Università degli Studi di Palermo

Corso di Laurea in Informatica

Appunti di

STATISTICA

Prof. Marcello Chiodi

Autori: Google&Wikipedia
Copy&Paste: Viviana Pezzano

Indice

Capitolo 1

Concetti di Calcolo delle Probabilità

1.1 Introduzione

Quello di **probabilità** è un concetto che, utilizzato a partire dal '600, è diventato con il passare del tempo la base di diverse discipline scientifiche. In particolare su di esso si basa una branca della *statistica* (la statistica inferenziale), cui faranno ricorso numerose scienze sia naturali che sociali.

Vi sono tre definizioni:

- **Definizione classica** (Laplace): si applica ad esperimenti casuali i cui eventi elementari sono ritenibili equiprobabili. La probabilità di un evento è il rapporto tra il numero dei casi favorevoli e il numero dei casi possibili, purchè questi ultimi siano ugualmente possibili.
- **Definizione frequentista** (von Mises): si applica ad esperimenti casuali i cui eventi elementari non sono ritenibili ugualmente possibili, ma l'esperimento è ripetibile più volte sotto le stesse condizioni. La probabilità di un evento è associata alla frequenza relativa del verificarsi dell'evento stesso, su un elevato numero di prove (tendenti all'infinito).
- **Definizione soggettiva** o soggettivista (de Finetti, Savage, Ramsey): si applica a esperimenti casuali i cui eventi elementari non sono ritenibili ugualmente possibili e l'esperimento non è ripetibile più volte sotto le stesse condizioni. La probabilità di un evento è fornita secondo l'esperienza personale e le informazioni disponibili.

Quindi se i casi possibili sono n e l'insieme dei casi favorevoli sono n_A, per la teoria classica la probabilità che accada l'evento A sarà:

$$p_A = \frac{n_A}{n}$$

mentre per la teoria frequentista essa sarà:

$$p_A = \lim_{n\to\infty} \frac{n_A}{n}$$

Bisogna però sottolineare che questo limite non va inteso come un'usuale convergenza puntuale in analisi matematica, bensì come una convergenza in probabilità. Si avrà cioè che:

$$\frac{n_A}{n} \to p_A \quad \text{per} \quad n \to \infty \quad \text{se}$$

$$\forall\, \varepsilon > 0 \quad p\left(\left|\frac{n_A}{n} - p_A\right|\right) \to 1 \quad \text{per} \quad n \to \infty$$

Cioè le fluttuazioni di $\frac{n_A}{n}$ lontano da p_A diventano sempre più improbabili al tendere all'infinito di n.

In entrambe le definizioni la probabilità è una funzione il cui insieme di definizione è un numero reale compreso fra 0 e 1, estremi inclusi.

La *teoria classica* considera che tutti i casi siano equiprobabili cosa che, invece, nella realtà non accade sempre. La *definizione frequentista* poggia invece su quella che è definita **legge empirica del caso** ovvero legge dei grandi numeri: in una successione di prove effettuate nelle medesime condizioni, la frequenza di un evento si avvicina alla probabilità dell'evento stesso, e l'approssimazione tende a migliorare con l'aumentare delle prove.

L'**impostazione assiomatica** della probabilità venne proposta da Andrey Nikolaevich Kolmogorov nel 1933, sviluppando la ricerca che era ormai cristallizzata sul dibattito fra quanti consideravano la probabilità come limiti di frequenze relative (cfr. impostazione frequentista) e quanti cercavano un fondamento logico della stessa.

La sua impostazione assiomatica si mostrava adeguata a prescindere dall'adesione a una o all'altra scuola di pensiero ed è basata sui seguenti punti:

1. Gli eventi sono sottoinsiemi di uno spazio S, e formano una classe additiva A.
2. Ad ogni a appartenente alla classe A è assegnato un numero reale non negativo $P(a)$ e mai superiore ad uno, detto probabilità di a.
3. $P(S) = 1$, ovvero la probabilità di un evento certo è pari ad 1
4. Se l'intersezione tra a e b è vuota, allora $P(a \cap b) = P(a) + P(b)$
5. Se $A(n)$ è una successione decrescente di eventi e al tendere di n all'infinito l'intersezione degli $A(n)$ tende a 0, allora $\lim P(A(n)) = 0$

Dai suddetti assiomi derivano alcuni *teoremi fondamentali*, quali

- il **teorema della probabilità totale**:

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

- il **teorema della probabilità composta**:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

- il **teorema della probabilità assoluta**, il quale afferma che, se $A_1, \ldots, A_n$ formano una partizione dello spazio campionario di tutti gli eventi possibili Ω (ossia $A_i \cap A_j = \varnothing \quad \forall\ i \neq j$ e $\cup_{i=1}^{n} A_i = \Omega$) e B è un qualsiasi evento, allora:

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

- il **teorema di Bayes** deriva dai due teoremi precedenti e afferma che, considerando un insieme di alternative $A_1, A_2, ..., A_n$ (partizione dello *spazio degli eventi*) si trova la seguente espressione per la probabilità condizionata:
$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

Dagli assiomi derivano anche concetti chiave come:

- la **probabilità condizionata**, ossia la probabilità che si verifichi A dato il verificarsi dell'evento B:
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
Una precisazione importante riguarda il fatto che la probabilità condizionata ha senso solo se l'evento B si può verificare (non è l'*evento impossibile*).

- l'**indipendenza stocastica** di due eventi A e B, che si ha quando il verificarsi di uno non modifica la probabilità di verificarsi dell'altro, ovvero quando:
$$P(A|B) = P(A) \qquad P(B|A) = P(B)$$
queste due condizioni si possono sintetizzare con la formula:
$$P(A \cap B) = P(A) \cdot P(B)$$
In altre parole, dire che due eventi sono indipendenti tra loro significa dire che il fatto di sapere che uno di essi si è verificato non modifica la valutazione di probabilità sul secondo.

1.2 Definizioni Utili

Definizione 1.2.1. Si definisce **spazio campionario** o insieme universo (generalmente indicato dalle lettere S, Σ o U) l'insieme dei possibili risultati di un esperimento casuale.

Definizione 1.2.2. Si definisce **evento** un insieme di risultati (un sottoinsieme dello spazio campionario) al quale viene assegnata una probabilità.

Due eventi si dicono *mutualmente esclusivi* o *incompatibili* se non possono essere contemporaneamente veri, cioè se $E \cap F = \varnothing$.

Una collezione di eventi $E_1, ..., E_n$ si dice *mutualmente esclusiva* se tutte le possibili coppie di eventi sono tra loro incompatibili, cioè $\forall\ i, j \quad E_i \cap E_j = \varnothing$.

Due eventi si dicono *necessari* o *esaustivi* se almeno uno dei due deve essere vero, cioè $E \cup F = \Omega$ (dove Ω è l'evento certo).

Definizione 1.2.3. Uno **spazio di probabilità** è una terna $(\Omega, \mathfrak{F}, \mathbb{P})$ dove Ω è un insieme non vuoto, $\mathfrak{F}$ è una σ-algebra su Ω e $\mathbb{P}$ è una misura positiva tale che $\mathbb{P}(\Omega) = 1$ (detta misura di probabilità). Più brevemente, uno spazio di probabilità è uno spazio di misura positiva, tale che l'intero spazio abbia misura 1.

Definizione 1.2.4. Si definisce **misura di probabilità** la funzione che assegna agli esiti di un determinato esperimento la probabilità che tali esiti si realizzino. E' importante osservare che la misura di probabilità non assegna la probabilità ai singoli punti dello spazio campionario (gli eventi elementari) bensì a sottoinsiemi di esso (gli eventi).

Definizione 1.2.5. Sia Ω un insieme non vuoto, e sia $\mathfrak{F}$ una famiglia di sottoinsiemi di Ω (ovvero sia un sottoinsieme dell'insieme delle parti di Ω). Diremo che $\mathfrak{F}$ è un'**algebra** su Ω se:

- L'insieme vuoto $\emptyset$ appartiene ad $\mathfrak{F}$: $\emptyset \in \mathfrak{F}$
- Se un insieme $A \in \mathfrak{F}$ allora anche il suo complementare vi appartiene:

$$A \in \mathfrak{F} \quad \Rightarrow \quad A^c \in \mathfrak{F}$$

- Se due insiemi A, B sono in $\mathfrak{F}$, allora la loro unione è in $\mathfrak{F}$:

$$A, B \in \mathfrak{F} \qquad \Rightarrow \qquad A \cup B \in \mathfrak{F}$$

Definizione 1.2.6. Una **σ-algebra** su di un insieme Ω è una famiglia di sottoinsiemi di Ω che abbia delle proprietà di stabilità rispetto ad alcune operazioni insiemistiche, in particolare l'operazione di *unione numerabile* e di passaggio al *complementare*. Diremo che $\mathfrak{F}$ è una σ-algebra su Ω se:

1. L'insieme vuoto $\emptyset$ appartiene ad $\mathfrak{F}$: $\emptyset \in \mathfrak{F}$
2. Se un insieme $A \in \mathfrak{F}$ allora anche il suo complementare vi appartiene:

$$A \in \mathfrak{F} \quad \Rightarrow \quad A^c \in \mathfrak{F}$$

3. Se gli elementi A_i di una famiglia numerabile di insiemi $\{A_i\}_{i\in\mathbb{N}}$ sono in $\mathfrak{F}$, allora la loro unione è in $\mathfrak{F}$:

$$A_i \in \mathfrak{F}, \quad \forall\, i \in \mathbb{N} \qquad \Rightarrow \qquad \bigcup_{i\in\mathbb{N}} A_i \in \mathfrak{F}$$

Se Ω è un insieme non vuoto, e $\mathfrak{F}$ è una σ-algebra su Ω, la coppia $(\Omega, \mathfrak{F})$ viene detta **spazio misurabile**, e gli elementi di $\mathfrak{F}$ (ossia i sottoinsiemi di Ω che sono in $\mathfrak{F}$) vengono detti **insiemi misurabili**.

Definizione 1.2.7. Si definisce **variabile aleatoria** una funzione a valori reali X definita su uno spazio campionario Ω, ossia: $X : \Omega \to \mathbb{R}$.

A ogni esperimento otteniamo un numero, $X(e)$, che è il valore che la variabile aleatoria assume sul risultato dell'esperimento, l'evento elementare e. Possiamo quindi considerare l'insieme di tutti i valori possibili (detto il *range* della variabile aleatoria) come un nuovo spazio campionario e assegnare una probabilità ai possibili valori della variabile aleatoria.

A ogni valore x nel range della variabile aleatoria X, assegnamo la probabilità che X assuma il valore x. Questo valore è dato dalla **probabilità** $P(E)$ dell'evento

$$E = \{e \in \Omega \mid X(e) = x\}$$

ovvero la retroimmagine di x tramite X.
Ad una variabile casuale X si associa dunque la sua **legge di distribuzione** di probabilità P_X, che assegna ad ogni sottoinsieme dell'insieme dei possibili valori di X, la probabilità che la variabile casuale X assuma valore in esso.
In formule, se X è una variabile casuale che ha valori in Ω ed E è un sottoinsieme di Ω, la *distribuzione di probabilità* di X in E vale:

$$P_X(E) := P(X \in E) = \nu(X^{-1}(E))$$

dove ν è la *misura di probabilità* definita sullo spazio campionario.

Otteniamo così, al posto dello spazio campionario Ω, che in genere è assai complesso, un semplice spazio campionario formato da un insieme di numeri. Il maggiore vantaggio di questa sostituzione è che molte variabili aleatorie, definite su spazi campionari anche molto diversi tra loro, danno luogo a una stessa *distribuzione* di probabilità sull'asse reale.
Denoteremo con lettere romane maiuscole le variabili aleatorie e con lettere romane minuscole i valori assunti da una variabile aleatoria.

