Spazi Campionari ed Eventi

Def: Esperimenti $\leftrightarrow$ Eventi

Chiameremo esperimenti i fenomeni aleatori di cui ci occupiamo e eventi i risultati di tali esperimenti

Def: Dato un esperimento, è detto spazio campionario l'insieme di tutti gli esiti possibili

Ω = generico spazio campionario

ω = generici esiti

Lancio di una moneta : $\Omega = \{T, C\}$

dado $\Omega = \{1, 2, \dots 6\}$

Risultati esperimento $\leftrightarrow$ insieme

esiti dell'esperimento $\leftrightarrow$ elementi dell'insieme

Def: Gli eventi sono sottoinsiemi di Ω

$\Omega = \{1 \dots 6\} \rightarrow \Omega' = \{3, 5\}$ è un evento

Ω è l'evento certo

$\emptyset$ è l'evento impossibile

Def: Due eventi si dicono esclusivi (mutuamente) se

$$E \cap F = \emptyset \quad E \subseteq \Omega, \ F \subseteq \Omega$$

Corrispondenze

Ω : spazio campionario certo

ω : evento elementare / esito

E : si verifica uno degli esiti contenuti in E / si verifica l'evento E

$\bar{E}$: non si verifica uno degli esiti contenuti in E / non si verifica l'evento E

$E \cap F$ si verificano entrambi gli eventi E ed F

$E \cup F$ si verifica almeno uno degli eventi E ed F

$E \setminus F$ si verifica E e non F

$E \subseteq F$ se si verifica E allora si verifica F ($E \Rightarrow F$)

$E \cap F = \emptyset$ eventi mutuamente esclusivi
Se si verifica uno dei due, è impossibile che si verifichino entrambi

$\emptyset$ evento impossibile

Ω insieme degli esiti

$\mathcal{F}$ insieme dei sottoinsiemi di Ω (insieme degli eventi)

Requisiti per $\mathcal{F}$

1) $\Omega \in \mathcal{F}$
2) Se $E \in \mathcal{F} \Rightarrow \bar{E} \in \mathcal{F}$
3) Se $E, F \in \mathcal{F} \Rightarrow E \cup F \in \mathcal{F}$

} Algebra

Dimostriamo che se $E, F \in \mathcal{F} \Rightarrow E \cap F \in \mathcal{F}$

- $E, F \in \mathcal{F} \Rightarrow \bar{E}, \bar{F} \in \mathcal{F} \Rightarrow \bar{E} \cup \bar{F} \in \mathcal{F} \Rightarrow$
$\Rightarrow \overline{\bar{E} \cup \bar{F}} \in \mathcal{F}$

Grazie alle leggi di De Morgan sappiamo che
$\overline{\bar{E} \cup \bar{F}} = E \cap F$

NB: Leggi di De Morgan

- $\overline{E \cup F} = \bar{E} \cap \bar{F}$
- $\overline{E \cap F} = \bar{E} \cup \bar{F}$

Esempio

T_n = all' n-esimo lancio esce testa
C_n = all' n-esimo lancio esce croce

E_n dice testa per la 1ª volta all'n-esimo lancio

Segue che $E_n = C_1 \cap C_2 \dots \cap T_n$

Requisiti per $\mathcal{F}$

1) =

2) =

3) Se $E_n \in \mathcal{F} \ \forall n \Rightarrow \bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$

σ-algebra

$\mathcal{F} \subseteq \mathcal{P}(\Omega)$

Esempio: Dato Ω la più piccola σ-algebra è $\{\emptyset, \Omega\}$ e la più grande è $\mathcal{P}(\Omega)$

Definizione

Dato un insieme Ω e dati di cui sono sottoinsiemi $E_1, E_2 \dots E_n$ si dice σ-algebra generata dagli E_i la più piccola σ-algebra di cui fanno parte tutti gli E_i

$$Prob = \frac{\#\text{Casi favorevoli}}{\#\text{Casi possibili}}$$ Definizione classica (a priori)

$$Prob = \frac{\#\text{successi}}{\#\text{prove}}$$ Definizione frequentista (a posteriori)

