corrispondono a n esperimenti di tipo bernoulliano. La variabile aleatoria descrive il numero di successi ottenuti.

Per esempio, con riferimento all'esperimento descritto per la distribuzione di Bernoulli, il numero di biglie che si fermano nel cerchio in 10 tentativi ha una distribuzione binomiale.

Una distribuzione binomiale ha le seguenti caratteristiche:

- $f_X(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0, 1, \ldots, n \\ 0, & \text{altrimenti} \end{cases}$
- $\boldsymbol{E}(X) = np$
- $\text{var}(X) = npq$

La distribuzione binomiale descrive gli esperimenti di estrazione con reimissione.

2.10.4 Distribuzione ipergeometrica

Modella gli esperimenti di estrazione senza reimissione.

Per esempio, vi sia un'urna contenente M oggetti, K dei quali difettosi. La distribuzione aleatoria ipergeometrica descrive la probabilità di estrarre x elementi difettosi estraendone n in totale, senza reimmissione.

Immaginiamo di dover cercare una chiave giusta per una serratura avendo a disposizione un mazzo di 23 chiavi, fra le quali sono presenti 5 copie della chiave cercata. Facendo le cose assennatamente, una volta provata, una chiave viene accantonata, e le prove successive vengono effettuate con le chiavi rimanenti.

Una distribuzione ipergeometrica ha le seguenti caratteristiche:

- $f_X(x) = \begin{cases} \frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}}, & x = 0, 1, \ldots, n \\ 0, & \text{altrimenti} \end{cases}$
- $\boldsymbol{E}(X) = n\frac{K}{M}$
- $\text{var}(X) = n \cdot \frac{K}{M} \cdot \frac{M-K}{M} \cdot \frac{M-n}{M-1}$

2.10.5 Distribuzione geometrica

La distribuzione geometrica (o di Pascal) descrive la probabilità di dover ripetere un esperimento bernoulliano un certo numero di volte prima di riuscire ad ottenere un successo. Tali prove di Bernoulli devono essere indipendenti. Per esempio, il numero di tentativi che un ubriaco deve fare per trovare la chiave di casa in un mazzo di chiavi. La variabile aleatoria descrive il numero di tentativi falliti prima di ottenere un successo. E' possibile vedere la distribuzione geometrica come un modello del tempo d'attesa di un evento.

Una distribuzione geometrica ha le seguenti caratteristiche:

- $f_X(x) = \begin{cases} p(1-p)^x, & x = 0, 1, \ldots \\ 0, & \text{altrimenti} \end{cases}$
- $\boldsymbol{E}(X) = \frac{q}{p}$
- $\text{var}(X) = \frac{q}{p^2}$

2.10.6 Distribuzione di Poisson

La distribuzione di Poisson può essere usata per modellare i conteggi di eventi con certe caratteristiche. Per esempio: numero di incidenti stradali in una settimana in una certa regione, numero di particelle radioattive emesse per unità di tempo, numero di organismi per unità di fluido, numero di imperfezioni per unità di lunghezza di un cavo.

I fenomeni devono avere le seguenti caratteristiche:

1. la probabilità che si verifichi esattamente un evento in un piccolo intervallo di tempo h è approssimativamente uguale a vh, per una costante v opportuna;
2. la probabilità che si verifichi più di un evento nell'intervallo di tempo di lunghezza h è trascurabile rispetto alla probabilità che se ne verifichi esattamente uno;
3. il numero di eventi in intervalli di tempo non sovrapposti sono indipendenti.

In termini matematici, le prime due condizioni si esprimono come segue:

1. $P(\text{un solo evento nell'intervallo lungo } h) = vh + o(h)$
2. $P(\text{due o più eventi nell'intervallo lungo } h) = o(h)$

Il termine $o(h)$ indica una funzione non specificata avente la proprietà: $\lim_{h\to 0} \frac{o(h)}{h} = 0$.

La quantità v può essere interpretata come il numero medio di eventi nell'unità di tempo.

Se i tre precedenti assunti sono soddisfatti, il numero di volte che un evento si verifica in un intervallo di tempo t è una variabile aleatoria poissoniana.

Per le caratteristiche sopra descritte, si dice che la distribuzione di Poisson modella fenomeni caratterizzati da eventi rari, nel senso che ad una adeguata scala temporale, non è possibile che due eventi si verifichino nello stesso istante di tempo. Inoltre, la poissoniana può essere usata per modellare una binomiale con n molto elevato.

Una distribuzione poissoniana ha le seguenti caratteristiche:

- $f_X(x;t) = \begin{cases} \frac{e^{-vt}(vt)^x}{x!}, & x = 0,\, 1,\, \ldots \\ 0, & \text{altrimenti} \end{cases}$
- $\boldsymbol{E}(X;t) = vt$
- $\text{var}(X;t) = vt$

dove t è la misura dell'intervallo considerato.

Esempio Un impianto di lavorazione del legno produce fogli di compensato che presentano, in media, tre imperfezioni ogni 50 m^2. Qual è la probabilità che un foglio 3 m $\times$ 4 m:

a) sia esente da imperfezioni?

b) presenti non più di una imperfezione?

Prendendo come unità di misura il metro quadro, il numero di imperfezioni per unità di misura sarà $v = \frac{3}{50} = 0.06$. Pertanto:

a) la probabilità che in $3\times 4 = 12$ m^2 vi siano zero impefezioni è $P(0;12) = \frac{e^{-0.06\cdot 12}\cdot(0.06\cdot 12)^0}{0!} = \frac{0.48675\cdot 1}{1} = 0.48675$

b) $P(0;12)+P(1;12) = 0.48675+\frac{e^{-0.06\cdot 12}(0.06\cdot 12)^1}{1!} = 0.48675+\frac{0.4867\cdot 0.72}{1} = 0.48675+0.35046 = 0.83721$

2.10.7 Distribuzione uniforme continua

La distribuzione uniforme continua descrive gli eventi che possono accadere con uguale probabilità in un intervallo continuo.

Per esempio, lanciamo una palla su un biliardo e, una volta fermatasi, misuriamone la distanza dal bordo del tavolo.

La distribuzione uniforme nell'intervallo $[ab] \subset \mathbb{R}$ avrà le seguenti caratteristiche:

- $f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{altrimenti} \end{cases}$
- $F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x \geq b \end{cases}$
- $\boldsymbol{E}(X) = \frac{(a+b)}{2}$
- $\text{var}(X) = \frac{(b-a)^2}{12}$

Data la forma del suo grafico, tale distribuzione viene anche detta *rettangolare*.

2.10.8 Distribuzione normale

La distribuzione normale viene anche detta *gaussiana* ed ha una notevole importanza in statistica.

Una distribuzione normale con media μ e deviazione standard σ ha le seguenti caratteristiche:

- $f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $\boldsymbol{E}(X) = \mu$
- $\text{var}(X) = \sigma^2$

Media, moda e mediana sono posti in μ, mentre in $\mu - \sigma$ e $\mu + \sigma$ la funzione di densità presenta dei flessi.

Per la sua importanza, è usata una notazione dedicata alla gaussiana. Una variabile aleatoria normale X, di media μ e varianza σ^2 viene indicata $X \sim N(\mu, \sigma^2)$. La funzione di densità di $X \sim N(\mu, \sigma^2)$ viene indicata con $\phi_{\mu,\sigma^2}(\cdot)$ e la funzione distribuzione cumulativa con $\Phi_{\mu,\sigma^2}(\cdot)$.

Di particolare importanza è anche la *normale standardizzata*, cioè la distribuzione gaussiana con media 0 e varianza 1 ($X \sim N(0, 1)$), per la quale vengono tralasciati gli indici:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{e} \quad \Phi(x) = \int_{-\infty}^{x} \phi(u) \mathrm{d}u$$

L'importanza della normale standardizzata è evidente dalla seguente proprietà:

$$\text{se } X \sim N(\mu, \sigma^2), \;\; P(a < X < b) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

Ciò consente di calcolare le probabilità riferite ad una qualsiasi gaussiana disponendo dei valori della gaussiana standardizzata. Tali valori sono generalmente tabulati.

Essendo, come si vedrà in seguito, la distibuzione limite della somma di un gran numero di distribuzioni, la gaussiana è utile per modellare le probabilità di fenomeni che sono la risultante di un gran numero di cause.

Per esempio, i fori sul bersaglio del gioco delle freccette, i voti di un esame (in condizioni normali), o le dimensioni dei pezzi di lavorazione si distribuiscono secondo una gaussiana.

2.10.9 Distribuzione esponenziale

La distribuzione esponenziale modella il tempo che intercorre tra gli eventi di una variabile aleatoria poissoniana.