Esempio 1.2.1. Considerando un lancio di un dado, è possibile creare una variabile casuale che associ il numero 1 se esce pari ed il numero 0 se esce dispari, si ha quindi:

$$f(\omega_1 = 1) = 0 \qquad f(\omega_2 = 2) = 1 \qquad f(\omega_3 = 3) = 0$$
$$f(\omega_4 = 4) = 1 \qquad f(\omega_5 = 5) = 0 \qquad f(\omega_6 = 6) = 1$$

da cui:

$$X = \begin{cases} 0 & 1 \\ P(X=0) & P(X=1) \end{cases}$$

Si ha quindi:

$$P(X = 0) = P(\omega_1) + P(\omega_3) + P(\omega_5) = \frac{1}{2}$$
$$P(X = 1) = P(\omega_2) + P(\omega_4) + P(\omega_6) = \frac{1}{2}$$

◇

Definizione 1.2.8. Una variabile aleatoria si dice **discreta** se essa può assumere solo un numero finito o numerabile di valori. In questo caso si definisce la funzione f, chiamata **funzione di densità discreta** o *funzione di distribuzione*, ossia una funzione di variabile reale che assegna ad ogni valore possibile di X la probabilità dell'evento elementare $(X = x)$, cioè la probabilità che la variabile X assuma esattamente quel valore:

$$f(x) := P(X = x)$$

La funzione f si estende a tutti i valori reali, ponendo il suo valore uguale a 0 al di fuori dei valori che può assumere X.
La funzione f soddisfa inoltre la condizione di normalizzazione

$$\sum_x f(x) = 1$$

dove la somma è estesa a tutti i possibili valori assunti da X, che ci dice che la probabilità che X assuma almeno uno dei valori possibili è 1.

Definizione 1.2.9. Si definisce **funzione di ripartizione**, o *funzione di distribuzione cumulativa* della variabile aleatoria X, la funzione definita da:

$$F(x) := P(X \leqslant x)$$

F quindi rappresenta la probabilità che la variabile aleatoria X assuma un qualunque valore minore o uguale a x.

La funzione di distribuzione gode della seguenti proprietà:

1. $F(x)$ è una funzione non decrescente di x
2. $\lim_{x\to+\infty} F(x) = 1$
3. $\lim_{x\to-\infty} F(x) = 0$
4. F è continua a destra, ovvero $\lim_{x\to x_0^+} F(x) = F(x_0) \quad \forall\, x_0 \in \mathbb{R}$

Le proprietà 2 e 3 sono "ovvie": esse ci dicono semplicemente che la probabilità di assumere un qualsiasi valore è 1 e quella di non assumere alcun valore è 0.

Definizione 1.2.10. Se la variabile casuale X è **continua**, cioè l'insieme dei possibili valori ha la potenza del continuo, la probabilità di un preciso valore x è nulla, mentre ha senso definire la probabilità che si verifichi un numero entro un intervallo. Si definisce dunque la **funzione di densità di probabilità**, cioè la funzione f non negativa tale per cui

$$P(X \in E) = \int_E f(x)\ dx$$

La funzione f deve essere tale da soddisfare la condizione di normalizzazione :

$$\int_{-\infty}^{+\infty} f(x)\ dx = 1$$

vale a dire che la probabilità di assumere un qualsiasi valore reale è 1.

Definizione 1.2.11. Come già fatto per le varibili discrete, possiamo definire la **funzione di distribuzione**:

$$F(x) := \int_{-\infty}^{x} f(x)\ dx = P(X \leqslant x)$$

La funzione di distribuzione così definita gode delle stesse proprietà 1, 2, 3 della funzione di distribuzione per le variabili aleatorie discrete. In più la funzione di distribuzione di una varibile aleatoria continua risulta essere una funzione continua (e non solo continua a destra)

Definizione 1.2.12. Si definisce **valore atteso** (chiamato anche aspettazione, attesa, media o speranza matematica) di una variabile casuale reale X, un numero $\mathbb{E}[X]$ che formalizza l'idea euristica di valore medio di un fenomeno aleatorio. In generale il valore atteso di una variabile casuale discreta (che assuma cioè solo un numero finito o una infinità numerabile di valori) è dato dalla somma dei possibili valori di tale variabile, ciascuno moltiplicato per la probabilità di essere assunto (ossia di verificarsi), cioè è la media ponderata dei possibili risultati.

Nel caso di variabile casuale *discreta* che ammette funzione di probabilità $p(x)$, può essere calcolata come:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \; p(x_i)$$

Nel caso di variabile casuale *continua* che ammetta densità $f(x)$ il calcolo diventa:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \; f(x) \; dx$$

Definizione 1.2.13. Quando della stessa grandezza si possiede un campione di valori frutto di n misurazioni si possono riassumere le informazioni inerenti alla grandezza derivanti dalle singole misure atrraverso la **media** che, in questo caso, costituisce la miglior stima possibile per la grandezza in esame.

Se le singole misure si possono considerare equivalenti l'una all'altra senza che ve ne siano di alcune più importanti o privilegiate, allora definiamo la media aritmetica come segue:

$$\bar{x} = \frac{x_i + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Nel caso in cui i singoli dati abbiano pesi diversi, allora si ricorre a quella che viene definita *media pesata*.

Definizione 1.2.14. Si definisce **moda** il valore più probabile che la variabile aleatoria può assumere: quando trattiamo campioni di dati frutto ad esempio di diverse misure della stessa grandezza, allora definiamo la moda come il valore più popolare del campione. Per capire meglio questa definizione pensiamo ad un istogramma: la moda è costituita in questo caso dal valore corrispondente alla colonna più alta.

Definizione 1.2.15. Si definisce **mediana**, ad esempio in un istogramma, l'ascissa corrispondente al punto in cui l'area delimitata dall'istogramma si divide in due parti uguali: in pratica il numero di dati che sta alla destra della mediana (quelli maggiori) è uguale al numero di dati alla sinistra della mediana (quelli minori).

Accanto alle cosiddette caratteristiche di posizione quali la media, la moda e la mediana, esistono altre caratteristiche, ciascuna delle quali descrive una proprietà della distribuzione in esame.

Definizione 1.2.16. I **momenti** forniscono indicazioni sulla dispersione rispetto al valore centrale, sull'asimmetria della distribuzione, ecc. Nelle applicazioni pratiche si ha a che fare con *momenti iniziali* e con *momenti centrali*.

Definizione 1.2.17. Si definisce **momento iniziale di ordine** s di una variabile aleatoria discreta la seguente espressione:

$$a_s = \sum_{i=1}^{N} x_i^s \; p_i$$

dove le x_i rappresentano gli N valori possibili della variabile mentre le p_i sono le probabilità che la variabile assuma i valori x_i. E' facile notare che nel caso in cui si stia trattando eventi equiprobabili su un intervallo di N eventi totali, le singole p_i assumono tutte valore uguale a $\frac{1}{N}$, per cui il momento iniziale di ordine 1 non è altro che la media della variabile.

Definiamo il corrispettivo momento nel caso continuo come:

$$a_s = \int_{-\infty}^{+\infty} x^s \; f(x) \; dx$$

dove $f(x)$ rappresenta la funzione densità di probabilità per la variabilie aleatoria continua.

Definizione 1.2.18. Il **momento centrale di ordine** s di una variabile aleatoria discreta si esprime attraverso la somma:

$$\mu_s = \sum_{i=1}^{N} (x_i - \bar{x})^s \; p_i$$

dove le x_i rappresentano gli N valori possibili della variabile, $\bar{x}$ rappresenta la media della variabile e le p_i sono le probabilità che la variabile assuma i valori x_i.

Il momento centrale di ordine 1 è nullo, mentre il momento centrale di ordine 2 prende il nome di *varianza*.

Definizione 1.2.19. Si definisce **varianza**, e viene solitamente indicata con σ^2 (dove σ è la deviazione standard), la seguente espressione:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

dove μ rappresenta la media aritmetica dei valori x_i, ossia:

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

Nel caso si tratti di *valori ponderati*, allora la definizione diventa:

$$\sigma^2 = \frac{\sum_{j=1}^{k} f_j (x_j - \mu)^2}{\sum_j f_j}$$

in questo caso μ è la media aritmetica ponderata, che vale:

$$\mu = \frac{\sum_i x_i f_i}{\sum_i f_i}$$

La varianza è un indicatore di dispersione in quanto è nulla solo nei casi in cui tutti i valori sono uguali tra di loro (e pertanto uguali alla loro media) e cresce con il crescere delle differenze reciproche dei valori.

Trattandosi di una somma di valori (anche negativi) al quadrato, è evidente che la varianza non sarà mai negativa.

Più in generale, se X è una variabile casuale, la sua varianza si definisce come:

$$\sigma^2[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Essendo $\mathbb{E}[X]$ il valore atteso della variabile casuale X. Si osservi che essendo la variabile casuale $(X - \mathbb{E}[X])^2$ sempre positiva, il suo valore atteso, ovvero la varianza di X, sarà anch'esso positivo.

Definizione 1.2.20. La **deviazione standard** o scarto quadratico medio, è un indice di dispersione (vale a dire una misura di variabilità di una popolazione o di una variabile casuale).

Se non indicato diversamente, è semplicemente la radice quadrata della varianza, la quale viene coerentemente rappresentata con il quadrato di sigma (σ^2).

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Definizione 1.2.21. L'indice di **curtosi** è uno degli indici relativi alla forma di una distribuzione, e costituisce una misura dello spessore delle code di una funzione di densità, ovvero il grado di appiattimento di una distribuzione. Il coefficiente di curtosi è dato dalla formula:

$$\gamma_2 = \beta_2 - 3 \qquad \text{dove} \qquad \beta_2 = \frac{m_4}{m_2^2}$$

è l'indice di curtosi, e m_4 e m_2 sono rispettivamente il *momento centrale* di ordine 4 e 2. Nel caso di una variabile casuale normale, $\beta_2 = 3$, così che il coefficiente di curtosi γ_2 risulta pari a zero. Se il coefficiente di curtosi è:

- > 0 la curva si definisce leptocurtica, cioè più appuntita di una normale.
- < 0 la curva si definisce platicurtica, cioè più piatta di una normale.
- $= 0$ la curva si definisce normocurtica, cioè piatta come una normale.

Definizione 1.2.22. In statistica una distribuzione, una funzione di probabilità, una funzione di densità o comunque una variabile casuale si dicono **simmetriche** quando esiste un valore X_m (che coincide con la media aritmetica ovvero con il valore atteso) per il quale a tutti i valori minori X_a (con $X_a = X_m - \Delta$, con $\Delta > 0$) corrisponde una frequenza o funzione di probabilità o funzione di densità identica a quella che corrisponde al valore $X_b = X_m + \Delta$. In altre parole, ciò è verificato laddove vale l'uguaglianza $f(X_m + \Delta) = f(X_m - \Delta)$, dove $f(\,\cdot\,)$ denota la funzione di densità di probabilità (nel caso di variabili casuali continue) o la funzione di massa di probabilità (nel caso di variabili casuali discrete). In generale viene usato la statistica di simmetria:

$$\beta_1 = \frac{m_3^2}{m_2^3}$$

ove m_3 e m_2 sono rispettivamente il momento centrale secondo e terzo.
Tale indicatore è:

- $= 0$, nel caso di perfetta simmetria;
- < 0, per l'asimmetria a destra;
- > 0, per l'asimmetria a sinistra.

Definizione 1.2.23. La **funzione generatrice dei momenti** $g(t)$ di una variabile casuale X è definita come il valore atteso di e^{tX}, dove esso è finito (e ciò può accadere solo in un intorno dello 0, in cui vale 1 indipendentemente da X). Infatti tale valore atteso potrebbe essere infinito e in tal caso si dice semplicemente che X non possiede funzione generatrice dei momenti.