Definizione Kolmogorov

Sia Ω uno spazio campionario e $\mathcal{F}$ una σ-algebra

Assioma 1: Per ogni evento $E \in \mathcal{F}$ esiste

$$P[E] \in \mathbb{R}, \quad 0 \leq P[E] \leq 1$$

Assioma 2: Per l'intero spazio campionario

$$P[\Omega] = 1$$

Assioma 3: Presa una qualsiasi successione di eventi tali che $E_i \cap E_j = \emptyset \ (i \neq j)$

$$P[\bigcup_i E_i] = \sum_i P[E_i]$$

La probabilità è una corrispondenza

$P: \mathcal{F} \to [0,1]$ che soddisfa gli assiomi precedenti

$(\Omega, \mathcal{F}, P)$ è detto <u>spazio delle probabilità</u>

<u>Proposizione 1</u>: $P[E]$, quanto vale $P[\bar{E}]$?

$1 - P[E]$

Dim: $P[\Omega] = 1$, $\Omega = E \cup \bar{E} \to P[E \cup \bar{E}] = 1$

$P[E \cup \bar{E}] = P[E] + P[\bar{E}]$

$P[\bar{E}] = 1 - P[E]$

NB

Ω: spazio degli esiti $\to \omega$: esito elementare

E: evento ovvero sottoinsieme di Ω

$\mathcal{F}$: famiglia degli eventi $\to \sigma$-algebra (con le 3 proprietà)

<u>Proposizione 2</u>: Se $F \subseteq E \to P[F] \leq P[E]$

Dim: Se $F \subseteq E \to E = F \cup (E \setminus F)$

$P[E] = P[F \cup (E \setminus F)] = P[F] + P[E \setminus F] \geq P[F]$

<u>Proposizione 3</u>: $P[E \cup F] = P[E] + P[F] - P[E \cap F]$

Dim: $P[E] = P[(E \setminus F)] + P[E \cap F]$

$P[E \setminus F] = P[E] - P[E \cap F]$

$E \cup F = (E \setminus F) \cup (E \cap F) \cup (F \setminus E)$

$P[E \cup F] = P[E \setminus F] + P[E \cap F] + P[F \setminus E] =$

$= P[E] - P[E \cap F] + P[E \cap F] + P[F] - P[E \cap F]$

$= P[E] + P[F] - P[E \cap F]$

Proposizione 4

$$P\left[\bigcup_{i=1}^{\infty} E_i\right] = \sum_i P[E_i] - \sum_{i<j} P[E_i \cap E_j] + (-1)^{n+1} \cdot P[E_1 \cap E_2 \dots \cap E_n]$$

Probabilità condizionata

Def: Siano E e F due eventi. Si chiama probabilità di E dato F, e scriveremo $P[E|F]$ la probabilità che si verifichi l'evento E supponendo che si sia verificato l'evento F

$$P[E|F] = \frac{P[E \cap F]}{P[F]} \qquad P[F] > 0$$

$\downarrow$

$$P[E \cap F] = P[F] \cdot P[E|F] \qquad P[F|E] = \frac{P[E \cap F]}{P[E]}$$

$$P[E \cap F] = P[E] \cdot P[F|E]$$

Teorema (Regola della catena)

Dati n eventi $E_1, E_2 \dots E_n$ la cui intersezione ha probabilità positiva ($P[E_1 \cap E_2 \dots \cap E_n] > 0$), si ha

$$P[E_1 \cap E_2 \dots \cap E_n] = \cancel{P[E_1]} \cdot$$

$$= P[E_1] \cdot P[E_2|E_1] \cdot P[E_3|E_1 \cap E_2] \cdot P[E_4|E_3 \cap E_2 \cap E_1] \dots$$

(fino a $\cdot P[E_n|E_{n-1} \cap E_{n-2} \dots \cap E_2 \cap E_1]$)

Def: Si dice **partizione di Ω** (insieme degli eventi) un insieme di eventi E_i (finito o numerabile) con le seguenti caratteristiche:

1) Ogni E_i ha probabilità diversa da $0 \rightarrow 0 < P[E_i] \leq$
2) Gli E_i sono mutuamente esclusivi
$$P[E_i \cap E_j] = 0 \qquad i \neq j$$
3) L'unione di tutti gli E_i copre Ω
$$\bigcup_{i=1}^{n} E_i = \Omega$$

Teorema (Formula delle probabilità totali oppure legge delle alternative)