Una distribuzione esponenziale relativa ad un fenomeno con numero medio di eventi nell'unità di tempo pari a v ha le seguenti caratteristiche:

- $f_X(x) = \begin{cases} v\, e^{-vx}, & x \geq 0 \\ 0, & \text{altrimenti} \end{cases}$
- $\boldsymbol{E}(X) = \frac{1}{v}$
- $\text{var}(X) = \frac{1}{(v)^2}$

2.10.10 Distribuzione gamma

Modella l'intervallo di tempo che bisogna attendere per la r-esima manifestazione di un evento poissoniano:

- $f_X(x; r) = \begin{cases} \frac{v}{\Gamma(r)}\, (vx)^{r-1}\, e^{vx}, & x \geq 0 \\ 0, & \text{altrimenti} \end{cases}$
- $\boldsymbol{E}(X; r) = \frac{r}{v}$
- $\text{var}(X; r) = \frac{r}{(v)^2}$

La funzione $\Gamma(\cdot)$ è così definita:

$$\Gamma(t) = \int_0^\infty x^{t-1} e^x \mathrm{d}x, \;\; t > 0$$

Se t è intero, $\Gamma(t+1) = t!$.

2.11 Distribuzioni congiunte

Una funzione di distribuzione cumulativa congiunta di k variabili aleatorie ha come dominio lo spazio euclideo k-dimensionale e come codominio l'intervallo [0, 1].

Nel caso monodimensionale, la variabile aleatoria serviva per associare ad ogni evento un numero. Nel caso monodimensionale ci sono più variabili aleatorie in gioco contemporaneamente. Esse servono per descrivere i diversi aspetti del fenomeno che si sta analizzando. Per esempio, consideriamo il caso in cui una variabile aleatoria, X, corrisponda ad un giorno dell'anno, mentre un'altra, Y, corrisponda alla temperatura durante il giorno. E' evidente la probabilità di avere almeno 20 C il 30 aprile è maggiore dalla probabilità di avere la stessa temperatura il 31 dicembre. Ciò può essere formalizzato come: $P[X = 30 \text{ aprile}, Y \geq textrm20C] \geq P[X = 31dicembre, Y \geq 20C]$.

Proprietà Le proprietà delle distribuzioni congiunte sono una generalizzazione delle proprietà delle distribuzioni monodimensionali. Vediamole per il caso in cui $k = 2$:

- $\lim_{x \to -\infty} F(x, y) = 0$
  $x \to -\infty$ descrive l'evento impossibile: la probabilità che avvenga un evento impossibile per X e, contemporaneamente, un qualsiasi evento per Y deve essere 0 per qualunque evento descritto da Y.

- $\lim_{y \to -\infty} F(x, y) = 0$
  Questa proprietà è analoga alla proprietà precedente, ma riferita ad un evento impossibile descritto da Y.

- $\lim_{x \to \infty, y \to \infty} F(x, y) = 1$
  Analogamente ai casi precedenti, $x \to \infty$ e $y \to \infty$ descrivono eventi certi sia per X che per Y. La probabilità di un evento certo deve essere 1.

- se $x_1 < x_2$ e $y_1 < y_2$, allora
  $P(x_1 \leq X \leq x_2,\, y_1 \leq Y \leq y_2) = F(x_2,\, y_2) - F(x_2,\, y_1) - F(x_1,\, y_2) + F(x_1,\, y_1)$

- $F(x,\, y)$ è continua da destra in ciascuna variabile:
  $\lim_{h \to 0^+} F(x + h,\, y) = \lim_{h \to 0^+} F(x,\, y + h) = F(x,\, y)$

2.11.1 Distribuzioni marginali

Sono le distribuzioni delle singole variabili aleatorie: $F_X(x) = F_{X,\,Y}(x, \infty)$ e $F_Y(y) = F_{X,\,Y}(\infty,\, y)$.

La conoscenza della distribuzione congiunta implica la conoscenza delle distribuzioni marginali. Non è vero, in generale, il contrario: dalle distribuzioni marginali non possiamo ricavare la distribuzione congiunta.

2.11.2 Densità congiunta

Estensione del concetto di funzione di densità di probabilità al caso multidimensionale.

Variabili aleatorie discrete La funzione densità di probabilità congiunta di variabili aleatorie discrete è definita come funzione che descrive la probabilità degli eventi congiunti

$$f_{X_1,\,\ldots,\,X_k}(x_1,\, \ldots,\, x_k) = P(X_1 = x_1,\, \ldots,\, X_k = x_k)$$

$f_{X_1,\,\ldots,\,X_k}(x_1,\, \ldots,\, x_k)$ è quindi la probabilità che contemporaneamente si verifichino gli eventi tali per cui le variabili aleatorie $X_1,\, \ldots,\, X_k$ assumano rispettivamente i valori $x_1,\, \ldots,\, x_k$.

$$\sum f_{X_1,\,\ldots,\,X_k}(x_1,\, \ldots,\, x_k) = 1$$

Come sempre, la somma delle probabilità di tutti gli eventi possibili deve essere 1.

Variabili aleatorie continue Per ogni variabile aleatoria k-dimensionale $(X_1,\, \ldots,\, X_k)$ la funzione densità di probabilità $f_{X_1,\,\ldots,\,X_k}(\cdot,\, \ldots,\, \cdot)$ è la funzione tale che:

$$F_{X_1,\,\ldots,\,X_k}(x_1,\, \ldots,\, x_k) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f_{X_1,\,\ldots,\,X_k}(u_1,\, \ldots,\, u_k)\, \mathrm{d}u_1 \cdots \mathrm{d}u_k$$

Come per il caso monodimensionale, valgono le seguenti proprietà:

- $f_{X_1,\,\ldots,\,X_k}(x_1,\, \ldots,\, x_k) \geq 0$, per ogni valore della variabile aleatoria k-dimensionale $(X_1,\, \ldots,\, X_k)$
- $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\,\ldots,\,X_k}(u_1,\, \ldots,\, u_k)\, \mathrm{d}u_1 \cdots \mathrm{d}u_k = 1$

2.11.3 Densità marginali

Sono le funzioni densità delle singole variabili aleatorie.

La conoscenza della distribuzione congiunta implica la conoscenza delle distribuzioni marginali. Infatti valgono le seguenti proprietà:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y \qquad \text{e} \qquad f_Y(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x$$

Non è vero, in generale, il contrario: dalle distribuzioni marginali non possiamo ricavare la distribuzione congiunta.

2.11.4 Indipendenza

Sia $(X_1, \ldots, X_k)$ una variabile k-dimensionale. $X_1, \ldots, X_k$, sono stocasticamente indipendenti se e solo se:

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \prod_{i=1}^{K} F_{X_i}(x_i)$$

per tutti i valori $x_1, \ldots, x_k$.

2.11.5 Media

Il valore atteso di una variabile aleatoria multidimensionale è dato dalla media dei suoi valori, pesata con la probabilità di ogni singolo evento. Per una variabile di tipo discreto:

$$\boldsymbol{E}(X_1, \ldots, X_k) = \sum_{X_1,\ldots,X_k} (x_1, \ldots, x_k) f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$$

Per una variabile continua:

$$\boldsymbol{E}(X_1, \ldots, X_k) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} (x_1, \ldots, x_k) f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)\,\mathrm{d}x_1 \ldots \mathrm{d}x_k$$

E' possibile dimostrare che la media di ogni componente, X_j, della variabile k-dimensionale $(X_1, \ldots, X_k)$ assume la forma già conosciuta:

$$\mu_j = \boldsymbol{E}(X_j) = \int_{-\infty}^{+\infty} x_j f_{x_j}(x_j)\,\mathrm{d}x_j$$

2.12 Correlazione

La varianza di una variabile monodimensionale misura la dispersione degli eventi intorno alla loro media. Questo concetto può essere generalizzato per il caso k-dimensionale. In questo caso, però, può essere utile sapere anche come la dispersioni degli eventi avviene rispetto alle singole componenti. Nel seguito, per semplicità, considereremo solo il caso bidimensionale.

2.12.1 Covarianza

$$\mathrm{cov}(X,Y) = \boldsymbol{E}((X - \mu_X)(Y - \mu_Y))$$

Variabili non correlate hanno $\mathrm{cov}(X,Y) = 0$ (ma non è vero il contrario).

2.12.2 Coefficiente di correlazione

$$\rho_{X,Y} = \frac{\text{cov}[X,Y]}{\sigma_X \sigma_Y}$$

Covarianza e coefficiente di correlazione descrivono una relazione lineare tra X e Y:

- positiva: $X - \mu_X$ e $Y - \mu_Y$ tendono (con alta probabilità) ad avere lo stesso segno, cioè è altamente probabile che si verifichino degli eventi per cui le variabili aleatorie sono entrambe o minori o maggiori delle loro medie;
- negativo: $X - \mu_X$ e $Y - \mu_Y$ tendono (con alta probabilità) ad avere segni negativi, cioè, al contrario del caso precedente, è altamente probabile che si verifichino eventi per cui se una variabile aleatoria assume un valore minore della sua media, l'altra assume un valore maggiore della sua media, e viceversa.

Il valore della covarianza, non è significativa, perché deve essere confrontata con la varianza delle singole variabili: un valore di covarianza pari a 100 indica una forte relazione tra X e Y la loro varianza è 1, ma, al contrario, indica l'assenza di relazione se la loro varianza è pari a 10000. Il valore del coefficiente di correlazione, invece, è normalizzato rispetto alle varianze delle singole variabili e, pertanto, soddisfa la relazione: $-1 \leq \rho_{X,Y} \leq 1$ Il $\rho_{X,Y}$ rimuove la variabilità di X e Y: un valore di 1 (-1) indica una correlazione diretta (inversa), mentre 0 indica generalmente assenza di correlazione.