Nel caso di variabili casuali discrete si ottiene

$$g(t) = \mathbb{E}[e^{tX}] = \sum_{i=1}^{n} p_i \; e^{tx_i}$$

mentre per la variabili casuali continue:

$$g(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{+\infty} e^{tx} \; f_X(x) \; dx$$

dove p_i, $i = 1, \ldots, n$, $f_X(x)$ denotano le funzioni di massa (densità nel caso continuo) della v.c. in questione.

Data la variabile aleatoria X con funzione generatrice di momenti $g(t)$, si ha che la derivata prima, con $t = 0$, coincide con il valore atteso di X:

$$g'(0) = \mathbb{E}(X)$$

In generale vale che, dalla f.g.m., è possibile ricavare i **momenti semplici** di ordine k derivando k volte $g(t)$ con $t = 0$, in formula:

$$\mu_n = g^n(0) = \frac{d^n g}{dt^n}|_{t=0} = \mathbb{E}(X^n)$$

1.3 Distribuzioni

1.3.1 Distribuzione Binomiale

Molti tipi di esperimenti hanno in comune la caratteristica che i loro risultati possono essere raggruppati in due classi, generalmente indicate con i nomi convenzionali di *successo* e *insuccesso*. L'esempio paradigmatico è quello del lancio di una moneta, dove si può considerare, p.e. successo l'uscita di una "testa" e insuccesso l'uscita di una "croce".

La variabile aleatoria rilevante in questo tipo di esperimenti è quella che conta il numero di successi su un dato numero di ripetizioni indipendenti dello stesso esperimento.

Un'altra importartante caratteristica di molte serie di esperimenti è che i singoli esperimenti della successione sono indipendenti, ovvero l'esito di un esperimento non influenza gli esperimenti precedenti.

Questo è quanto avviene in una serie di lanci di una moneta. Il risultato del lancio $n + 1$-esimo non è influenzato dai precedenti n lanci, nel senso che la probabilità di ottenere una testa o una croce non dipende da quante teste e

quante croci si sono ottenute nei lanci precedenti. Inoltre, come è lecito assumere se si lancia sempre la stessa moneta, la probabilità di successo rimane invariata per ogni esperimento della succesione.

Questo tipo di esperimenti ripetuti è comunemente indicato con il nome di **prova di Bernoulli**.

La **variabile casuale binomiale** (o aleatoria) è una variabile casuale discreta che descrive una sequenza di prove dicotomiche stocasticamente indipendenti tra loro.

La distribuzione di probabilità di tale variabile aleatoria è detta **distribuzione binomiale**.

La variabile aleatoria binomiale è caratterizzata da due parametri:

- p: la probabilità di successo della singola prova ($0 < p < 1$).
- $q = 1 - p$: la probabilità di insuccesso della singola prova
- n: il numero di prove effettuate.

Definito k il numero di successi ottenuti in n prove, la distribuzione di probabilità associa ad ogni possibile valore di k (da 0 ad n) è la seguente:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Il coefficiente binomiale:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

esprime i modi distinti (combinazioni) in cui possono essere ripartiti i k successi negli n tentativi.

Considerando il binomio di Newton, si dimostra che la somma di tutte le probabilità della distribuzione è uguale ad 1:

$$\sum_{k=0}^{n} P(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1$$

La funzione generatrice dei momenti risulta:

$$g(t) = \left(1 - p + pe^t\right)^n$$

Il valore atteso μ e la varianza σ^2, considerando che essa è pari alla somma di n variabili casuali di Bernoulli, sono:

$$\mu = P(X) = \sum_{k=1}^{n} k \cdot P(X = k) = np$$

$$\sigma^2(X) = P(X^2) - (P(X))^2 = np\,(1-p)$$

Se n è molto grande (orientativamente $n > 50$) e p molto piccolo, tale che np è, orientativamente, minore di 10 e $p(1-p)$ quasi uguale a p, allora la binomiale può essere approssimata con una variabile casuale poissoniana ove $\lambda = np$.

Se n è molto grande, ma $np > 10$ (e dunque non vale l'approssimazione con la poissoniana), allora la binomiale può essere approssimata con una variabile casuale normale con valore atteso pari a np e varianza uguale a npq : $N(np; npq)$.

1.3.2 Distribuzione di Poisson

La **variabile casuale poissoniana** è una variabile casuale discreta, detta pure degli *eventi rari*. Esprime la probabilità di un numero di eventi che si verificano in un periodo di tempo fissato, se questi eventi hanno una media conosciuta e accadono indipendentemente dall'evento precedente.

La v.c. poissoniana è definita con la funzione di probabilità:

$$P(x) = \frac{e^{-\lambda}\ \lambda^x}{x!}$$

dove:

- λ è un qualsiasi valore positivo ($\lambda > 0$) equivalente al numero di successi che ci si aspetta che si verifichino in un dato intervallo di tempo.
- x è il numero delle occorrenze (successi) per cui si vuole prevedere la probabilità (deve essere intero non negativo)

La funzione generatrice dei momenti è pertanto:

$$g(t) = e^{\lambda(e^t - 1)}$$

Il valore atteso e la varianza coincidono

$$\mu = \sigma^2 = \lambda$$

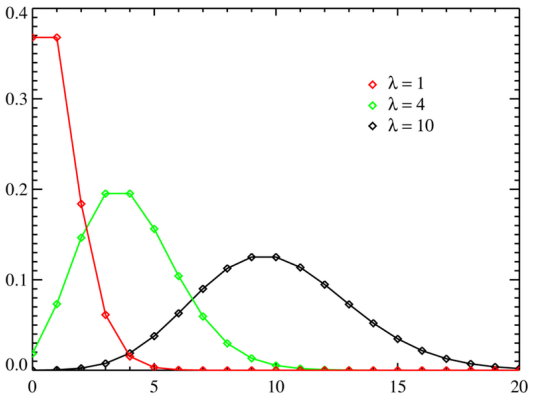

Figura 1.1: *Funzione di probabilità di una variabile casuale poissoniana*

La poissoniana è detta pure legge degli eventi rari, in quanto può essere applicata al posto della variabile casuale binomiale $B(p; n)$ quando la probabilità p di un evento è molto bassa e contemporaneamente la grandezza del campione n è molto alta, ovvero quando un evento è raro, ma il numero di eventi che si verificano ($\lambda = np$) è comunque finito.

1.3.3 Distribuzione Normale

La **variabile casuale Normale** (detta anche variabile casuale Gaussiana) è una variabile casuale continua con due parametri, indicata tradizionalmente con:

$$N(\mu; \sigma^2)$$

Si tratta di una delle più importanti variabili casuali, soprattutto continue, in quanto è, o la base di partenza per le altre v.c. (vedasi la *Chi Quadrato*, la *t di Student*, e la *F di Snedecor*) o la v.c. alla quale altre possono essere approssimate in certe situazioni limite (vedasi la *bernoulliana* e la *poissoniana* e il *teorema del limite centrale*).

La variabile casuale gaussiana, o normale, è caratterizzata dalla seguente funzione di densità di probabilità, cui spesso si fa riferimento con la dizione curva di Gauss o gaussiana:

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(t-\mu)^2}{2\sigma^2}} \qquad \text{con} \qquad \infty < t < \infty$$

L'equazione della funzione di densità è costruita in modo tale che l'area sottesa alla curva rappresenti la probabilità. Perciò, l'area totale è uguale a 1.

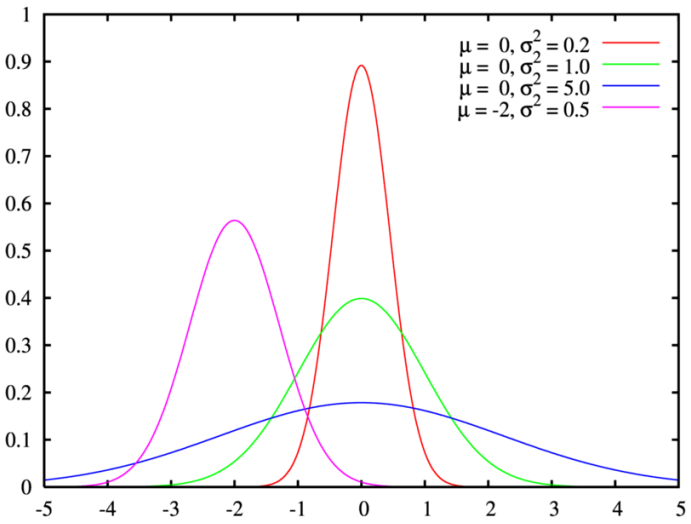

Figura 1.2: *Funzione di probabilità di una variabile casuale normale*

La sua funzione generatrice dei momenti è:

$$g(t) = e^{\mu t + \sigma^2 \frac{t^2}{2}}$$

Il valore atteso e la varianza (che sono gli unici due parametri di questa variabile casuale) sono appunto μ e σ^2.

Un piccolo appunto va fatto sulla standardizzazione. Data la distribuzione di probabilità:

$$X \sim N(\mu, \sigma)$$

inoltre definiamo la funzione di ripartizione di una $N(0,1)$ con la seguente simbologia:

$$\Phi(x) = \int_{-\infty}^{x} N(0,1)\; dx$$

definiamo standardizzazione la seguente operazione:

$$Z = \frac{x-\mu}{\sigma} \sim N(0,1)$$

La motivazione per cui si ricorre al processo di standardizzazione sta nel fatto che l'integrale di una generica distribuzione normale non è risolvibile analiticamente, tuttavia è sempre possibile ricondurre qualsiasi tipo di distribuzione normale alla $N(0,1)$, per la quale sono disponibili tavole che ne riportano la soluzione numerica dell'integrale.

1.3.4 Distribuzione Gamma

La **variabile casuale Gamma** o variabile casuale *erlanghiana* è una variabile casuale continua che viene definita da due parametri (indicati qui di seguito con a e p).

La funzione di densità di probabilità è:

$$f(x) = \frac{a^p e^{-ax} x^{p-1}}{\Gamma(p)} \qquad \text{con} \qquad (0 < x < \infty,\ a > 0,\ p > 0)$$

dove Γ è la funzione Gamma definita da:

$$\Gamma(p) = \int_0^\infty e^{-y}\ y^{p-1}\ dy$$

Integrando per parti $\Gamma(p)$ si ottiene, per $p > 1$:

$$\Gamma(p) = (p-1)\Gamma(p-1)$$

Dato che:

$$\Gamma(1) = \int_0^\infty e^{-x}\ dx = 1$$

per valori interi di p, diciamo per $p = n$, si ha:

$$\Gamma(n) = (n-1)!$$

Va notato che, se p è un intero positivo, la distribuzione Gamma di parametri (p, a) è spesso utilizzata nella pratica come variabile aleatoria del tempo di attesa per la realizzazione di n eventi. Più precisamente, se gli eventi avvengono casualmente, allora il tempo di attesa affinchè si realizzino n di questi eventi è una variabile aleatoria di tipo Gamma di parametri (n, a).

La funzione generatrice dei momenti è:

$$g(t) = \left(\frac{a}{a-t}\right)^p$$

Si ha inoltre:

- media $\mu = p/a$
- varianza $\sigma^2 = p/a^2$
- simmetria $\beta_1 = 4/p$
- curtosi $\beta_2 = 3 + 6/p$
- moda $\nu_0 = (p-1)/a$ per $p \geq 2$

Si ricava che se $p \to +\infty$, allora β_1 tende a zero e β_2 tende a 3 (come la v.c. normale), infatti per $p \to +\infty$ la v.c. Gamma tende ad una Normale $N(p/a, p/a^2)$.