Sia $\{E_i\}$ una partizione di Ω, sia F un evento qualsiasi contenuto in Ω.
Allora la
$$P[F] = \sum_{i=1}^{n} P[F|E_i]\, P[E_i]$$

Dim:
$$P[F \cap E_i] = P[F|E_i] \cdot P[E_i]$$
$$F = F \cap \Omega = F \cap \left(\bigcup_i E_i\right) = \bigcup_i (F \cap E_i)$$
$$P[F] = P\left[\bigcup_i (F \cap E_i)\right] = \sum_i P[F \cap E_i]$$
$$P[F] = \sum_i P[F|E_i]\, P[E_i]$$

Esempio

Le lampadine di una certa marca vengono prodotte in 3 stabilimenti distinti $\rightarrow$ A, B e C
A $\rightarrow$ 50% di produzione, B $\rightarrow$ 30% C $\rightarrow$ 20%

L'1% delle lampadine prodotte da A sono difettose mentre le lampadine difettose prodotte da B e da C sono rispettivamente il 5% e il 10%

In fine tutte le lampadine vengono raggruppate in un solo stabilimento

Prendendo a caso una lampadina qual'è la probabilità che sia difettosa?

<u>Soluzione</u>

$P[A] = 50\%$ $\quad P[B] = 30\%$ $\quad P[C] = 20\%$

F lampadine difettose

$P[F|A] = 1\%$ $\quad P[F|B] = 5\%$ $\quad P[F|C] = 10\%$

$P[F] = ?$

$$P[F] = P[F|A] \cdot P[A] + P[F|B]\, P[B] + P[F|C] \cdot P[C]$$

$$= 0{,}375\ \%$$

<u>Teorema</u> (Legge condizionata o delle alternative)

Sia $\{E_i\}$ una partizione di Ω e siano G ed F due eventi qualsiasi, segue

$$P[F|G] = \sum_i P[F|E_i \cap G] \cdot P[E_i|G]$$

dove la sommatoria è estesa unicamente ai casi $P[E_i \cap G] \neq \emptyset$

<u>Teorema</u> (Formula di Bayes)

Dato un evento F e data una partizione $\{E_i\}$ di Ω, si ha

$$P[E_h|F] = \frac{P[F|E_h] \cdot P[E_h]}{\sum_i P[F|E_i] \cdot P[E_i]}$$

Dim: $P[E_h \cap F] = P[F \mid E_h]\, P[E_h] =$

$= P[E_h \mid F] \cdot P[F]$

$$P[E_h \mid F] = \frac{P[F \mid E_h] \cdot P[E_h]}{P[F]} =$$

$$= \frac{P[F \mid E_h] \cdot P[E_h]}{\sum_i P[F \mid E_i] \cdot P[E_i]}$$

Per il teorema delle prob. totali

Eventi indipendenti

Osservazione: Se $P[E \mid F] > P[E]$ l'evento F è favorevole all'evento E

Se $P[E \mid F] < P[E]$ l'evento F è sfavorevole all'evento E.

Se $P[E \mid F] = P[E]$ l'evento F è ininfluente all'evento E o indipendente da F

Proprietà: Se E è indipendente da F allora F è indipendente da E

$$P[E \mid F] = P[E] \Longrightarrow P[F \mid E] = P[F]$$

Dim

$P[E \cap F] = P[E \mid F]\, P[F] = P[E] \cdot P[F]$

$P[F \cap E] = P[F \mid E]\, P[E] = P[F] \cdot P[E]$

* $\downarrow$

$\cancel{P[F \mid E] \cdot P[E] =}$ $P[F \mid E] = P[F]$

$P[E \mid F] = P[E]$

Conseguenza $\rightarrow$ $P[E \cap F] = P[E] \cdot P[F]$

* Sappiamo $P[E \cap F] = P[E] \cdot P[F]$

Sappiamo che $P[F \cap E] = P[F \mid E] \cdot P[E]$ $\rightarrow$ $P[F \mid E] \cdot \cancel{P[E]} = \cancel{P[E]}\, P[F]$