2.12.3 Curva di regressione

Curva di regressione di Y su x è definita come:

$$\boldsymbol{E}(Y|X = x)$$

Essa indica, per ogni x, il valore che possiamo aspettarci per Y quando X assume il valore x. Dovendo dare una descrizione sintetica (non stocastica) della relazione che lega due variabili aleatorie, questa è la migliore. Torneremo su questo argomento nel seguito.

3 Richiami di statistica

3.1 Statistica

Definizione di statistica come scienza (da http://www.garzantilinguistica.it):

> **Statistica**: analisi quantitativa dei fenomeni collettivi che hanno attitudine a variare, allo scopo di descriverli e di individuare le leggi o i modelli che permettono di spiegarli e di prevederli

3.2 A cosa serve la statistica?

Nel mondo reale abbiamo un modello dei fenomeni e la conoscenza limitata di alcuni eventi. La statistica dice quanto possiamo estendere la conoscenza che abbiamo ai fenomeni che l'hanno generata.

La totalità degli elementi in esame, dei quali si vogliono ottenere informazioni, viene chiamato *popolazione oggetto*. Un *campione casuale* di dimensione n è un'osservazione di n elementi aventi la stessa distribuzione. Tale elementi devono quindi essere stocasticamente indipendenti gli uni dagli altri. Spesso non è possibile accedere all'intera popolazione oggetto, ma si può accedere ad una popolazione ad essa attinente. Tale popolazione viene chiamata *popolazione campionata*. L'estensione alla popolazione oggetto delle informazioni inferite sulla popolazione campionata non sempre è possibile ed è comunque un'operazione critica. Tuttavia, i meccanismi

alla base dei fenomeni fisici (chimici, biologici) sono piuttosto stabili e riproducibili. Lo stesso non si può dire per i fenomeni sociologici. Conclusioni valide per una determinata popolazione campione non sempre sono estendibili a una popolazione più ampia, ma diversamente dislocata geograficamente, temporalmente o socialmente.

L'inferenza statistica si può suddividere nei seguenti argomenti:

- stima di parametri: siamo interessati a qualche valore numerico caratteristico di una distribuzione (e.g., la media);
- regressione: siamo interessati alla relazione che lega due grandezze (e.g, al variare della temperatura, come varia la velocità di fermentazione?);
- classificazione: siamo interessati a suddividere le osservazioni in classi, con omogeneità tra gli elementi della stessa classe, ma con forte differenziazione tra gli elementi di classi diverse;
- test d'ipotesi: siamo interessati a valutare la verosimiglianza di alcune ipotesi.

All'interno della scienza statistica, una *statistica* è una funzione di variabili casuali osservabili che non contiene alcun parametro incognito. Essendo una funzione di variabili casuali, la statistica è a sua volta una variabile casuale: ha, cioè, una sua distribuzione.

Per esempio, se $X_1, \ldots, X_n$ è un campione casuale, la quantità $\frac{1}{2}(\min_i X_i + \max_i X_i)$ è una statistica. Per convincersi che tale grandezza è una variabile aleatoria, basta immaginarsi di estrarre un campione e calcolarne il valore e ripetere poi tale operazione con un altro campione. Per esempio, la popolazione campione potrebbero essere gli esiti del lancio di tre dadi. E' chiaro che la media tra il più alto e il più basso dei dadi può cambiare ad ogni lancio.

3.3 Media campionaria

Dato un campione casuale $X_1, \ldots, X_n$, la *media campionaria* è la statistica definita come:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Questa statistica deve il suo nome al fatto che il suo valore atteso è proprio la media della distribuzione dalla quale sono stati estratti i campioni:

$$\boldsymbol{E}(\bar{X}) = \mu$$

La varianza della media campionaria è legata alla varianza σ^2 della distribuzione campione dalla seguente relazione:

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

Maggiore è il numero di elementi del campione, tanto più precisa è la stima della media.

3.4 Varianza campionaria

Dato un campione casuale $X_1, \ldots, X_n$, la *varianza campionaria* è la statistica definita come:

$$\boldsymbol{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Come la media campionaria, anche la varianza campionaria deve il suo nome al fatto che il suo valore atteso è la varianza della distribuzione dalla quale è stato estratto il campione:

$$\boldsymbol{E}(\boldsymbol{S}^2) = \sigma^2, \quad n > 1$$

NB : La varianza campionaria utilizza $n-1$ e non n. Ciò è dovuto al fatto che la stima di $\bar{X}$ fa perdere un grado di libertà.

3.5 Legge dei grandi numeri

Legge dei grandi numeri debole Dice che scelti comunque due numeri ϵ e δ ($\epsilon > 0$ e $0 < \delta < 1$), esiste un intero n tale per cui, la probabilità che la media campionaria calcolata su un campione di dimensione maggiore o uguale a n non si discosti di meno di ϵ dalla media della popolazione è minore di δ, e che tale n deve essere maggiore di $\frac{\sigma^2}{\epsilon^2\delta}$:

$$\forall \epsilon, \delta : \epsilon > 0,\ 0 < \delta < 1,\ \ n \geq \frac{\sigma^2}{\epsilon^2 \delta} \Rightarrow P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \delta$$

dove $\bar{X}_n$ è la media campionaria calcolata su un campione di dimensione n, μ e σ^2 sono rispettivamente la media e la varianza della distribuzione da cui sono stati estratti i campioni.

Questo teorema esprime una convergenza in probabilità.

Esempio Una termostato è in grado di mantenere la temperatura di un contenitore con una varianza di 1 K. Quante letture si devono fare per essere sicuri almeno al 95% che la media campionaria non si discosti di più di 0.5 K dalla temperatura media del contenitore?

Si ha, quindi: $\sigma^2 = 1$, $\epsilon = 0.5$ e $\delta = 1 - 0.95 = 0.05$; perciò:

$$n > \frac{\sigma^2}{\delta \epsilon^2} = \frac{1}{0.05 \cdot 0.5^2} = 80$$

Legge dei grandi numeri forte Esprime la convergenza della media campionaria alla media della distribuzione campione:

$$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$$

3.6 Teorema centrale della statistica

Questo teorema descrive il comportamento della media campionaria all'aumentare del numero di campioni. In particolare, esso dice che al crescere del numero di campioni, n, la media campionaria di una variabile aleatoria con media μ e varianza σ^2 tende asintoticamente ad una normale con media μ e varianza σ^2/n:

$$\lim_{n\to\infty} \bar{X}_n = N\left(\mu, \frac{\sigma^2}{n}\right)$$

NB : Il teorema non fa menzione della particolare distribuzione da cui si campiona: basta che sia nota la media e la varianza. La figura 6 esemplifica questo concetto tramite un esperimento. Da una distribuzione uniforme continua nell'intervallo [1, 7] si estraggono N campioni di dimensione 10. Ogni campione di dimensione 10 viene utilizzato per calcolare la media campionaria, ottenendo così un campionamento di dimensione N della media campionaria della distribuzione originale (uniforme [1 7]). Gli N valori della media campionaria vengono utilizzati per tracciare gli istogrammi (opportunamente riscalati — l'area dell'istogramma deve essere 1!) nei grafici (a)–(c). I valori usati per N nei grafici (a)–(c) sono, rispettivamente, 100, 1000 e 10000. In ogni grafico viene riportata la funzione di densità di probabilità di una normale con media 4 e varianza 0.3 (la distribuzione dalla quale e' stato effettuato il campionamento ha media 4 e varianza 3).

Una obiezione! L'esempio illustrato con la figura 6 si presta ad una obiezione basata su due osservazioni:

- la distribuzione normale ammette tutti i valori reali come valori possibili (sebbene valori lontani dalla media siano altamente improbabili);
- la media campionaria dell'esempio non può assumere valori esterni all'intervallo [1, 7].

Come è quindi possibile che la distribuzione della media campionaria sia una normale? In altri termini, la probabilità che la media campionaria dell'esempio assuma un valore maggiore di 9 è zero (è impossibile che la media di valori nell'intervallo [1, 7] sia 9 o più), mentre $P(N(4, 0.3) \leq 9) = 0.0019462$. Il teorema centrale della statistica fallisce?

La risposta è che il teorema descrive una convergenza asintotica, non un'identità. Ripetendo l'esperimento utilizzando 100 campioni, anziché 10, la distribuzione della media campionaria assume la forma di una campana molto affusolata ($N(4, 0.003)$) e la probabilità che la media valga 9 o più, cioè $P(N(4, 0.03) \leq 9)$ diviene così piccola che anche numericamente è difficile da calcolare. La figura 7 illustra lo stesso esperimento della figura 6, ma utilizzando campioni di dimensione 100 anziché 10.

Al tendere del numero di campioni verso l'infinito, la distribuzione tende ad una funzione che ha valore non nullo solo per la media.

4 Regressione lineare

La regressione vista al paragrafo 2.12.3 descriveva la relazione che lega due variabili aleatorie.