Alcuni casi particolari:

- Se $a = 1/2$ e $p = g/2$ allora siamo nel caso della *Chi-Quadrato.*

- Se $p = 1$, allora siamo nel caso della variabile *casuale esponenziale* negativa

Un'altra delle caratteristiche è che se $a = 1$ e $p - 1 = k$ intero, allora la funzione di densità di probabilità diventa

$$f(x) = \frac{e^{-x}x^k}{k!}$$

che è proprio la Bayesiana della v.c.Poissoniana.

1.3.5 Distribuzione Esponenziale

La variabile casuale esponenziale negativa (o semplicemente esponenziale) è un caso particolare della variabile casuale Gamma in cui il parametro p è posto uguale a 1.

E' spesso usata per modellare il tempo tra eventi indipendenti che avvengono con una frequenza media costante, risulta dunque utile in situazioni riguardanti il tempo di attesa prima che un evento accada.

1.3.6 Distribuzione Chi-Quadro

La variabile casuale Chi Quadrato (χ^2) è un caso particolare della v.c. Gamma con $a = \frac{1}{2}$ e $p = \frac{g}{2}$ (ove g sono i gradi di libertà), quindi è una variabile casuale assolutamente continua.

Nel caso di χ^2 con un grado di libertà la funzione di densità di probabilità è:

$$f(x) = \frac{e^{-\frac{x}{2}}\ x^{-\frac{1}{2}}}{\sqrt{2\pi}} \quad \text{per} \quad x > 0$$

La χ^2 oltre ad essere un caso particolare della distribuzione Gamma, è anche ricavabile come trasformata di una variabile casuale normale. Considerando una variabile casuale normale di media nulla $N(0; \sigma^2)$, sarà: $X = N^2$.

Per generalizzare il numero dei gradi di libertà, si considerino n distribuzioni normali $N_1(0; 1); N_2(0; 1); ...N_n(0; 1)$ con media nulla e varianza unitaria indipendenti tra loro. Sarà:

$$\chi^2_n = N_1^2 + N_2^2 + ... + N_n^2$$

La funzione di densità $f(x)$ per χ^2_n (chi quadro con n gradi di libertà) sarà:

$$f(x) = \frac{e^{-\frac{x}{2}}\ x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\ \Gamma\left(\frac{n}{2}\right)}, \quad \text{per} \quad (0 < x < \infty)$$

Dove $\Gamma()$ è la funzione Gamma.

Pertanto si ottiene:

- valore atteso: $\mu = g$ (dove g sono i gradi di libertà)

- varianza: $\sigma^2 = 2g$
- simmetria: $\beta_1 = \frac{8}{g}$
- curtosi: $\beta_2 = 3 + \frac{12}{g}$
- moda: $\nu_0 = g - 2$ (per $n \geq 3$)

La variabile aleatoria χ^2 può essere assunta come misura dello scostamento tra frequenze osservate $A_1, A_2, ..., A_k$ e frequenze teoriche $np_1; np_2; ....; np_k$, associate a una distribuzione di probabilità ipotizzata per le k modalità $(p_1; p_2; ...; p_k)$.

1.3.7 Teorema del Limite Centrale

La più nota formulazione di un teorema del limite centrale è quella dovuta a Lindeberg e Lévy: si consideri una successione di variabili casuali $\{x_i\}_{j=1}^n$ indipendenti e identicamente distribuite, e in particolare tali che esistano, finiti, i loro momenti di ordine primo e secondo, e sia in particolare

$$E[x_j] = \mu < \infty \qquad \text{e} \qquad var(x_j) = \sigma^2 < \infty \qquad \forall\ j$$

. Definita allora la nuova variabile casuale:

$$S_n = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

dove $\bar{x} = \frac{\sum_{i=1}^n x_j}{n}$ è la media aritmetica degli x_j, si ha che S_n converge in distribuzione a una variabile casuale normale avente valore atteso 0 e varianza 1, ossia la distribuzione di S_n, al limite per n che tende a infinito, coincide con quella di una tale variabile casuale normale.

Un'altra formulazione del teorema dice che, qualunque sia la distribuzione della variabile casuale (Bernoulli, Binomiale,...) la rispettiva media campionaria tende ad una distribuzione normale centrata sulla vera media μ e con varianza sempre più piccola per campioni sempre più ampi.

$$\overline{X}_n \xrightarrow[n\to\infty]{} N\left(\mu, \frac{\sigma^2}{n}\right)$$

con

- $\overline{X}_n = \frac{\sum_{i=1}^n x_i}{n}$
- x_i i.i.d. (indipendenti ed identicamente distribuite)
- μ e σ^2 finite

Capitolo 2

Statistica

2.1 Introduzione

La **scienza statistica** è comunemente suddivisa in due branche principali:

- la **statistica descrittiva**, che studia i criteri di rilevazione, di classificazione e di sintesi, attraverso i suoi strumenti grafici (diagrammi a barre, a torta, istogrammi) e indici (indicatori statistici, indicatori di posizione come la media, di variazione come la varianza e la concentrazione, di correlazione, ecc.), delle informazioni relative a una popolazione oggetto di studio.
- la **statistica inferenziale**, che è fortemente legata alla teoria della probabilità ed ha come obiettivo quello di fare affermazioni, con una possibilità di errore controllata, riguardo la natura teorica (la legge probabilistica) del fenomeno che si osserva. La conoscenza di questa natura permetterà poi di fare previsioni. La statistica inferenziale si suddivide poi in altri capitoli, di cui i più importanti sono la stima puntuale, la stima intervallare e la verifica delle ipotesi.

Nella statistica descrittiva si opera dunque disponendo di tutte le manifestazioni del fenomeno d'interesse, al contrario nella statistica inferenziale non si dispone di tutte le manifestazioni del fenomeno ma soltanto di un sottoinsieme di queste, si dispone cioè di un *campione*.

2.2 Statistica Inferenziale

2.2.1 Modello Campionario

L'obiettivo di ogni procedimento statistico è quello di formulare deduzioni riguardanti le leggi che regolano i fenomeni collettivi, partendo da un insieme iniziale di dati.
Per determinare le caratteristiche fondamentali di una popolazione statistica non è sempre necessario analizzare tutta la popolazione, ma risulta sufficiente esaminare un campione statistico. Un **campione statistico** è un sottoinsieme opportunatamente scelto dall'intera popolazione.

Partendo dai dati di questo sottoinsieme, con un'alternanza di ragionamenti induttivi e deduttivi, è possibile trarre conclusioni riguardanti l'intera popolazione.
La procedura base consiste nel:

- formulare un'ipotesi con la massima precisione;
- trarre le conseguenze logiche che da essa scaturiscono;
- confrontare tali deduzioni con i dati sperimentali ottenuti dal campione;

Le ragioni per cui può essere preferibile analizzare i dati di un campione, piuttosto di quelli dell'intera popolazione, possono essere diverse:

1. la popolazione può essere molto vasta, risulta allora troppo costoso e lungo analizzare tutte le unità statistiche;
2. le misure possono essere distribuite;
3. le unità statistiche non presentano variabilità;
4. non tutti gli elementi della popolazione sono disponibili.

Per poter estendere le informazioni ottenute da un campione all'intera popolazione da cui il campione è stato estratto, è indispensabile che il campione sia rappresentativo di tutta la popolazione. Se ad esempio si effettua la rilevazione su un gruppo di persone in cui sono molto numerosi i liberi professionisti e i commercianti, mentre sono poco numerosi gli operai, certamente i risultati non saranno rappresentativi di tutti i lavoratori.
Per ottenere campioni rappresentativi si può procedere ad un **campionamento casuale**, ovvero un campionamento in cui ogni elemento della popolazione ha la stessa probabilità di tutti gli altri di essere estratto, in tal modo si può sperare che tutte le caratteristiche della popolazione siano rappresentate, con le stesse proporzioni, nel campione.
Si deve espressamente notare che campione casuale non significa campione preso a caso, ma campione formato da elementi scelti in modo equiprobabile, se questa condizione risulta soddisfatta può essere applicata la teoria delle probabilità per estendere i risultati campionari all'intera popolazione. Per ottenere un campionamento realmente casuale si può procedere nel seguente modo:

- si assegna ad ogni abitante della città un numero progressivo;
- si pongono in un'urna tutti i numeri assegnati;
- si estraggono tanti numeri quante sono le persone da intervistare;
- si intervistano le persone che corrispondono ai numeri estratti.

I risultati campionari non interessano di per se ma solo perchè consentono di trarre conclusioni generali valide per tutta la popolazione da cui il campione è stato estratto.
Questo processo si chiama **inferenza statistica**.

Il percorso dell'inferenza statistica si svolge secondo le seguenti fasi:

1. estrazione di un campione della popolazione
2. calcolo delle statistiche campionarie, cioè dei valori corrispondenti ai dati contenuti nel campione
3. stima dei parametri nella popolazione in base ai risultati forniti dal campione

In particolare può essere utile ricordare che:

- una misura che descrive una caratteristica della popolazione sarà chiamata **parametro**;
- una misura che descrive una caratteristica di un campione sarà chiamata **statistica**.

	Popolazione	*Campione*
Definizione	Insieme di tutte le unità statistiche	Sottoinsieme di unità selezionate in modo casuale dalla popolazione

Simbolismo utilizzato:

Simbologia	*Parametri*	*Statistiche*
Media	μ	$\bar{x}$
Deviazione standart	σ	S
Varianza	σ^2	S^2
Frequenza relativa	p	fr
Ampiezza	N	n
Indicazione generica	θ	T

I parametri della popolazione sono *valori numerici costanti*, mentre le statistiche campionarie sono variabili aleatorie che presentano una *distribuzione campionaria*.

Se conosciamo la *legge di distribuzione* della popolazione dalla quale è preso il campione, possiamo determinare la legge di distribuzione della statistica, detta *distribuzione campionaria*.

Definizione 2.2.1. I **dati** osservati $x^{oss} = (x_1^{oss}, \cdots, x_n^{oss}),\ n \geqslant 1$, sono riferiti a caratteristiche di interesse rilevate sulle n unità statistiche, che costituiscono il campione selezionato; in particolare, x_i^{oss}, con $i = 1, \cdots, n$, indica l'osservazione effettuata sulla i-esima unità statistica.

L'idealizzazione fondamentale su cui poggia l'inferenza statistica è che x^{oss} è una realizzazione di un **vettore casuale** (variabile casuale multivariata) $X = (X_1, \cdots, X_n)$.

La distribuzione di probabilità di X è, almeno in parte, ignota e si utilizza l'informazione ricavabile dai dati per ottenerne una ricostruzione.

Spesso si assumerà che $x^{oss} = (x_1^{oss}, \cdots, x_n^{oss})$ sia un **campione casuale semplice** (c.c.s.), ossia che il vettore $X = (X_1, \cdots, X_n)$ sia costituito da n componenti X_i, con $i = 1, \cdots, n$, *indipendenti e identicamente distribuite* (i.i.d.).

Nell'ambito del *campionamento casuale semplice* si ipotizzerà sempre (almeno a livello teorico) l'esistenza di un *modello probabilistico* capace di rappresentare adeguatamente il fenomeno che interessa analizzare. In altre parole, si assumerà che la popolazione P sia rappresentata da una variabile casuale semplice o multipla con una propria funzione di distribuzione non completamente nota. Ovviamente, se la funzione di distribuzione fosse completamente nota si tornerebbe al caso di disponibilità completa di tutte le possibili manifestazioni del fenomeno d'interesse.

Se si fa riferimento al caso univariato la situazione di riferimento è quella di una variabile casuale X con funzione di distribuzione $F(x;\theta_1,\theta_2,...,\theta_d) = F(x;\theta)$ dove $(\theta_1,\theta_2,...,\theta_d) = \theta$ è l'insieme (vettore) dei parametri caratteristici del modello definiti nello spazio parametrico Θ_d, $(\theta \in \Theta_d)$, cioè dei parametri che caratterizzano lo specifico modello, rappresentativo della specifica situazione reale, nell'ambito della famiglia di distribuzioni espressa dalla funzione $F(\,\cdot\,,\,\cdot\,)$.