✓

Proposizione

Se E ed F sono indipendenti allora sono indipendenti anche E ed $\bar{F}$

Dim

$$E = (E \cap F) \cup (E \cap \bar{F})$$

$$P[E] = P[E \cap F] + P[E \cap \bar{F}] = P[E] \cdot P[F] + P[E \cap \bar{F}]$$

$$P[E \cap \bar{F}] = P[E] - P[E] \cdot P[F] =$$

$$= (1 - P[F])\, P[E] = P[E] \cdot P[\bar{F}]$$

Osservazione

$$P[E \cap F \cap G] = P[E]\, P[F] \cdot P[G]$$

eventi indipendenti

Definizione

Una famiglia di n eventi $E_1, E_2 \dots E_n$ si dice famiglia di eventi indipendenti se gli eventi $\{E_i\}$ sono a 2 a 2, a 3 a 3 ... e così via fino a n a n indipendenti

evento A: passaggio di corrente nel circuito

$P[A] = ?$ è in funzione degli eventi interruttori

$A = E_1 \cup E_2 \qquad \bar{A} = \bar{E}_1 \cap \bar{E}_2$

$P[A] = 1 - P[\bar{A}] = 1 - P[\bar{E}_1] \cdot P[\bar{E}_2]$

p_1: prob pos corrente in 1 $\qquad p_2$: in 2

$$P[A] = 1 - (1 - p_1)(1 - p_2) = p_1 + p_2 - p_1 p_2$$

$p_1 \quad p_2$

$$P[A] = P[E_1 \cap E_2] = P[E_1]\,P[E_2] = p_1\,p_2$$

Calcolo combinatorio

Def: Si dice che n oggetti sono ordinati o disposti in un allineamento quando sono collocati in n posti numerati da 1 a n

Esempio: $O_1, O_2 \dots O_n$

Permutazione

Si dice permutazione di n oggetti distinti ogni allineamento degli oggetti stessi; due permutazioni sono distinte quando differiscono per il posto occupato da almeno un oggetto

Indichiamo con P_n le permutazioni (il numero)

Teorema $\quad P_n = n!$ $\quad (n! = 1 \cdot 2 \cdot 3 \dots (n-1)n)$

Permutazioni con oggetti non necessariamente distinti

Teorema

Il numero delle permutazioni distinte di n oggetti dei quali q_1 uguali fra loro, q_2 uguali fra loro e distinti dai precedenti ... q_h uguali fra loro e distinti dai precedenti

$(q_1 + q_2 \dots + q_h = n)$ è

$$P^*_{q_1, q_2 \dots q_h} = \frac{n!}{q_1!\, q_2! \dots q_h!}$$

Oss: se gli n oggetti sono solo di 2 tipi $(k, n-k)$

$$P^*_{k, n-k} = \frac{n!}{k!\,(n-k)!} = \binom{n}{k}$$

Disposizione semplice

Si dice disposizione semplice di n oggetti di classe k ogni allineamento di k oggetti scelti fra gli n

$D_{n,k}$ disposizione di k oggetti scelti tra gli n

Teorema

$D_{n,k}$ è il prodotto di k interi consecutivi ~~positivi~~ il cui massimo è n

$$D_{n,k} = n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!}$$

Def: si dice disposizione con ripetizione di n oggetti di classe k ogni allineamento di k oggetti scelti fra gli n con la convenzione che ogni oggetto può essere ripetuto 1 o più volte

Si dimostra $D^*_{n,k} = n^k$

Combinazioni semplici ~~di~~ con ripetizione

Si dice combinazione semplice, o più brevemente combinazione, di n oggetti di classe k, ogni raggruppamento di k oggetti comunque scelti tra gli n:

Teorema

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

Si dice combinazione con ripetizione di n oggetti di classe k ogni raggruppamento di k oggetti comunque scelti fra gli n omogenei con la convenzione che un oggetto può essere ripetuto più volte

Teorema

$$C^*_{n,k} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Variabili Aleatorie

Consideriamo lo spazio di probabilità $(\Omega, \mathcal{F}, P)$

$$X : \Omega \to \mathbb{R} \qquad X(\omega) = x \qquad x \in \mathbb{R}$$

La variabile aleatoria è dunque una corrispondenza.