In ambito statistico, questo concetto viene ripreso cercando di affrontare il seguente problema: dato un insieme di n osservazioni di due variabili aleatorie, $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ cosa possiamo dire del legame che intercorre tra le due variabili? Le due grandezze osservate hanno una qualce interazione?

La forma più naturale per cercare di descrivere la dipendenza che lega le due variabili è la regressione: $\boldsymbol{E}(Y|X = x)$.

Esistono almeno due modi di utilizzare la conoscenza di una tale relazione:

- sfruttare la conoscenza di X per fare delle ipotesi su Y; per esempio per rispondere alla domanda "dato che la pressione ha il valore di 30 Pa, quanto potrà valere la temperatura?"
- potendo agire sulla grandezza X, pilotare tramite essa il valore di Y; per esempio, "quanta acqua aggiungere per mantenere costante la crescita della tal pianta?".

Quando si ha una conoscenza completa della relazione che lega due variabili, la risposta è relativamente semplice. Il problema che stiamo descrivento ha però le seguenti caratteristiche:

- si ha la conoscenza delle variabili sono in un numero finito di punti
- tale conoscenza non è precisa: le misure sono sempre affette da errore.

In caso di assenza di informazioni, si possono solo fare delle ipotesi (e delle assunzioni) realistiche. Tali ipotesi sono volte a semplificare il problema, in modo da renderlo matematicamente trattabile, ma senza pregiudicare l'utilità della soluzione trovata.

Il modello di *regressione lineare semplice* è basato sulle seguenti ipotesi:

- X e Y sono due variabili univariate;
- $Y_i = \beta_0 + \beta_1 x + e_i$, per tutti i campioni osservati (X_i, Y_i): X e Y hanno un legame lineare, a meno di un disturbo aleatorio;
- $\boldsymbol{E}(e_i) = 0$, per tutti i disturbi e_i: i disturbi aleatori hanno media nulla;

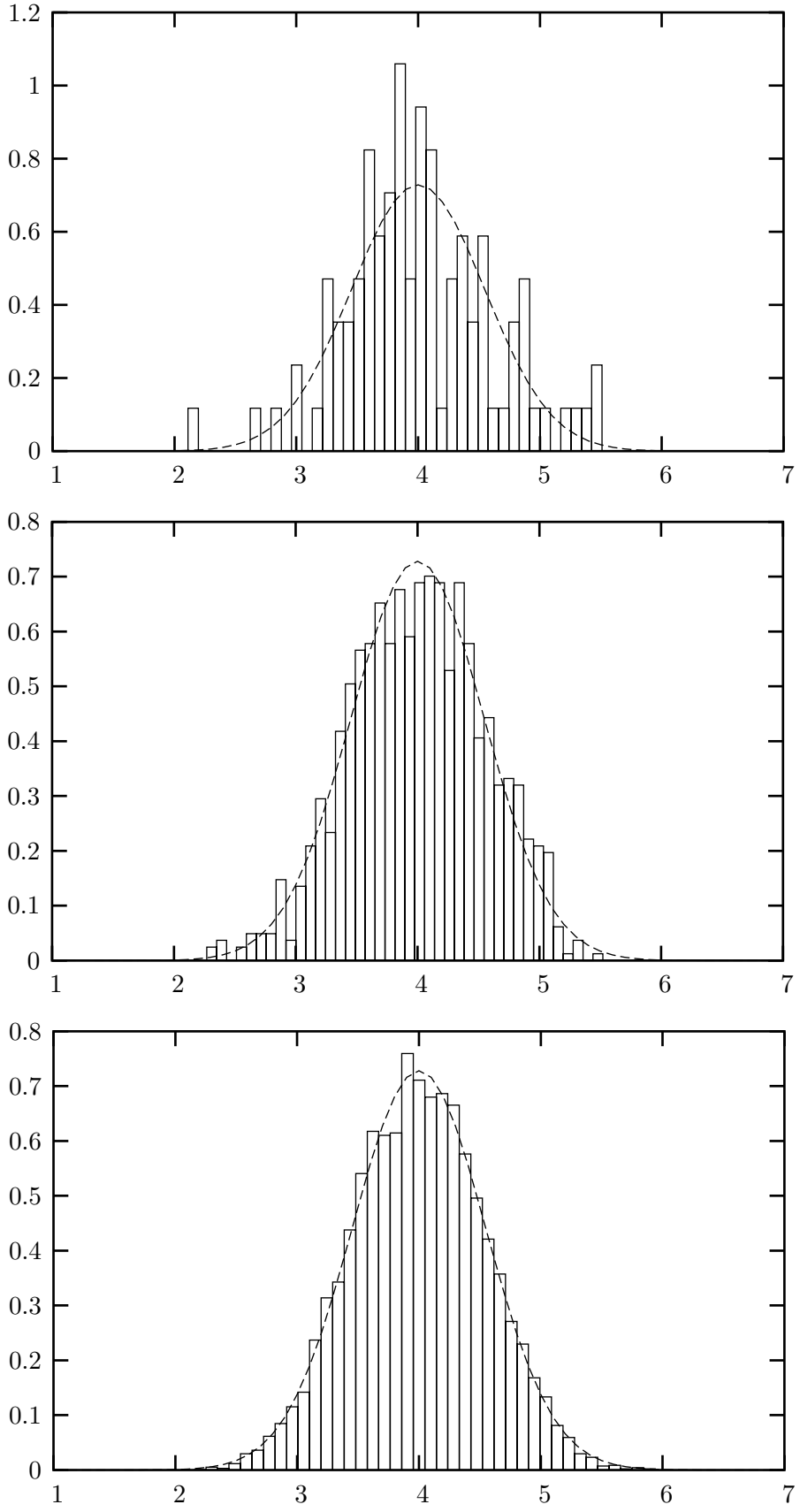


Figura 6: Verifica empirica del teorema centrale della statistica.

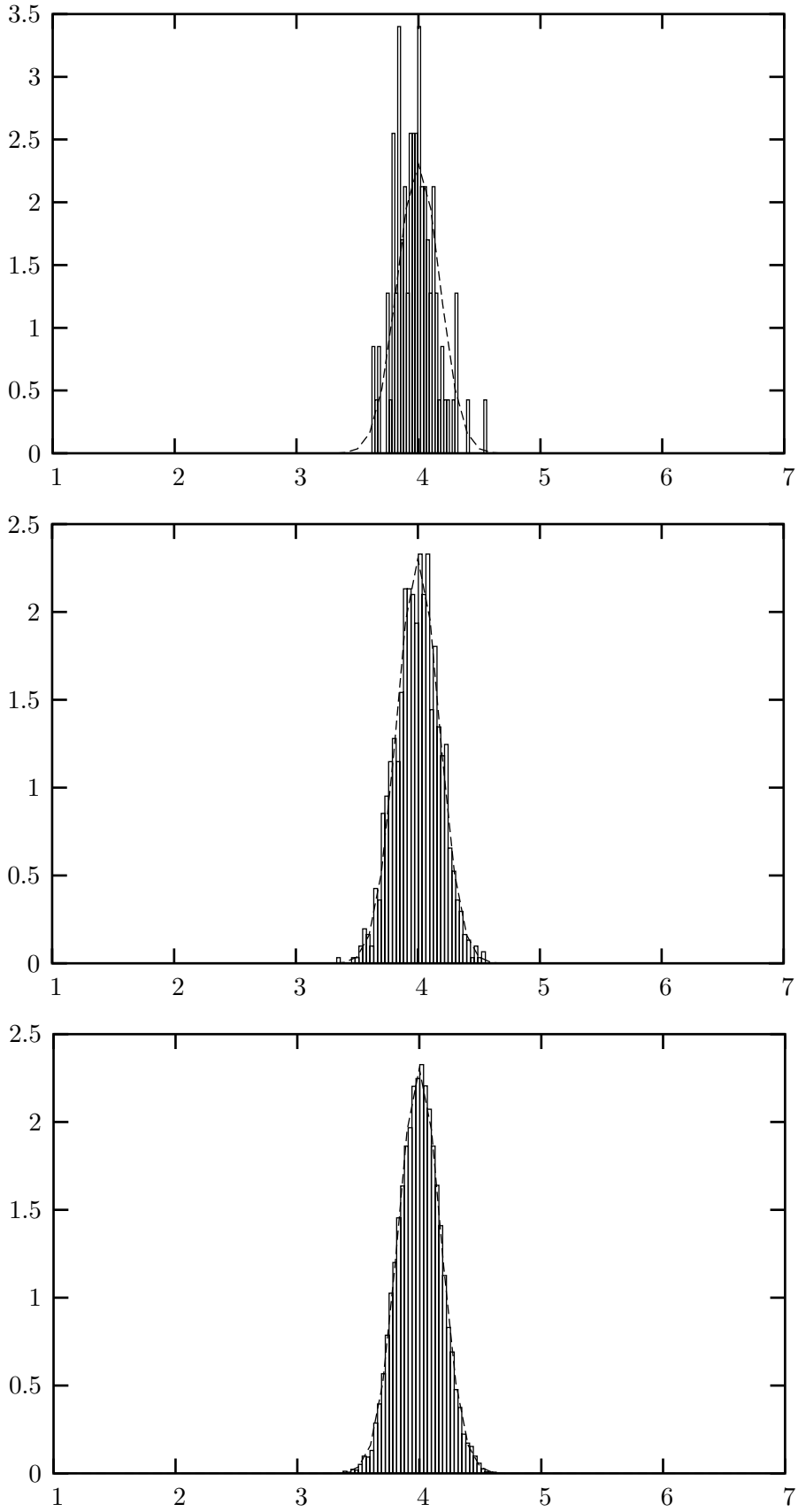


Figura 7: Verifica empirica del teorema centrale della statistica: al crescere del numero di campioni, la distribuzione della media campionaria si addensa sempre più intorno alla media della distribuzione campionata.