Se, come avviene usualmente, si considera la funzione di massa (caso discreto) o di densità (caso continuo) di probabilità della variabile casuale X, si dirà che si sta trattando della variabile casuale semplice X con funzione di massa o di densità di probabilità $f(x;\theta_1,\theta_2,...,\theta_d) = f(x;\theta)$.

Si è detto che esiste un problema di induzione statistica quando la funzione di distribuzione $F(\,\cdot\,,\,\cdot\,)$ non è completamente nota; ovviamente, tale affermazione vale anche nei confronti della funzione $f(\,\cdot\,,\,\cdot\,)$.

In proposito si possono distinguere almeno due situazioni di mancanza di conoscenza: la prima situazione è quella caratterizzata da una conoscenza parziale della funzione $f(x;\theta_1,\theta_2,...,\theta_d) = f(x;\theta)$ nel senso che si conosce la forma analitica della funzione ma non si conosce il valore di tutti o di alcuni parametri caratteristici della funzione stessa, in questa circostanza si parla di **inferenza statistica parametrica**.

La seconda situazione è quella d'ignoranza completa: non si conosce ne il valore dei parametri ne la forma analitica della funzione di massa o di densità di probabilità; in questa circostanza si parla di **inferenza statistica non parametrica**.

Una terza situazione, intermedia rispetto alle due precedenti, è quella in cui si specificano certe componenti del modello (ad esempio si suppone che la v.c. appartenga alla famiglia esponenziale ma non si specifica la sottofamiglia: forma funzionale della funzione di massa o di densità). Se si opera in tale contesto si parla di **inferenza statistica semi-parametrica**, nel senso che il modello statistico per l'analisi del fenomeno è specificato solo parzialmente.

Nel seguito, si considererà principalmente l'**inferenza statistica parametrica**: si suppone che la distribuzione di probabilità di $X = (X_1,\cdots,X_n)$ sia nota a meno di una quantità $\theta = (\theta_1,\cdots,\theta_d)$, che è un vettore numerico non noto di dimensione $d \geqslant 1$.

θ è detto **parametro** e corrisponde tipicamente a quelli che sono gli *aspetti di interesse*, con riferimento alla popolazione da cui i dati sono tratti come campione casuale.

Definizione 2.2.2. Un **modello statistico** è una collezione di distribuzioni di probabilità per $X = (X_1,\cdots,X_n)$, che siano compatibili con i dati osservati x^{oss}.

Definizione 2.2.3. Un **modello statistico parametrico** è

- una collezione di *funzioni di probabilità congiunte*

$$f_X(x_1, \cdots, x_n; \theta) \qquad \text{con} \qquad \theta \in \Theta$$

per X indicizzate dal parametro θ, nel caso discreto;

- una collezione di *funzioni di densità di probabilità congiunte*

$$f_X(x_1, \cdots, x_n; \theta) \qquad \text{con} \qquad \theta \in \Theta$$

per X indicizzate dal parametro θ, nel caso continuo.

Definizione 2.2.4. Il supporto S_X di X è detto **spazio campionario** e rappresenta l'insieme dei possibili campioni $x = (x_1, \cdots, x_n)$, cioè delle possibili relizzazioni di $X = (X_1, \cdots, X_n)$.

Definizione 2.2.5. L'insieme $\Theta \subseteq \mathbb{R}^d$, con $d \geqslant 1$, è detto **spazio parametrico** e contiene i possibili valori per il parametro θ.

Si suppone che il modello sia correttamente specificato e che esista uno e un solo valore θ^0, detto **vero valore del parametro**.

Definizione 2.2.6. Se $X_1, X_2, ..., X_n$ costituiscono un insieme di variabili casuali indipendenti e identicamente distribuite (i.i.d.), la loro funzione di massa o di densità di probabilità congiunta soddisfa l'uguaglianza:

$$f(x_1, x_2, ..., x_n; \theta_1, \theta_2, ..., \theta_d) = f(x; \theta)$$

$$= f(x_1; \theta) \ \cdot \ f(x_2; \theta) \ \cdot \ ... \ \cdot \ f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

allora si dice che l'insieme di variabili casuali i.i.d. $X_1, X_2, ..., X_n$ costituisce un **campione casuale semplice** di n osservazioni indipendenti relativo alla variabile casuale X che ha funzione di massa o di densità di probabilità equivalente a quella (comune) di ciascuna componente X_i del campione. Il **punto campionario** $X = (X_1, X_2, ..., X_n)$ è definito nello spazio dei campioni ad n dimensioni.

Utilizzando il campione osservato x^{oss}, alla luce del modello statistico parametrico prescelto, si vogliono ricavare informazioni sul parametro ignoto θ, che individua alcuni aspetti di interesse sulla popolazione (processo di generazione dei dati sperimentali o modello probabilistico) di riferimento.

In sostanza, si vuole ricavare informazioni su θ_0, utilizzando i dati x^{oss}.

Una buona procedura statistica deve fornire buoni risultati qualsiasi sia il vero valore del parametro θ_0 e deve essere utilizzabile con riferimento ad ogni possibile campione osservato x^{oss}.

Per questo motivo, al posto di θ_0 e x_{oss}, si adotterà la scrittura θ e $x = (x_1, \cdots, x_n)$, per indicare un qualsiasi vero valore e un generico campione osservato.

Dalla definizione risulta che se, ad esempio, si volesse estrarre un campione di n elementi da una popolazione **distribuita normalmente**, con media μ e varianza σ^2, la funzione di densità di probabilità del campione casuale è

$$f(x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n; \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \cdot e^{-\left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(x_i-\mu)^2}$$

Se l'estrazione del campione di n elementi riguardasse una **popolazione poissoniana** caratterizzata dal parametro λ, la funzione di massa di probabilità del campione casuale è:

$$f(x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n; \lambda) =$$

$$= \prod_{i=1}^{n} f(x_i; \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

Alle due funzioni $f(x_1, x_2, ..., x_n; \lambda)$ e $f(x_1, x_2, ..., x_n; \mu, \sigma^2)$ sopra riportate e, in generale, ad ogni funzione di massa o di densità di probabilità campionaria

$$f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

dove θ rappresenta uno o più parametri caratteristici della distribuzione di riferimento, può essere associata una seconda interpretazione, che introduce nella trattazione un concetto di estrema rilevanza: la **funzione di verosimiglianza**.

Si tratta di una funzione del tutto equivalente, in termini formali, alla funzione di massa o di densità di probabilità campionaria sopra introdotta, ma che da questa si diversifica sostanzialmente. Infatti, la funzione

$$f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

viene detta **funzione di verosimiglianza** se la si interpreta come funzione del parametro (o dei 2 parametri) θ per un campione prefissato e non come funzione degli elementi campionari.

Per evidenziare questa particolare interpretazione si può rappresentare algebricamente la funzione di verosimiglianza con l'espressione:

$$L(\theta) = L(\theta \mid X = x) = \prod_{i=1}^{n} f(\theta \mid x_1, x_2, ..., x_n)$$

dove $X = (X_1, X_2, ..., X_n)$ rappresenta la variabile casuale ad n dimensioni (vettore casuale) associata alle n rilevazioni campionarie, mentre $x = (x_1, x_2, ..., x_n)$ rappresenta il punto campionario, cioè una specifica determinazione del vettore casuale X, definito nello spazio o universo dei campioni a n dimensioni.

Più formalmente si ha la seguente definizione.

Definizione 2.2.7. Si consideri un modello statistico parametrico per i dati $x = (x_1, \cdots, x_n)$, riferiti a un campione casuale $X = (X_1, \cdots, X_n)$ con funzione (di densità) di probabilità congiunta $f_X(x_1, \cdots, x_n; \theta)$ e $\theta \in \Theta$ parametro non noto. La funzione $L : \Theta \to \mathbb{R}^+$, nella variabile θ, definita da:

$$L(\theta) = L(\theta; x) = f_X(x_1, \cdots, x_n; \theta)$$

è detta **funzione di verosimiglianza** (likelihood) di θ basata sui dati x.

Va notato che $L(\theta)$ non è una funzione (di densità) di probabilità.

Alla luce delle osservazioni x, θ_1 è più credibile di θ_2, come indicatore del modello generatore dei dati, se :

$$L(\theta_1) > L(\theta_2) \qquad \text{cioè se} \qquad \frac{L(\theta_1)}{L(\theta_2)} > 1$$

Si operano confronti tra coppie di valori per θ e si valuta la loro adeguatezza relativa.

In pratica, nella definizione di $L(\theta)$, si possono trascurare i fattori moltiplicativi che non dipendono da θ.

Spesso conviene considerare la trasformata logaritmica di $L(\theta)$:

$$\ell(\theta) = \ell(\theta; x) = \log L(\theta) = \log f_X(x_1, \cdots, x_n; \theta),$$

chiamata **funzione di log-verosimiglianza**.

L'interpretazione è analoga, con l'unica differenza che i confronti di credibilità tra coppie di valori θ_1 e θ_2 si basano sulle differenze $\ell(\theta_1) - \ell(\theta_2)$.

Nella definizione di $\ell(\theta)$ si possono trascurare le costanti additive che non dipendono da θ.

Dato che vale la regola ($\log \prod = \sum \log$) [1], nel caso di c.c.s. si ha:

$$\ell(\theta) = \sum_{i=1}^{n} \log \ f(x_i; \theta)$$

Ricapitolando, si ha che, nella prima interpretazione, la funzione

$$f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

fa riferimento all'universo dei campioni, si tratta quindi di un riferimento a priori, cioè prima dell'effettiva estrazione del campione. In questo contesto, le variabili che interessano sono $X_1, X_2, \ldots, X_n$, associate a ciascun punto campionario.

Nella seconda interpretazione, la variabile di riferimento è il parametro, o il vettore dei parametri incognito θ, in quanto si assume l'avvenuta estrazione campionaria delle unità statistiche di osservazione e le variabili associate a ciascuna unità (punto campionario) hanno assunto una specifica determinazione, sono cioè delle costanti note, mentre assume la natura di variabile θ (parametro o vettore dei parametri) essendo tale entità un'incognita del problema.

Esempio 2.2.1. Per effettuare un controllo di qualità si analizzano n oggetti, scelti a caso tra quelli prodotti da un certo macchinario. Il campione osservato $x = (x_1, \cdots, x_n)$ sarà costituito da una sequenza di valori 0 o 1, che indicano, rispettivamente, se l'oggetto è o non è conforme agli standard di qualità.

Se le n osservazioni sono state effettuate in modo indipendente e nelle medesime condizioni, $X_1, \cdots, X_n$ costituisce un c.c.s.

[1] da "i ricordi di Aldo", edizione "NeuroniSolitari"

E' ragionevole ipotizzare che le variabili casuali X_i, con $i = 1, \cdots, n$, seguano una distribuzione bernoulliana o binomiale elementare, con funzione di probabilità

$$f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$$

se $x_i = \{0, 1\}$, e nulla altrove; $p \in (0, 1)$.

Il modello statistico parametrico è dato dalla famiglia delle funzioni di probabilità congiunte

$$\begin{aligned} f_X(x_1, \cdots, x_n; p) &= \prod_{i=1}^{n} f(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^{n} x_i} \ (1-p)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

con $p \in (0, 1)$.

In questo caso, $\theta = p$ e corrisponde alla probabilità che un singolo oggetto sia difettoso, cioè alla porzione di oggetti difettosi prodotti dal macchinario.

Inoltre, $\Theta = (0, 1)$ e $S_X = \{0, 1\} \times \cdots \times \{0, 1\} = \{0, 1\}^n$.