Dato uno spazio di probabilità $(\Omega, \mathcal{F}, P)$ si dice variabile aleatoria una corrispondenza tra gli elementi di Ω e i reali, tale corrispondenza deve soddisfare una condizione che ne definisce l'insieme

$$A_t = \{ \omega \mid X(\omega) \leq t \} \in \mathcal{F} \quad \forall t \in \mathbb{R}$$

È detto supporto l'insieme dei valori che la variabile aleatoria può assumere

Definizione Funzione di ripartizione

Data la variabile aleatoria X, è detta funzione di ripartizione di X la funzione

$$F_X(t) = P[X \leq t] = P[A_t]$$

$$F_X(t) : \mathbb{R} \to [0, 1]$$

Proprietà

1) F_X è monotona non decrescente, cioè

$$F_X(t_1) \leq F_X(t_2) \quad \text{se} \quad t_1 \leq t_2$$

infatti $A_{t_1} \subseteq A_{t_2} \Rightarrow P[A_{t_1}] \leq P[A_{t_2}]$

2) $\lim_{t \to -\infty} F_X(t) = 0 \qquad \lim_{t \to +\infty} F_X(t) = 1$

3) Per ogni valore $x_0 \in \mathbb{R}$, $P[X = x_0] = \lim_{t \to x_0^+} F_X(t) - \lim_{t \to x_0^-} F_X(t)$

4) La funzione di ripartizione F_X è continua a destra ossia:

per ogni valore $x_0 \in \mathbb{R}$

$$F_X(x_0) \left(= P[X \leq x_0]\right) = \lim_{t \to x_0^+} F_X(t)$$

Definizioni

- Se la variabile aleatoria X può assumere solo un numero finito (o al più numerabile) di valori, la relativa F_X sarà costante a tratti e la variabile viene detta discreta.

- Se X può assumere i valori di un intervallo $I \in \mathbb{R}$ allora F_X è generalmente continua e la variabile è detta continua.

- Densità di variabili aleatorie

1) Caso discreto

Una variabile aleatoria X discreta può assumere solo alcuni valori x_i

$$P[X = x_i] \rightarrow \sum_i P[X = x_i] = 1$$

<u>Densità di probabilità</u> (per il caso discreto)

La funzione $p_X : \mathbb{R} \rightarrow [0, 1]$ definita dalla relazione

$$p_X(x) = P[X = x] = \begin{cases} P[X = x_i] & x = x_i \text{ per un certo } i \\ 0 & x \neq x_i \ \forall i \end{cases}$$

Calcolare $P[a \leq X \leq b] = ?$

$$P[a < X \leq b] = \sum_i P[X = x_i]$$

$A = \{x_i \mid a < x_i \leq b\} \longrightarrow$

$$P[a < X \leq b] = \sum_{x_i \in A} P[X = x_i] =$$

$$= \sum_{x_i \in A} p_X(x_i)$$

Oss: $P[a < X \leq b] = P[a < X < b] + P[X = b]$

$P[a < X \leq b] \neq P[a < X < b]$

$F_X(t) = P[X \leq t] = \sum_{x_i \leq t} p_X(x_i)$

Quindi in generale

$P[X \leq x_i] = P[X < x_i] + P[X = x_i]$

Esempio

Sia X la variabile aleatoria "numero di occasioni in cui è uscita testa dopo 2 lanci di moneta"

$\Omega = \{(T,T), (T,C), (C,T), (C,C)\}$

$X : \Omega \to \cancel{[0,1]}\ \mathbb{R}$

$X((T,T)) = 2 \leftarrow$ È uscita testa 2 volte

$X((T,C)) = X((C,T)) = 1 \leftarrow$ 1 volta

$X((C,C)) = 0 \leftarrow$ 0 volte

$P[X=2] = \frac{1}{4}$ $\quad P[X=1] = \frac{1}{2}$ $\quad P[X=0] = \frac{1}{4}$

$P[-2 < X < 0,5] = P[X=0] = \frac{1}{4}$

$P[-2 < X \leq 0,5] = P[-2 < X < 0,5]$

$P[0 < X \leq 2] = P[X=2] + P[X=1] = \frac{1}{4} + \frac{1}{2}$

$P[0 < X < 2] = P[X=1]$

$F_X(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{4} & 0 \leq t < 1 \\ \frac{3}{4} & 1 \leq t < 2 \\ 1 & t \geq 2 \end{cases}$