- $\text{var}(e_i) = \sigma^2 < \infty$: i disturbi aleatori hanno tutti la stessa varianza (ed essa è finita);
- $\text{cov}(e_i, e_j) = 0$: i disturbi casuali non sono tra loro correlati;
- il valore assunto da X in ogni osservazione è noto senza alcun errore;
- i valori e_i e X_i sono tra loro indipendenti.

Per semplificare il problema, si suppone che il legame tra le due grandezze X ed Y sia di tipo deterministico, a meno di un termine aleatorio, e, che racchiude tutte le oscillazioni casuali dei valori di X e Y. Inoltre, si suppone che il legame tra X e Y sia lineare, e quindi la funzione di regressione sia una retta: $\boldsymbol{E}(Y|X = x) = \beta_0 + \beta_1 x$. Per effetto del disturbo aleatorio, Y_i sono anch'esse variabili aleatorie, tali che $\text{var}(Y_i) = \sigma^2$. Infine, condizionatamente a X_i e X_j rispettivamente, le variabili aleatorie Y_i e Y_j sono indipendenti: se conosciamo X_i e X_j la conoscenza di Y_i non ci dice nulla su Y_j e viceversa.

Un esempio di una situazione ben descritta dalle seguenti ipotesi è riportata in figura 8. Nella figura 8a sono riportati delle misure delle variabili X e Y. Tutta l'informazione che abbiamo sul fenomeno descritto da queste variabili risiede in questo insieme di punti. Il loro andamento è abbastanza lineare (ad occhio, si vede che i punti si ammassano lungo una retta). Il problema è ora individuare un criterio per decidere quale è la retta che meglio approssima i nostri dati. In figura 8b sono riportate due curve. E' evidente che la retta con il tratteggio lungo approssima meglio i dati della retta con tratteggio corto, ma calcolarla a partire dai dati campionati?

4.1 Il metodo dei minimi quadrati

Il metodo dei minimi quadrati (*least squares*) fornisce un criterio per trovare la retta migliore (nel senso dei minimi quadrati, appunto). Tale metodo è basato sulla ricerca dei valori dei parametri $\hat{\beta}_0$ e $\hat{\beta}_1$ tali da minimizzare la seguente funzione:

$$S(\beta_0, \beta_1) = \sum_i (Y_i - \beta_0 - \beta_1 X_i).^2$$

In pratica, si definisce ottima quella retta che permette di minimizzare la distanza media tra la retta ed i punti dati. Oltre ad utilizzare una formalizzazione ragionevole di retta ottima, questa formulazione ha anche il vantaggio di avere la soluzione nella forma:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_i) Y_i}{\sum_{i=1}^n (X_i - \bar{X}_i)^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

$$\hat{\beta}_0 = \sum_{i=1}^n \frac{Y_i}{n} - \hat{\beta}_1 \sum_{i=1}^n \frac{X_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{X}$$

dove $\text{cov}(X, Y)$ e $\text{var}(X)$ indicano rispettivamente la covarianza e la varianza campionaria, e $\bar{X}$ e $\bar{Y}$ le medie campionarie dei dati.

E' possibile dimostrare che il valore atteso dello scostamento dei dati dalla retta è nullo.

5 Modelli di ordine superiore

Non sempre la relazione tra due grandezze è di tipo lineare. Per trattare tali casi, le soluzioni sono due:

- rendere lineare il problema: per esempio, i dati in figura 9a non hanno una relazione lineare; tuttavia, considerando non la variabile Y, ma la variabile $Z = \log_2(X)$, la relazione con X torna lineare;

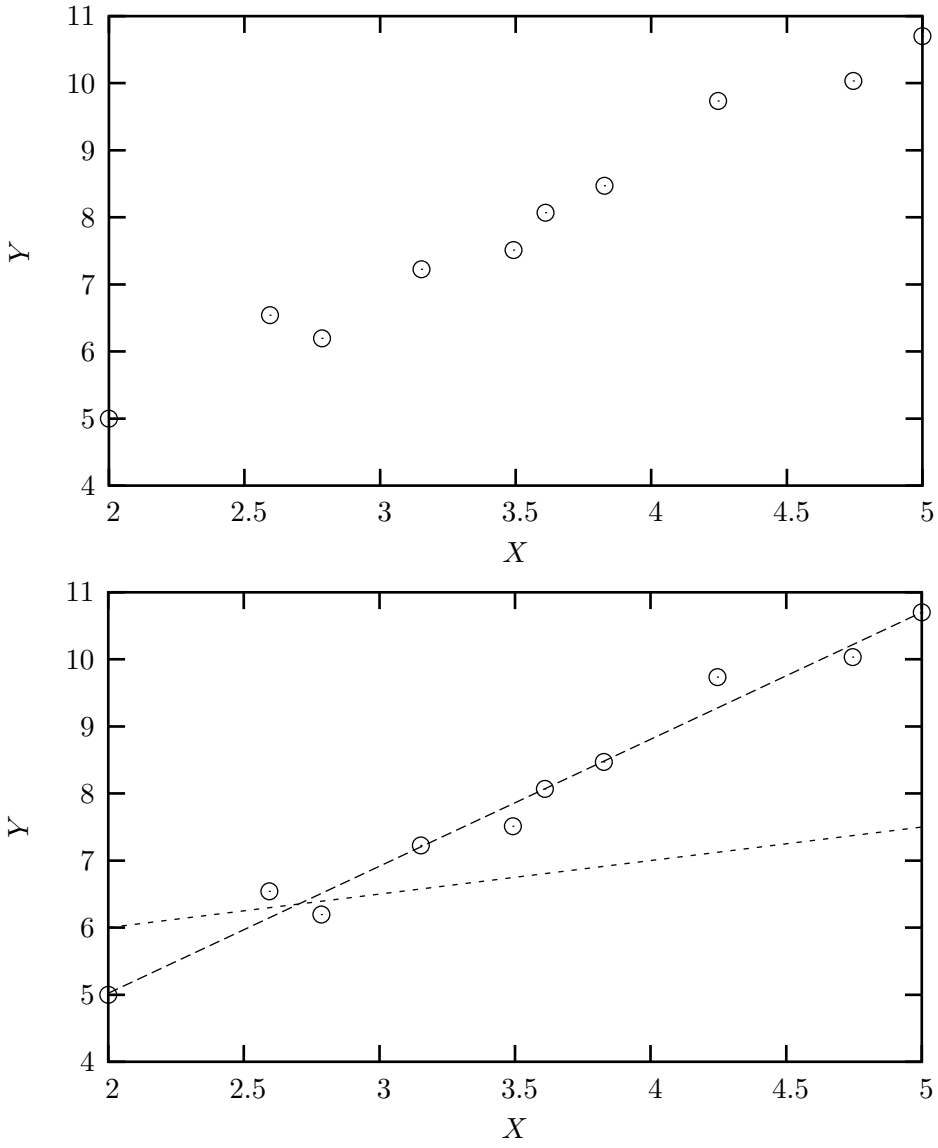


Figura 8: Esempio di un problema di regressione lineare.