Per sintetizzare, si dice che $X_1, \cdots, X_n$ è un c.c.s. tratto da una popolazione bernoulliana, con parametro p ignoto, oppure che il campione $X_1, \cdots, X_n$ è costituito da copie indipendenti di una variabile casuale $Y \sim Ber(p)$, con parametro p ignoto.

$\diamond$

2.2.2 Distribuzioni Campionarie

Ogni analisi statistica inferenziale è caratterizzata da una componente di incertezza, poichè i dati x sono interpretati come realizzazione di un vettore casuale X. Se si ripete l'esperimento, nelle medesime condizioni, si ottengono dei dati x', tipicamente diversi da x.
Ogni inferenza sulla popolazione (sul parametro di interesse) va accompagnata da una valutazione sul suo grado di affidabilità/incertezza. Nell'effettuare una analisi statistica, i dati x non vengono considerati così come sono ma vengono opportunamente sintetizzati.

Definizione 2.2.8. Si definisce **distribuzione campionaria**, ogni distribuzione di probabilità che evidenzia la relazione esistente tra i possibili valori che possono essere assunti (nell'universo dei campioni) da una qualsiasi funzione

$$T(X_1, X_2, \ldots, X_n)$$

applicata agli n elementi campionari (casuali) e la distribuzione di massa o di densità di probabilità associata agli n elementi costituenti il campione stesso. Essa descrive quindi come variano le statistiche dei campioni, se campioni casuali aventi la stessa grandezza n vengono ripetutamente estratti dalla popolazione.

La distribuzione di probabilità, che è una funzione di $X = (X_1, ..., X_n)$, dipende dal parametro ignoto θ. Quindi, la distribuzione campionaria va intesa sotto θ, cioè nell'ipotesi che θ sia il vero valore del parametro, qualunque esso sia.

Le statistiche campionarie più usate in statistica inferenziale sono le seguenti:

- La **media campionaria** $\overline{X}$ definita da:

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

In particolare, si consideri la funzione, definita sugli elementi $X_1, X_2, ..., X_n$, di un campione casuale semplice con ripetizione relativo ad una certa variabile X che ha momento s-esimo (s = 1,2,3,...) pari a μ_s e varianza pari a σ^2:

$$\overline{X}_s = T_s(X_1, X_2, ..., X_s) = \frac{1}{n}\sum_{i=1}^{n} X_i^s$$

che viene usualmente detto **momento campionario di ordine s** rispetto all'origine, o momento empirico. Evidentemente tale momento, varierà al variare del campione e descriverà una variabile casuale, la cui funzione di massa o di densità di probabilità dipenderà dalla funzione di massa o di densità di probabilità delle variabili casuali $X_1, X_2, \ldots, X_n$, e qundi, dalla funzione di massa o di densità di probabilità della variabile casuale X.

È facile verificare che il valore medio di $\overline{X}_s$ è pari al momento s-esimo della variabile X, infatti:

$$E(\overline{X}_s) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^s\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^s) = E(X^s) = \mu_s$$

e quindi, per $s = 1$ si avrà:

$$E(\overline{X}_s) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X) = E(X) = \mu_1 = \mu$$

cioè il valor medio della media campionaria è uguale alla media della popolazione.

La varianza della media campionaria è data da:

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n} = \sigma_{\overline{x}}^2$$

cioè, la varianza della media campionaria è pari alla varianza della popolazione divisa per la dimensione del campione.

Se la popolazione da cui fa il campionamento è distribuita secondo una distribuzione normale di media μ e di varianza σ^2, e il campionamento consiste di n osservazioni indipendenti, allora la variabile aleatoria X è distribuita normalmente con media μ e varianza $\frac{\sigma^2}{n}$.

- La **mediana campionaria**, cioè l'elemento centrale della lista ordinata dei dati statistici. Per determinarla è necessario disporre i dati in ordine non decrescente:

- se il numero delle unità del collettivo n è dispari, la mediana della distribuzione è la modalità relativa al dato statistico che occupa il posto di mezzo nella graduatoria ordinata:

$$X_{\frac{n+1}{2}}$$

- se n è pari non esiste un posto centrale ma ne esistono due, in questo caso la mediana viene definita come la media aritmetica delle due modalità mediane:

$$X_{\frac{n}{2}} \qquad X_{\frac{n}{2}+1}$$

- La **varianza campionaria** S^2, definita da:

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

Si può verificare, nell'ipotesi di campionamento bernoulliano (campione casuale semplice con ripetizione), che $E(S^2) = \sigma^2$, cioè il valor medio della varianza campionaria è pari alla varianza della popolazione.

Nel caso in cui il c.c.s. $X_1, ..., X_n$ sia tratto da una popolazione $N(\mu, \sigma^2)$, si verifica inoltre che $\overline{X_n}$ e S^2 sono variabili casuali indipendenti ed inoltre:

$$\frac{n-1}{\sigma^2}S^2 \sim \chi^{2(n-1)}$$

2.2.3 Stima

Nell'ambito dell'inferenza statistica parametrica si possono individuare tre classi generali di procedure che affrontano i seguenti problemi inferenziali, con riferimento al parametro di interesse θ:

- la **stima puntuale**: si vuole ottenere, sulla base dell'osservazione x, una congettura puntuale su θ;
- la **stima intervallare** o **regione di confidenza**: si vuole ottenere, sulla base dell'osservazione x, un sottoinsieme di Θ in cui è plausibilmente incluso θ;
- **verifica di ipotesi**: data una congettura o un'ipotesi su θ, si vuole verificare, sulla base dell'osservazione x, se essa è accettabile (cioè in accordo con i dati x).

Stima Puntuale

Le procedure di stima puntuale assegnano, sulla base delle informazioni contenute nel campione osservato, un singolo valore al parametro ignoto della popolazione oggetto di studio.

Definizione 2.2.9. Si consideri un modello statistico parametrico per i dati $x = (x_1, ..., x_n)$, riferiti a un campione casuale $X = (X_1, ..., X_n)$. Si definisce

stimatore o *statistica* una funzione $\hat{\theta} = T(X)$ dallo spazio campionario allo spazio parametrico:

$$\hat{\theta} : S_X \to \Theta$$

T, la funzione applicata al campione X, è detta **stimatore** di θ. Il valore $\hat{\theta} = \theta(x) \in \Theta$, ossia il risultato ottenuto applicando lo stimatore al campione osservato, è detto **stima** di θ (il parametro θ è ignoto).

Dunque, uno stimatore di θ è una opportuna *statistica campionaria* utilizzata per stimare θ, mentre la stima di θ è il suo valore osservato in corrispondenza ai dati x.

Si utilizzerà la notazione sintetica $\hat{\theta}$ sia per lo stimatore che per la stima di θ, poichè il significato appropriato è chiaro dal contesto.

Usualmente, θ sarà un parametro unidimensionale, e quindi $\hat{\theta}$ una variabile casuale univariata. Spesso si utilizzerà la scrittura $\hat{\theta}_n$ per evidenziare la numerosità n del campione.

Proprietà degli stimatori

Le proprietà desiderabili per uno stimatore sono:

- **Correttezza** (assenza di deviazioni).
  Un *campione distorto* è un campione statistico in cui la probabilità di inclusione nel campione di individui appartenenti alla popolazione dipende dalle caratteristiche della popolazione oggetto di studio. Uno *stimatore distorto* è uno stimatore che per qualche ragione ha valore atteso diverso dalla quantità che stima; uno stimatore non distorto è detto *stimatore corretto.*

  Uno stimatore è corretto quando, applicando $T(X)$ a svariati campioni, la media delle stime $\hat{\theta}$ coincide con il vero (e incognito) valore θ del meccanismo generatore della probabilità:

  $$E[T(X)] = \theta$$

  Si dice che $T(X)$ è **asintoticamente corretto** per θ se:

  $$\lim_{n\to\infty} E[T(x)] = \theta \qquad \Rightarrow \qquad \lim_{n\to\infty} E[T(X) - \theta] = 0$$

- **Consistenza**.
  Uno stimatore è *consistente* se, all'aumentare dell'informazione, ossia della numerosità del campione, la sua distribuzione di probabilità si concentra in corrispondenza al valore del parametro da stimare. Più formalmente, uno stimatore si definisce consistente quando è **consistente in probabilità**, ossia:

  $$\lim_{n\to\infty} P\{|T_n(X) - \theta| < \epsilon\} = 1$$

  vale a dire che la probabilità di selezionare un campione per cui la stima $\hat{\theta}$ è vicina al vero valore ignoto θ, in misura minore di ϵ, tende a 1 per ogni ϵ piccolo a piacere, basta infatti aumentare l'ampiezza campionaria.

E' diffcile dimostrare la convergenza in probabilità di uno stimatore, più semplice è invece la dimostrazione di **consistenza in media quadratica**, che implica quella in probabilità, quindi si utilizzerà spesso questa:

$$\lim_{n\to\infty} E[(T_n(X)-\theta)^2] = 0 \qquad \Rightarrow \qquad P[|T_n(X)-\theta| < \epsilon] = 1$$

è possibile scomporre $E\left[(T_n(X)-\theta)^2\right] = \{E[T_n(X)]-\theta\}^2 + V[T_n(X)]$ il primo addendo a destra è noto come *bias*^{2} (bias è la distorsione dello stimatore) mentre il secondo rappresenta la *varianza dello stimatore*, se entrambe queste quantità $\to 0$ allora si ha convergenza in media quadratica:

$$\begin{aligned} lim_{n\to\infty}\{E[T_n(X)]-\theta\}^2 = 0 \qquad & \textit{correttezza asintotica} \\ lim_{n\to\infty} V[T_n(X)] = 0 \qquad & \textit{varianza asintotica nulla} \end{aligned}$$

quindi:

- se T_n è corretto basta calcolare $lim_{n\to\infty}V[T_n(X)]$ per dimostrare la consistenza in media quadratica
- altrimenti bisogna dimostrare che valgono le due proprietà di *correttezza asintotica* e *varianza asintotica nulla*.

- **Efficiente**:

  E' sicuramente auspicabile che una stima sia il più vicino possibile al valore vero del parametro da stimare. Quindi se su campioni ripetuti della stessa ampiezza lo stimatore t è più vicino al parametro θ rispetto allo stimatore t', lo stimatore t sarà da preferire. Poichè non è possibile dare una definizione di vicinanza in grado di stabilire se uno stimatore è meglio di un altro, si preferisce utilizzare una misura di variabilità di t rispetto a θ.

  Alla luce di questa scelta verrebbe spontaneo utilizzare come misura della variabilità la varianza o la deviazione standard come si fa di solito. Tuttavia, a meno che il parametro θ non rappresenti il valore medio della distribuzione t, la varianza non misura la variabilità attorno a θ. Questa difficoltà è superabile prendendo come misura il momento del secondo ordine rispetto a θ; tale misura sì riduce alla varianza di t quando θ è il valore atteso di t.

  Consideriamo ora due stimatori di θ, t_1 e t_2: diremo che lo stimatore t_1 è migliore dello stimatore t_2 se

  $$E(t_1-\theta)^2 \leqslant E(t_2-\theta)^2$$

  per tutti i valori di θ, e a patto che la disuguaglianza stretta valga almeno per un valore di θ.

  Senza entrare nei dettagli delle motivazioni diciamo che, affinchè il procedimento seguito abbia effettivamente senso e per evitare situazioni paradossali, è consuetudine considerare l'efficienza solo in relazione a stimatori non distorti.

  Si possono dare due definizioni di efficienza, una assoluta e una relativa:

- **Definizione assoluta**: uno stimatore è detto efficiente se è consistente, asintoticamente normale e tale che per $n \to \infty$ ha minima varianza.
- **Definizione relativa**: tra due stimatori consistenti e asintoticamente normali è più efficiente quello con minore varianza.

$$E[(t_1 - \theta)^2] \leqslant E[(t_2 - \theta)^2]$$

- **Sufficienza**.
  La *sufficienza* di una statistica (intesa come funzione di un campione di osservazioni) definisce formalmente la capacità di tale funzione di rappresentare in maniera sintetica l'informazione contenuta nel campione.