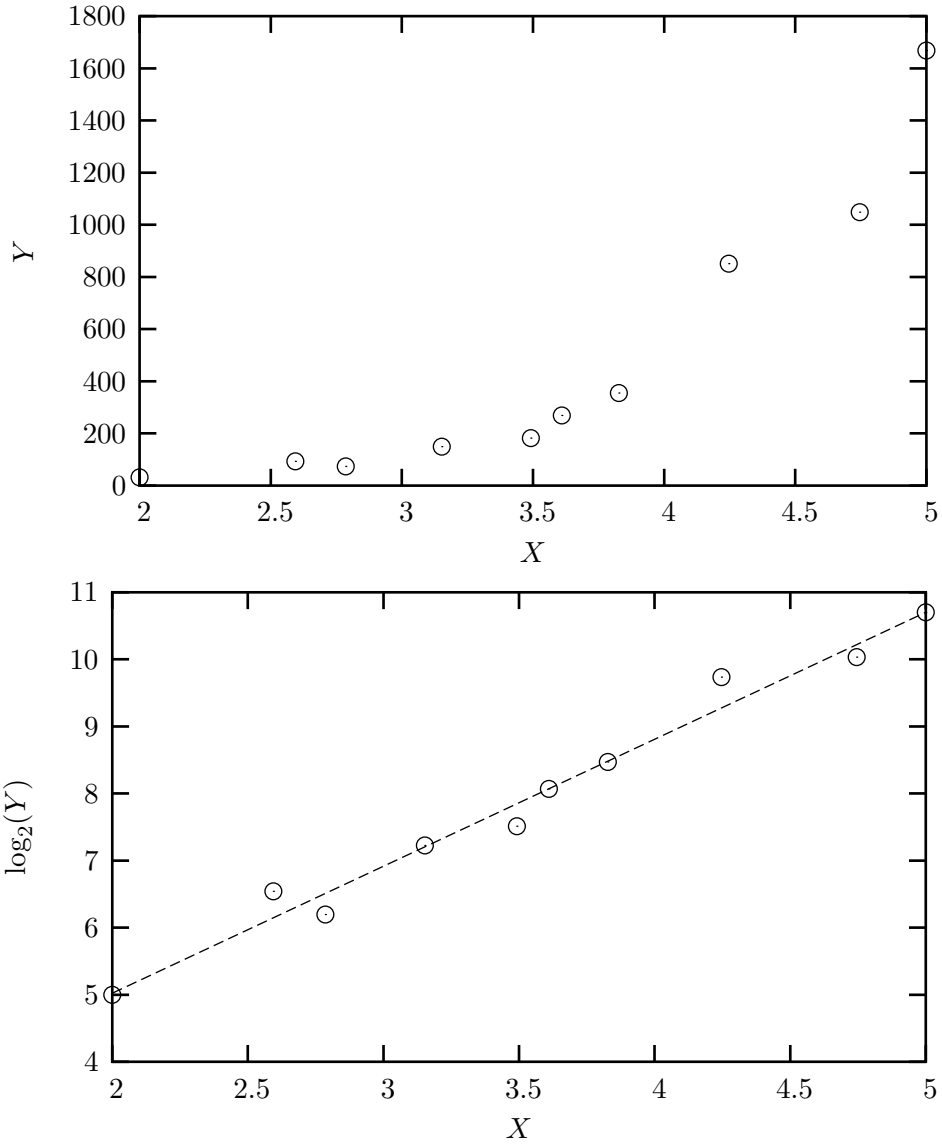


Figura 9: Esempio di un problema di regressione non lineare.

- usare un modello di ordine superiore, generalmente, un polinomio.

In quest'ultima soluzione, si sceglie come modello di regressione una curva del tipo:

$$\boldsymbol{E}(Y|X=x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

In figura 10 viene riportato un esempio di un problema risolta con una regressione polinomiale di ordine 2. Le tecniche di stima dei parametri di curve di regressione vanno oltre gli scopi di queste note.

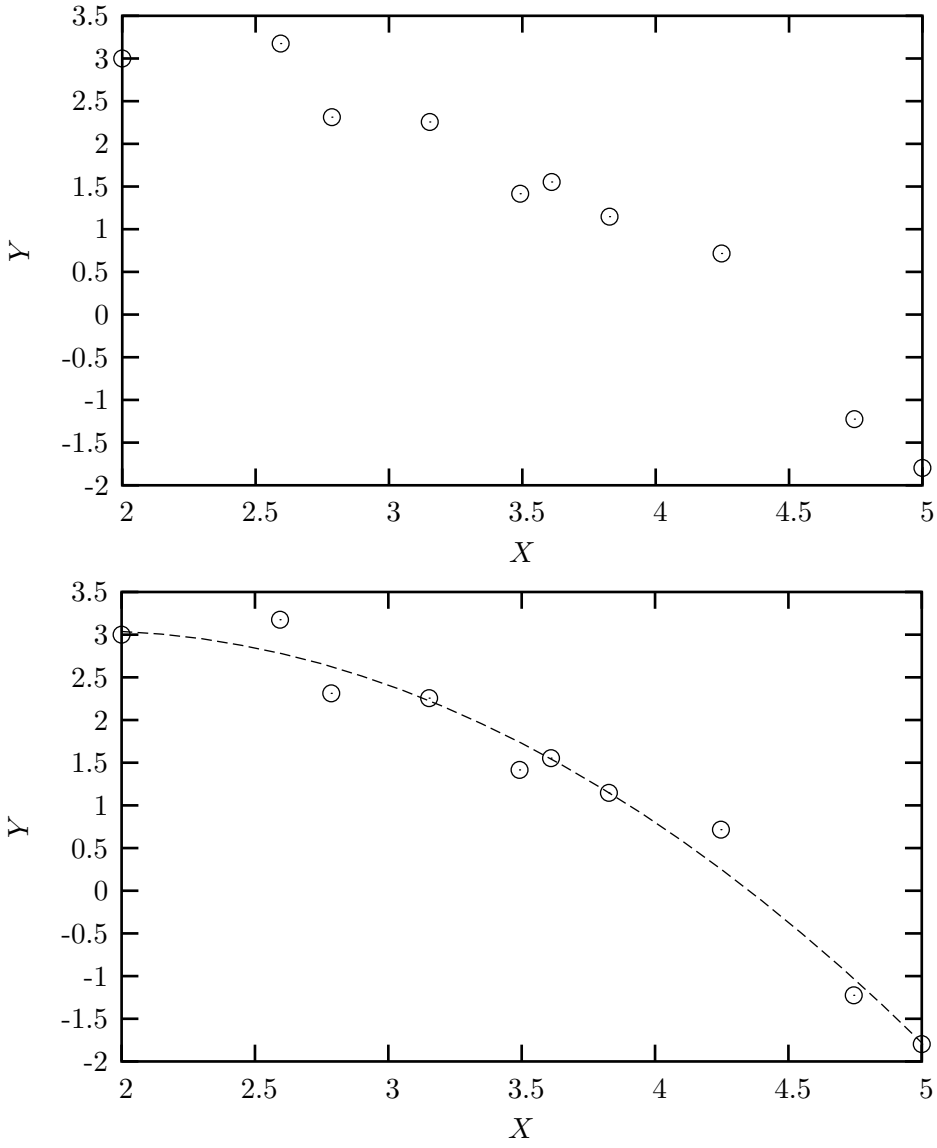


Figura 10: Esempio di un problema di regressione non lineare.

6 Dilemma bias/variance

Nel paragrafi precedenti si è visto come sia possibile costruire un approssimatore che ricostruisca l'andamento di una curva dato un numero finito di campioni affetti da incertezza.

Limitando per ora l'analisi alle curve polinomiali, risulta evidente che più alto sarà il grado del polinomio utilizzato, maggiore sarà la capacità della curva di seguire l'andamento dei dati, e quindi, tanto minore sarà lo scostamento della curva dai dati. Per illustrare questo concetto,

è stato approssimato lo stesso insieme di dati (composto da 10 campioni) con curve polinomiali di diverso grado (vedi figura 11). La vicinanza della curva ai punti è stata misurata con l'errore quadratico medio (*mean square error*, mse), espresso come:

$$\text{mse}(Y, g(X)) = \frac{1}{N}\sum_{i=1}^{N}(y_i - g(x_i))^2$$

dove N è il numero di campioni che compongono gli insiemi $X = \{x_1, \ldots, x_N\}$ e $Y = \{y_1, \ldots, y_N\}$, e $g(\cdot)$ è la curva usata per approssimare i dati. La seguente tabella riporta per ogni grado dei polinomi usati, l'mse corrispondente:

grado	1	2	3	4	5	6	7	8
mse	0.49200	0.28593	0.23448	0.066471	0.066469	0.059234	0.038639	0.037873

All'aumentare del grado, l'mse diminuisce.

Perché, dunque, non si utilizza sempre un polinomio di grado elevato per approssimare la curva di regressione? La risposta è semplice: se cercassimo la vicinanza ai dati, basterebbe scegliere una curva interpolante, ma tale curva descriverebbe, al più l'andamento di un particolare campionamento, e non il valore atteso delle grandezze campionate. In altri termini, se si dà troppa fiducia ai dati di un particolare campionamento, si rischia di perdere in generalità. Per illustrare questo concetto, proviamo a confrontare le curve ottenute con un altro insieme di dati (sempre composto da 10 elementi) campionati dalle stesse variabili aleatorie da cui sono stati ottenuti i campioni della figura 11.

L'impressione visiva è che le approssimazioni di grado più elevato non sono necessariamente le migliori. Tale impressione è confermata dal calcolo dell'mse:

grado	1	2	3	4	5	6	7	8
mse	1.2568	0.44922	0.49014	1.6778	1.6733	1.1147	0.24857	0.25354

Ripetendo la procedura più volte, il risultato non cambia:

grado	1	2	3	4	5	6	7	8
mse	0.68054	0.29312	0.27657	0.64926	0.64717	0.38396	0.34462	0.32457
mse	0.67028	0.20411	0.26397	0.46688	0.46690	0.38128	0.39837	0.34936
mse	0.74688	0.25148	0.14936	0.66293	0.66250	0.44320	0.23777	0.28492

Questo fatto deve essere messo in relazione con l'utilizzo della curva di regressione: la capacità di predire il valore di Y anche per valori sconosciuti (cioè non presenti nel campione) di X. Se cambiando campione la capacità della curva di approssimare i dati peggiora notevolmente, tale curva di regressione non è utilizzabile. Quello che si cerca non è quindi la curva con il minor mse, ma la curva che mediamente approssimi meglio i dati, dove l'insieme degli eventi su cui valutare la media sarà l'insieme di tutti i campioni possibili.

Il fenomeno per cui se una curva ha ottime prestazioni di approssimazione su un campione, la stessa curva avrà un mse elevato quando verrà valutato su un altro campione, è noto in letteratura con il nome di *dilemma bias-variance*. Esso è matematicamente formalizzato dal seguente teorema (Geman et al., 1992):

$$\boldsymbol{E}((y(\cdot) - g(\cdot))^2) = (y(\cdot) - \boldsymbol{E}(g(\cdot)))^2 + \boldsymbol{E}((g(\cdot) - \boldsymbol{E}(g(\cdot)))^2)$$

Esso descrive l'mse di uno stimatore, $g(\cdot)$, di una funzione, $y(\cdot)$, come somma di due termini. Nel nostro caso, $y(\cdot)$ è la (vera) curva di regressione, mentre $g(\cdot)$ è la stima che costruiamo a partire dal campione disponibile. Il teorema dimostra che l'mse, espresso come $\boldsymbol{E}((y(\cdot) - g(\cdot))^2)$ è sempre composto dalla somma di un termine, $(y(\cdot) - \boldsymbol{E}(g(\cdot)))^2$, detto *bias* e di un termine, $\boldsymbol{E}((g(\cdot) - \boldsymbol{E}(g(\cdot)))^2)$, detto *variance*. Il bias esprime lo scostamento che possiamo attenderci in media utilizzando lo stimatore $f(\cdot)$, dove la media va immaginata su tutti i campioni possibili. Il secondo termine, variance, esprime la variabilità dello stimatore al variare del campione.

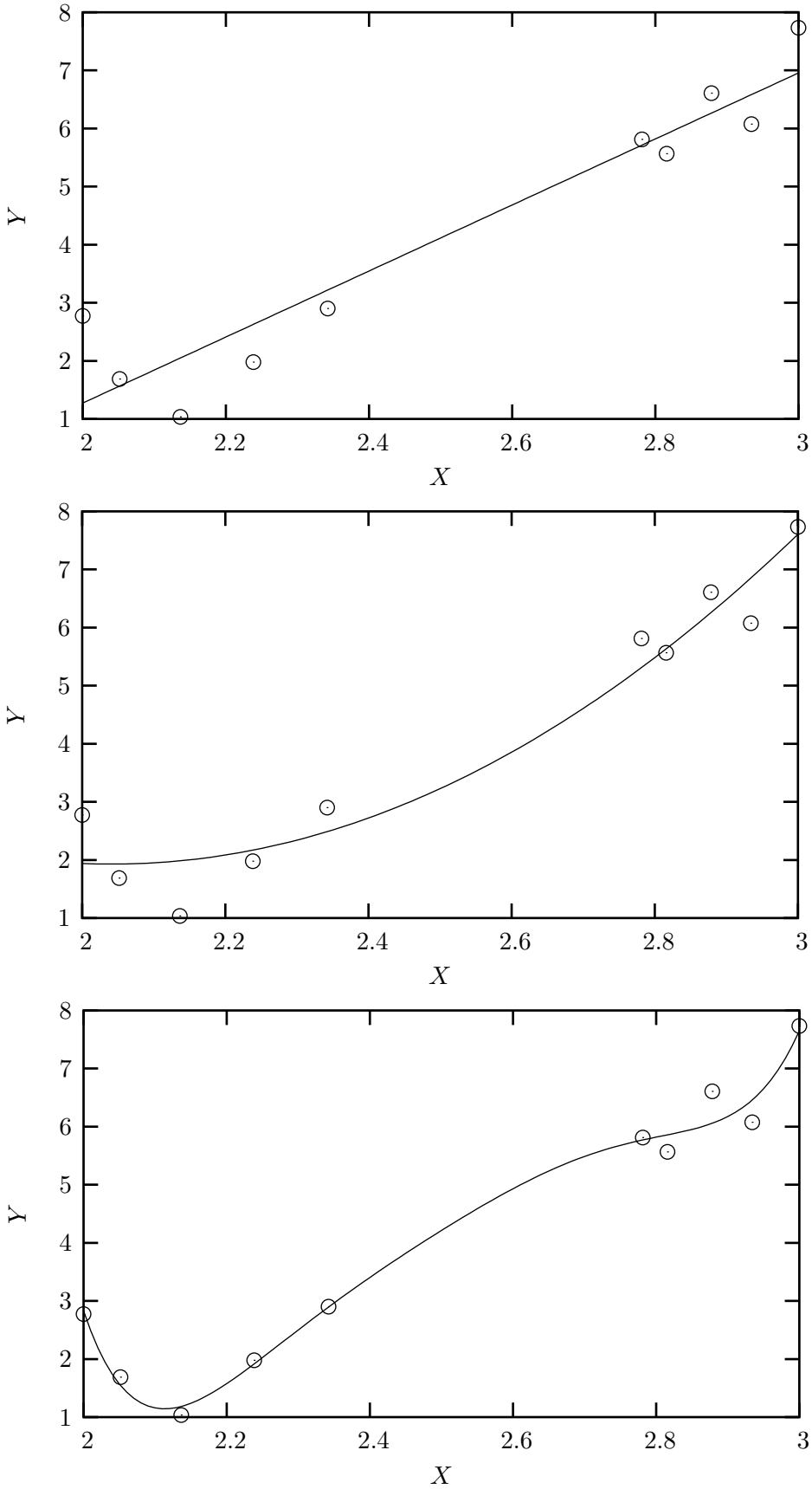
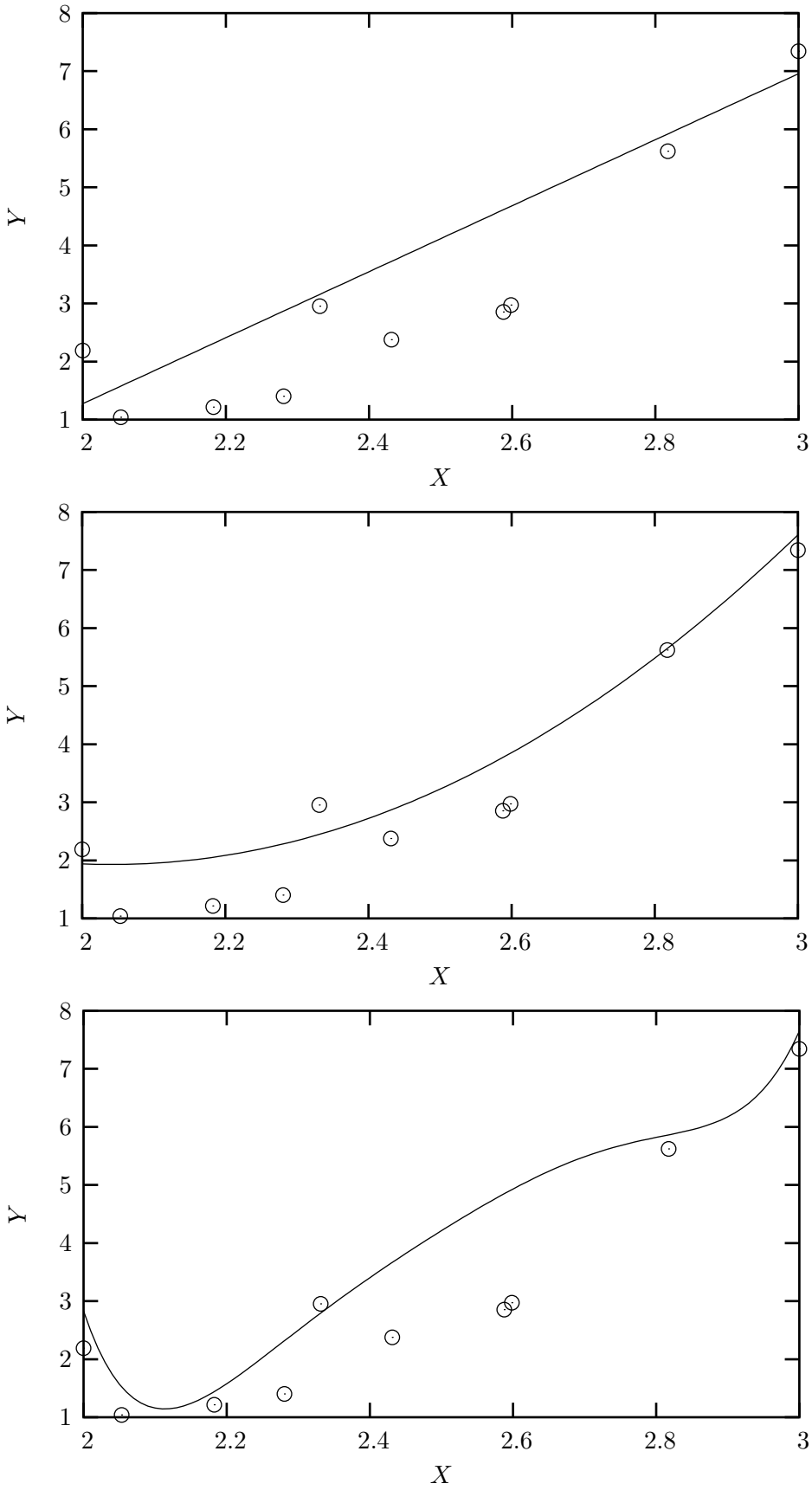


Figura 11: Approssimazione dei dati campionati con una retta (a), con una parabola (b) e con un polinomio di grado 6 (c)



(a)

(b)

(c)

Figura 12: Confronto tra un nuovo campione di dati e le approssimazioni ottenute precedentemente: una retta (a), una parabola (b) e un polinomio di grado 6 (c)

Il teorema fa capire che, anche se nel caso ottimale lo stimatore in media coincidesse con la funzione di regressione, lo stimatore che effettivamente riusciamo a costruire utilizzando un numero limitato di dati affetti da incertezza può discostarsi anche di molto dallo stimatore medio, se la variance è elevata.