Stimatori di massima verosimiglianza

Fondamentalmente gli approcci che si possono seguire per la stima puntuale di parametri sono:

- **Metodo dell'analogia**, in base al quale, per stimare una certa quantità di popolazione (ad esempio, un parametro) si utilizza la corrispondente quantità campionaria (statistica campionaria). Ad esempio, un valore atteso si stima con una media campionaria, una varianza con una varianza campionaria (corretta), una covarianza con una covarianza campionaria, ecc.
- **Metodo dei momenti**, che si applica quando le quantità a cui si applica il metodo dell'analogia sono momenti (ad es. il valore atteso, la varianza, ecc.)
- **Metodo della massima verosimiglianza**, in cui si cerca un valore del parametro θ che massimizzi la funzione di verosimiglianza.

Col metodo della **massima verosimiglianza**, si cerca un valore $u(x)$ del parametro θ che massimizzi $L(\theta; x)$ per ogni $x \in S$. Se riusciamo a trovare tale valore, $u(X)$ è detto **stimatore di massima verosimiglianza** di θ.

Deve essere nota, in quanto deve essere utilizzata, la funzione di verosimiglianza, che esprime la plausibilità dei diversi valori del parametro rispetto alle osservazioni ottenute, e che porta ad utilizzare il principio di massima verosimiglianza, attribuendo a θ il valore che massimizza la $L(x; \theta)$ (come se fosse x costante e θ variabile).

Il valore di stima, funzione dei dati osservati, costituisce la **stima di massima verosimiglianza** di θ, mentre la statistica di cui è realizzazione è lo **stimatore di massima verosimiglianza** di θ.

Intuitivamente il metodo consiste nel cercare di trovare i valori dei parametri che possono aver prodotto, con la maggiore probabilità, i dati osservati e si articola nelle seguenti fasi:

- Si attribuiscono ai parametri i valori che possono aver prodotto, con la maggiore probabilità, i dati osservati;

- I parametri sono determinati risolvendo il problema del massimo. La soluzione di $max\ L$ è ottenuta eguagliando a 0 la derivata 1^a della $f(X, \theta)$ e risolvendo rispetto a θ, purché la derivata 2^a sia negativa.

  Essendo la funzione di verosimiglianza una produttoria e la funzione logaritmo naturale strettamente crescente, il valore massimo di $L(\theta\ ;\ x)$, se esiste, si ha allo stesso punto in cui è massima $log[L(\theta; x)]$, funzione di log-verosimiglianza, in molti casi più semplice da trattare.

Lo stimatore di massima verosimiglianza gode delle seguenti **proprietà**:

- *Invarianza*: se U è uno stimatore di massima verosimiglianza per θ, allora $V = h(U)$ è uno stimatore di massima verosimiglianza per $B = h(\theta)$.
- proprietà asintotiche di consistenza, efficienza e sufficienza
- se esiste lo stimatore a varianza minima, questo coincide con lo stimatore di massima verosimiglianza
- al tendere di $n \to \infty$, converge in distribuzione a $N\ (0, 1/\ I_n\ (\theta))$.

Esempio 2.2.2. Al fine di illustrare il metodo della massima verosimiglianza, si consideri un campione $\{x_i\}_{i=1}^n$ di variabili casuali identicamente e indipendentemente distribuite, con distribuzione normale: $x_i \sim N(\mu, \sigma^2)$.

La **funzione di verosimiglianza** associata è:

$$L(\mu, \sigma^2|\{x_i\}_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2 \right\}$$

La massimizzazione della funzione di verosimiglianza è equivalente a massimizzarne il logaritmo:

$$\begin{aligned}
L(\mu, \sigma^2|\{x_i\}_i) &= \ln L(\mu, \sigma^2|\{x_i\}_i) \\
&= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{n} \ln e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \\
&= \sum_{i=1}^{n} \left(\ln(1) - \frac{1}{2}\ln(2\pi\sigma^2)\right) + \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^{n} \ln(1) - \sum_{i=1}^{n} \frac{1}{2}\ln(2\pi\sigma^2) + \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2
\end{aligned}$$

I parametri μ e σ^2 sono determinati risolvendo il problema di massimo:

$$\{\mu, \sigma^2\} = \arg\max_{\mu,\sigma^2} L\left(\mu, \sigma^2|\{x_i\}_i\right)$$

Le condizioni del primo ordine per un massimo definiscono il seguente sistema di equazioni in μ e σ^2:

$$\frac{\partial L}{\partial \mu} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \hat{\mu}) = \frac{1}{\hat{\sigma}^2} \left(\sum_{i=1}^{n} (x_i) - n\hat{\mu} \right) = 0$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i} (x_i - \hat{\mu})^2 = 0$$

dove i segni di apice sopra i parametri denotano i loro stimatori.

Dalla prima equazione discende immediatamente lo **stimatore di massima verosimiglianza per la media**, dato che, la condizione affinchè si verifichi $\frac{\partial L}{\partial \mu} = 0$ si ha solo se uno dei due termini del prodotto è 0, ma $\frac{1}{\sigma^2}$ sarà ovviamente positivo, si dovrà quindi avere:

$$\sum_{i=1}^{n} (x_i) - n\hat{\mu} = 0 \qquad \text{da cui segue} \qquad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

cioè la **media campionaria**. La **varianza dello stimatore** è data dalla seguente espressione:

$$var(\hat{\mu}) = var\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} var(x_i) = \frac{\sigma^2}{n}$$

Sostituendo $\hat{\mu}$ nella seconda equazione, si ha lo **stimatore di massima verosimiglianza per la varianza**:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

cioè la **varianza campionaria**.

L'esempio è particolarmente calzante, perché consente di illustrare alcune proprietà degli stimatori di massima verosimiglianza. È immediato verificare la correttezza di $\hat{\mu}$:

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \mu$$

D'altra parte, $\hat{\sigma}^2$ non gode di tale proprietà. Ricordando che:

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \hat{\mu})^2 + n(\hat{\mu} - \mu)^2$$

segue che:

$$E[\hat{\sigma}^2] = \frac{1}{n} E\left(\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \right) = \frac{1}{n} E\left[\sum_{i=1}^{n} (x_i - \mu)^2 + n(\hat{\mu} - \mu)^2 \right]$$

Dunque $\hat{\sigma}^2$ non è uno stimatore corretto; un tale stimatore sarebbe dato dalla statistica:

$$\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Val la pena d'altra parte di osservare che lo stimatore di massima verosimiglianza è comunque uno stimatore asintoticamente corretto, infatti:

$$\lim_{n\to\infty} E[\hat{\sigma}^2] = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$$

In particolare, qualunque stimatore di massima verosimiglianza è asintoticamente corretto e asintoticamente normalmente distribuito.

$\diamond$

Esempio 2.2.3. Considerando una variabile aleatoria X con distribuzione di Poisson $X \sim P(\lambda)$, determinare lo stimatore di massima verosimiglianza per λ.

Come primo passo dobbiamo giungere alla forma semplificata della log-verosimiglianza:

$$\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
\ell(\lambda) &= \log\left[L(\lambda)\right] = \log\left\{\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right\} \\
&= \sum_{i=1}^{n} \log\left\{\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right\} \\
&= \sum_{i=1}^{n} \left\{\log\left(e^{-\lambda}\right) + \log\left(\lambda^{x_i}\right) - \log\left(x_i!\right)\right\} \\
&= \sum_{i=1}^{n} \left\{-\lambda \underbrace{\log\left(e\right)}_{=1} + x_i \log\left(\lambda\right) - \log\left(x_i!\right)\right\} \\
&= -n\lambda + \log\left(\lambda\right)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log\left(x_i!\right)
\end{aligned}$$

Procediamo ora all'annullamento delle derivata prima:

$$\begin{aligned}
\frac{\partial \ell(\lambda)}{\partial \lambda} &= 0 \\
-n + \frac{\sum_{i=1}^{n} x_i}{\lambda} &= 0 \\
\frac{-\lambda n + \sum_{i=1}^{n} x_i}{\lambda} &= 0
\end{aligned}$$

Tale equazione è soddisfatta quando il numeratore assume valore zero ovvero:

$$-\lambda n + \sum_{i=1}^{n} x_i = 0 \qquad \Rightarrow \qquad \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Si può dimostrare adesso la correttezza e la consistenza dello stimatore di massima verosimiglianza.

La correttezza è verificata in quanto vale quanto segue:

$$E[\lambda_{mv}] = \left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{n\cdot\lambda}{n} = \lambda$$

Anche la consistenza dello stimatore di MV è verificata:

$$V[\lambda_{mv}] = V\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n^2} V\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V(x_i) = \frac{n\cdot\lambda}{n^2} = \frac{\lambda}{n}$$

$$\Rightarrow \qquad \lim_{n\to\infty} V[\lambda_{mv}] = 0$$

$\diamond$

2.3 Regressione e Correlazione

2.3.1 Relazioni di dipendenza

Lo studio della **dipendenza statistica** tra due caratteri si occupa di accertare se esiste una relazione fra due caratteri X ed Y. Si distingue in:

- **dipendenza logica** se tra due caratteri esistono relazioni di causa ed effetto; (un esempio di dipendenza logica è la statura di un essere umano, che dipende dall'alimentazione)
- **dipendenza statistica** se tra due caratteri esistono delle regolarità nell'associazione tra le modalità dei caratteri.

Due caratteri indipendenti da un punto di vista logico, possono presentare una associazione statistica.

Definizione 2.3.1. Un carattere statistico X è **indipendente** dal carattere Y se assume la medesima *distribuzione relativa* per ciascuna modalità di Y.

Definizione 2.3.2. Un carattere statistico si definisce **dipendente** se la sua variazione è causata dalla variazione da una o più variabili, dette indipendenti.

2.3.2 Introduzione alla Regressione e alla Correlazione

Quando si prendono in considerazione congiuntamente due o più variabili quantitative, oltre alla media e alle varianza, per ognuna di esse, è possibile:

- esaminare anche il tipo e l'intensità delle relazioni che sussistono tra loro; per esempio, quando di un individuo si misurano contemporaneamente il peso e l'altezza, è possibile verificare statisticamente se queste due variabili cambino simultaneamente, valutando direzione e intensità della loro relazione;
- predire il valore di una variabile quando l'altra è nota (ad esempio come determinare in un gruppo d'individui il peso di ognuno sulla base della loro altezza).

Per rispondere a questa serie di domande, nel caso della rilevazione congiunta di due variabili, è possibile ricorrere:

- all'analisi della **regressione**;
- analisi della **correlazione**.

Queste sono da considerare tra loro concettualmente alternative, seppure fondate su principi e metodi simili. In particolare:

- si ricorre all'analisi della regressione quando dai dati campionari si vuole ricavare un modello statistico che predica i valori di una variabile Y, detta *dipendente*, individuata come effetto, a partire dai valori dell'altra variabile X, detta *indipendente* o esplicativa, individuata come causa;
- si ricorre all'analisi della correlazione quando si vuole misurare l'intensità dell'associazione tra due variabili quantitative (X_1 e X_2) che variano congiuntamente, senza che tra esse esista una relazione diretta di causa-effetto.

E' sempre importante saper distinguere tra:

- casualità o legame di causa-effetto, che richiede l'esame della regressione
- l'associazione o evoluzione temporale simile, che richiede la correlazione

2.3.3 Modello Lineare di Regressione

La regressione formalizza dunque il problema di una relazione funzionale della misurazione tra variabili, sulla base di dati campionari estratti da un'ipotetica popolazione infinita.

Originariamente Galton utilizzava il termine come sinonimo di correlazione, tuttavia oggi in statistica l'analisi della regressione è associata alla risoluzione del **modello lineare**.

La regressione lineare rappresenta un metodo di stima del valore atteso condizionato ad una variabile dipendente, Y, dati i valori di altre variabili indipendenti, $X_1, X_2, ..., X_k$, formalmente: $E[Y|X_1, X_2, ..., X_k]$.

Definizione 2.3.3. Il **modello di regressione lineare** è:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

dove:

- $i = 1, 2, ..., n$ varia tra le osservazioni
- Y_i è la *variabile dipendente*
- X_i è la *variabile indipendente* o *regressore*
- $\beta_0 + \beta_1 X$ è la *retta di regressione* o *funzione di regressione della popolazione*
- β_0 è l'*intercetta* della retta di regressione della popolazione
- β_1 è il *coefficiente angolare* della retta di regressione della popolazione
- u è l'*errore statistico*

Per studiare la relazione tra due variabili è utile il **diagramma di dispersione**, in cui si riportano i valori della variabile esplicativa X sull'asse delle ascisse e i valori della variabile dipendente Y sull'asse delle ordinate.

La scelta del modello matematico appropriato è suggerita dal modo in cui si distribuiscono i valori delle due variabili nel diagramma di dispersione (figura 2.1).

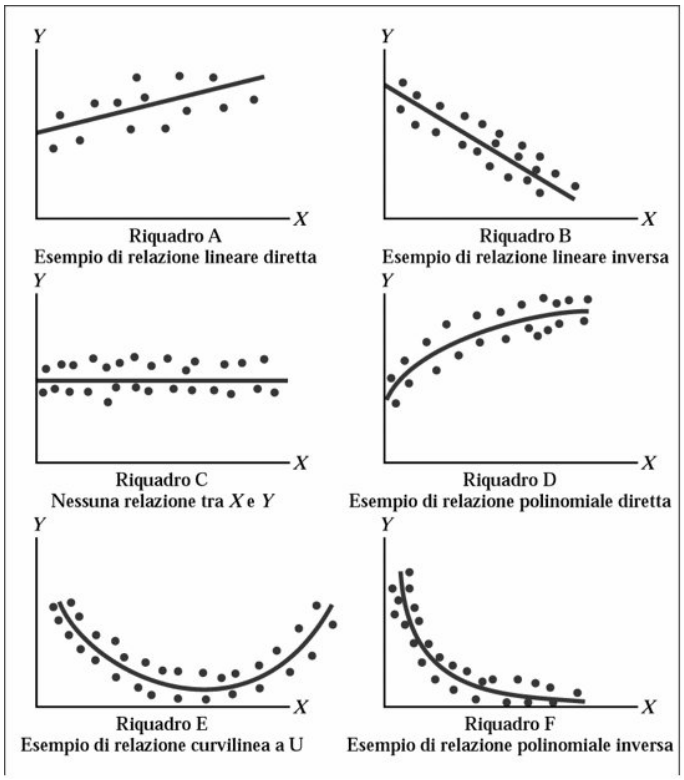

Figura 2.1: *Esempio di diagramma di dispersione*

Illustrazione del metodo

Per ogni osservazione campionaria si dispone di una determinazione Y e di k determinazioni non stocastiche $X_1, X_2, ..., X_k$.

Si cerca quindi una relazione di tipo lineare tra la variabile Y e le k variabili deterministiche.

Una prima analisi può essere condotta considerando un modello semplice a due variabili (si suppone in pratica che k sia pari a 1).

Un tipico esempio è riscontrabile dall'esperienza economica considerando la relazione tra Consumi (C) e Reddito (Y).

Ricercando una relazione funzionale in cui i consumi siano spiegati dal reddito si può ricorrere alla relazione lineare:

$C = f(Y)$ generica relazione che caratterizza i consumi;

$C = a + bY$ relazione lineare;

dove a rappresenta l'intercetta e b il coefficiente angolare della retta interpolatrice.

Stime dei parametri nel caso bivariato

Generalizzando il problema a due variabili x e y, scriveremo:

$$y_i = a + bh(x_i) + \epsilon_i$$

dove $h(x)$ è una generica funzione di x e comunemente si assume $h(x) = x$.
Ponendo, senza perdita di generalità, tale condizione la formula diviene:

$$y_i = a + bx_i + \epsilon_i$$

Quindi la variabile dipendente y viene *spiegata* attraverso una relazione lineare della variabile indipendente x (cioè: $a + bx$) e da una quantità casuale ϵ_i.

Il problema della regressione si traduce nella determinazione di a e b in modo da esprimere al 'meglio' la relazione funzionale tra y e x.

Per avvalorare di un significato statistico la scelta dei coefficienti occorre realizzare alcune ipotesi sul modello lineare di regressione:

x è una variabile deterministica

$E(\epsilon_i) = 0$

$var(\epsilon_i)$ costante per ogni i

$cov(\epsilon_i; \epsilon_j) = 0 \quad \forall i \neq j$

Date queste ipotesi si calcolano i coefficienti a e b secondo il **metodo dei minimi quadrati** (in inglese Ordinary Least Squares, o OLS, da cui il riferimento agli stimatori di seguito ottenuti come agli stimatori OLS) proposto da Gauss; detta:

$$S = S(a, b) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

le stime si ottengono risolvendo:

$$\{a, b\} = \arg\min_{a,b} S(a, b)$$

Le soluzioni si ricavano uguagliando a zero le derivate parziali di S rispetto ad a e b:

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0$$

Dove n denota il numero delle osservazioni; segue:

$$an + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

da cui si ricavano le soluzioni:

$$b = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} = \frac{S_{xy}}{S_{xx}} = \frac{cov(x, y)}{var(x)}$$

$$a = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i x_i \sum_i x_i y_i}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} = \bar{y} - b\bar{x}$$

Essendo la varianza osservata data da:

$$S_{xx} = var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

e la covarianza osservata da:

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

dove $\bar{x}, \bar{y}$ denotano le medie osservate, si possono scrivere i parametri nella forma:

$$b = \frac{S_{xy}}{S_{xx}} \qquad a = \bar{y} - b\bar{x}$$

Giustificazione probabilistica del metodo di regressione con i minimi quadrati

Si consideri il seguente problema teorico: date due variabili casuali X e Y, quale è il migliore stimatore per il valore atteso di Y, ossia quale stimatore presenta lo scarto quadratico medio (o MSE, dall'inglese Mean Squared Error) minimo?

Se si utilizza uno stimatore affine che sfrutta l'informazione relativa alla variabile casuale X, $Y = a + bX$, è possibile dimostrare che lo scarto quadratico medio $E[(Y - a - bX)^2]$ è minimizzato se:

$$b = \frac{cov(X, Y)}{var(X)} \qquad a = E[Y] - bE[X] = E[Y] - \frac{cov(X, Y)}{var(X)}E[X]$$

2.3.4 Correlazione

Per correlazione si intende una relazione tra due variabili casuali tale che a ciascun valore della prima variabile corrisponda con una certa regolarità un valore della seconda.

Non si tratta necessariamente di un rapporto di causa ed effetto ma semplicemente della tendenza di una variabile a variare in funzione di un'altra. Talvolta le variazioni di una variabile dipendono dalle variazioni dell'altra (relazione tra la statura dei padri e quella dei figlio ad esempio), talvolta sono comuni (relazioni tra la statura e il peso di un individuo); talvolta sono reciprocamente dipendenti (relazione tra prezzo e domanda di una merce: il prezzo influisce sulla domanda e la domanda influisce sul prezzo).

La correlazione si dice **diretta** o positiva quando variando una variabile in un senso anche l'altra varia nello stesso senso (alle stature alte dei padri corrispondono stature alte dei figli); si dice **indiretta** o inversa quando variando una variabile in un senso l'altra varia in senso inverso (a una maggiore produzione di grano corrisponde un prezzo minore).

La correlazione si dice **semplice** quando i fenomeni posti in relazione sono due (per esempio, numero dei matrimoni e il numero delle nascite); **doppia** quando i fenomeni sono tre (per esempio, circolazione monetaria, prezzi e risparmio); tripla quando sono quattro ecc...

Il grado di correlazione fra due variabili viene espresso mediante i cosiddetti **indici di correlazione**. Questi assumono valori compresi tra meno uno (quando le variabili considerate sono *inversamente correlate*) e l'unità (quando vi sia

correlazione assoluta cioè quando alla variazione di una variabile corrisponde una variazione rigidamente dipendente dall'altra), ovviamente un indice di correlazione pari a zero indica un'assenza di correlazione e quindi le variabili sono indipendenti l'una dall'altra.

I **coefficienti di correlazione** sono derivati dagli indici di correlazione tenendo presenti le grandezze degli scostamenti dalla media. In particolare, il **coefficiente di correlazione di Pearson** è calcolato come rapporto tra la covarianza delle due variabili ed il prodotto delle loro deviazioni standard:

$$-1 \leqslant \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y)^2}} \leqslant +1$$

Se:

- $\rho_{xy} > 0$, le variabili x e y si dicono direttamente correlate, oppure correlate positivamente;
- $\rho_{xy} = 0$, le variabili x e y si dicono indipendenti;
- $\rho_{xy} < 0$, le variabili x e y si dicono inversamente correlate, oppure correlate negativamente.

Nel caso di indipendenza lineare il coefficiente assume valore zero, mentre non vale la conclusione opposta, ovvero dal coefficiente nullo non si può desumere l'indipendenza lineare i.e. *la condizione è necessaria ma non sufficiente per l'indipendenza delle due variabili.* L'ipotesi di assenza di autocorrelazione è più restrittiva ed implica quella di indipendenza fra due variabili.

2.3.5 Analisi empirica dei residui

Il **residuo** e è uguale alla differenza tra valore osservato e il valore previsto di Y: $e_t = Y_t - \hat{Y}_t$.

Per stimare la capacità di adattamento ai dati della retta di regressione è opportuna un'analisi grafica (grafico di dispersione dei residui) e dei valori di X (ascisse).

Se si evidenzia una relazione particolare il modello non è adeguato.

Ad esempio, in figura 2.2 è mostrato un modello di regressione lineare non appropriato. Il grafico a destra evidenzia lo scarso adattamento ai dati del modello. Quindi il modello polinomiale è più appropriato.

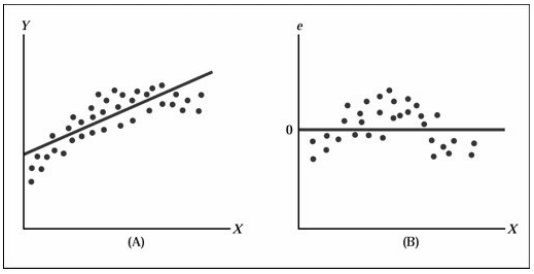

Figura 2.2: *Regressione lineare non appropriata*

Il grafico dei residui rispetto ad X consente di stabilire anche se la variabilità degli errori varia a seconda dei valori di X.

Ad esempio in figura 2.3 è mostrato un grafico in cui si evidenzia che la variabilità dei residui aumenta all'aumentare dei valori di X.

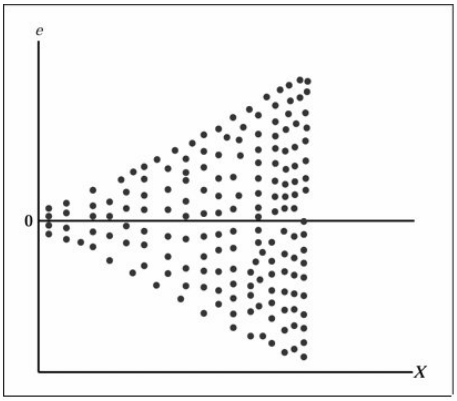

Figura 2.3: *Variabilità dei residui*