Nel caso che abbiamo esaminato precedentemente, significa che un basso valore dell'mse per la curva di grado 8 è stato ottenuto a spese di un comportamento insoddisfacente nel caso generale.

La figura 13 riporta gli approssimatori medi e la dispersione degli approssimatori polinomiali di grado 1, 2 e 6. La linea a tratto continuo rappresenta la vera curva di regressione, $y(\cdot)$, la linea a tratto lungo rappresenta l'approssimatore medio, $\boldsymbol{E}(g(\cdot))$ e le linee a tratto corto rappresentano l'ampiezza della deviazione standard degli approssimatori, $\sqrt{\boldsymbol{E}((g(\cdot) - \boldsymbol{E}(g(\cdot)))^2)}$. Gli approssimatori sono stati stimati utilizzando campioni di dimensione 100. Media e deviazione standard sono state stimate su 100 sessioni di calcolo dei regressori. Si può notare che, nel caso medio, sia il polinomio di grado 2 che quello di grado 6 approssimano ragionevolmente bene la vera curva di regressione (valore di bias contenuto). Tuttavia, è evidente che la curva di grado 6 ha una dispersione (e qundi un valore di variance) molto più elevato rispetto al polinomio di grado 2. In generale, quindi, usando come approssimatore un polinomio di grado 2 si otterrà un mse più contenuto rispetto all'uso di un polinomio di grado 6.

6.1 I fattori in gioco

Quali sono le cause dell'mse e cosa influenza la ripartizione dell'mse in bias e variance? Essenzialmente, tutto è riconducibile all'incertezza.

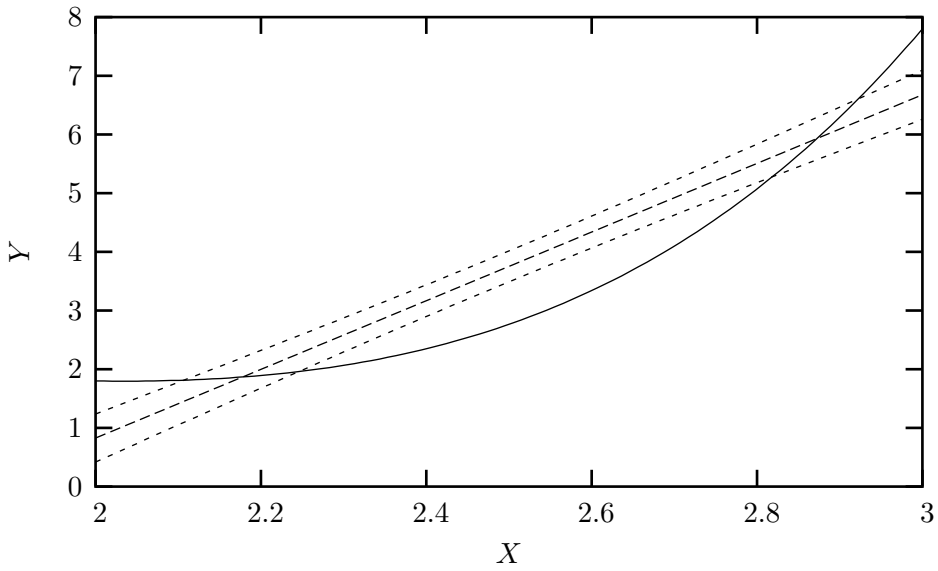
A sua volta, l'incertezza è legata ai seguenti fattori:

casualità I dati del campione sono affetti da un disturbo stocastico. Esso può essere dovuto sia al fenomeno misurato che allo strumento usato per misurarlo (dal punto di vista pratico, non vi è differenza tra i due);

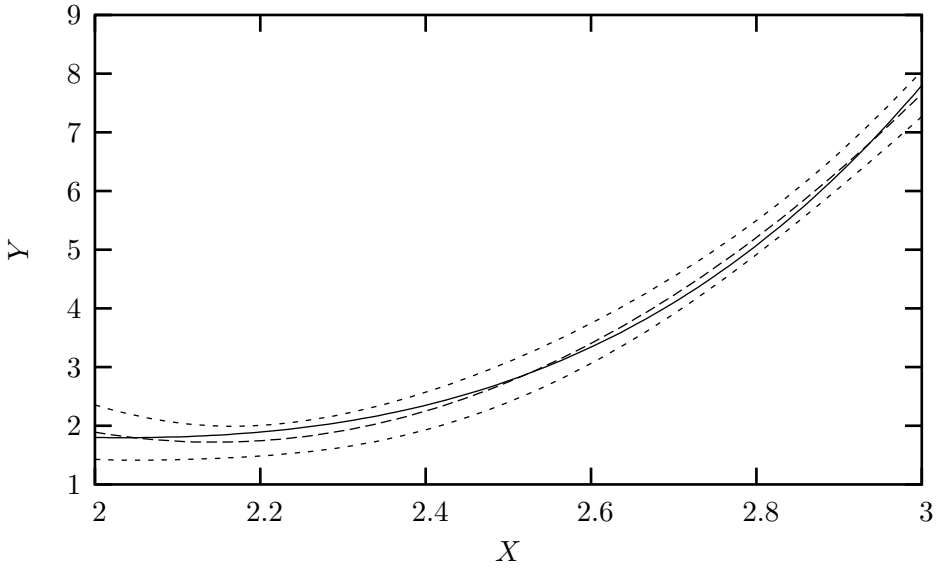
numero di campioni I campioni disponibili nella pratica sono necessariamente in numero finito. Il numero di campioni potrebbe essere insufficiente per stimare tutti i parametri che caratterizzano la curva di regressione. Per esempio, con due punti non possiamo costruire una parabola.

modello La natura della relazione che lega le variabili è generalmente sconosciuta. Il modello di curva che si sceglie per stimare la regressioe potrebbe non essere quello vero. Tipicamente, è solo quello che meglio spiega il comportamento dei dati nell'intervallo da essi coperto.

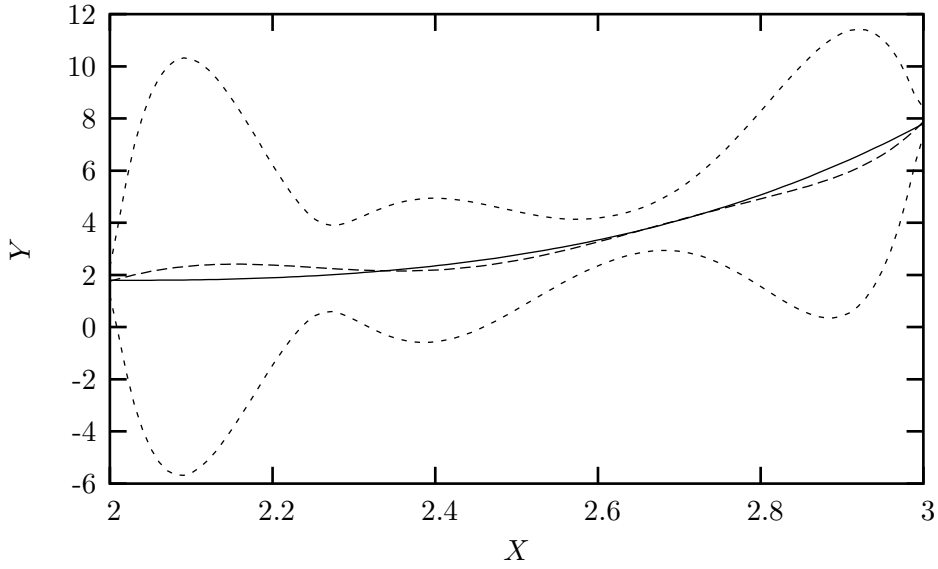
A parità di disturbo aleatorio sui dati campionati e di numero di campioni, l'uso di un modello con un numero di gradi di libertà elevato causa un basso valore di bias, ma un alto valore di variance. Al contrario, un modello con un basso numero di parametri avrà un valore di bias molto elevato, ma un valore di variance contenuto.



(a)



(b)



(c)

Figura 13: Bias-variance dei regressori: retta (a), parabola (b) e polinomio di grado 6 (c). La linea continua, quella a tratto lungo e quelle a tratto corto rappresentano, rispettivamente, la vera curva di regressione, l'approssimatore medio e l'ampiezza della deviazione standard degli approssimatori.