$(= P[X \leq t])$

0,25

Caso continuo

Per definire la probabilità consideriamo la funzione di ripartizione

$$F_X(a) = P[X \leq a]$$

$$F_X(t) = \int_{-\infty}^{t} f_X(x)\,dx$$

La funzione $f_X(x)$ è detta funzione densità

- $P[a \leq X \leq b] = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx$

Dim

$$I_a = (-\infty, a) \qquad I_b = (-\infty, b) \qquad I_{ab} = [a, b]$$

$$P[X \in I_b] = P[I_b] = P[I_a \cup I_{ab}] =$$

$$= P[I_a] + P[I_{ab}]$$

Segue $P[I_{ab}] = P[I_b] - P[I_a] =$

$$= P[X \leq b] - P[X \leq a] = F_X(b) - F_X(a) =$$

$$= \int_{-\infty}^{b} f_X(x)\,dx - \int_{-\infty}^{a} f_X(x)\,dx = \int_a^b f_X(x)\,dx$$

Def

Sia X una variabile aleatoria continua, se esiste una funzione f_X tale per cui, $\forall a, b \in \mathbb{R}$

$$P[a < X \leq b] = \int_a^b f(x)\,dx$$

allora $f(x)$ è detta funzione densità di probabilità

Oss:

1) $f_X(x)$ è in teoria definita nel supporto, Estendiamo

$$f_X(x) = 0 \quad \text{se} \quad x \notin I \qquad (I \text{ supporto di } X)$$

2) $P[a < x \leq b] = P[a < x < b]$

Proprietà

Se f_X è la funzione densità di una variabile aleatoria continua, allora

1) $f_X(x) \geq 0 \quad \forall x \in \mathbb{R}$ e $f_X(x) > 0$ se $x \in I$

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

ESEMPIO

Supponiamo sia possibile estrarre un numero reale a caso nell'intervallo $[1,3]$ in modo che nessuna zona dell'intervallo sia privilegiata.

Sia Y la variabile aleatoria "numero estratto"

Definire la densità di probabilità

$I = [1,3]$

$$f_Y(x) = \begin{cases} 0 & \text{se } x < 1 \\ c & \text{se } 1 \leq x \leq 3 \\ 0 & \text{se } x > 3 \end{cases}$$

Per calcolare c ricordare che

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

$$\int_1^3 f_Y(x)\,dx = \int_1^3 c\,dx = 2c = 1 \Rightarrow c = \frac{1}{2}$$

- $F_X(t) = P[X \leq t] = \int_{-\infty}^{t} f_X(x)\,dx$

$$f_Y(t) = \begin{cases} 0 & \text{se } x < 1 \\ \frac{1}{2} & \text{se } 1 \leq x \leq 3 \\ 0 & \text{se } x > 3 \end{cases}$$

$t < 1$

$$F_Y(1) = \int_{-\infty}^{1} f_Y(x)\,dx = \int_{-\infty}^{1} f_Y(x)\,dx = 0$$

$$F_Y(t) = \int_{-\infty}^{t} f_Y(x)\,dx = \int_1^t f_Y(x)\,dx = \qquad 1 \le t \le 3$$

$$= \frac{1}{2}\int_1^t dx = \frac{1}{2}(t-1)$$

$$t > 3$$

$$F_Y(3) = \int_{-\infty}^{3} f_Y(x)\,dx = \int_1^3 f_Y(x)\,dx = \frac{1}{2}\int_1^3 dx = 1$$

$$F_Y(t) = \begin{cases} 0 & t < 1 \\ \frac{1}{2}(t-1) & 1 \le t \le 3 \\ 1 & t > 3 \end{cases}$$

NB

$F_X(t) = P[X \le t]$ sia in discreto che continuo

Discreto

$$F_X(t) = \sum_{x \le t} p_X(x)$$

Continuo

$$F_X(t) = \int_{-\infty}^{t} f_X(x)\,dx$$

- $F_X(t)$ definita $\forall t \in \mathbb{R}$
- $0 \le F_X(t) \le 1$
- F_X è monotona non decrescente
- $\lim_{t \to -\infty} F_X(t) = 0 \qquad \lim_{t \to +\infty} F_X(t) = 1$

Come calcolare $f_X(t)$?

$$F_X(t) = \int_{-\infty}^{t} f_X(x)\,dx \rightarrow f_X(t) = \frac{dF_X}{dt}$$

- Se X è discreta

$$P[a < X \le b] = F_X(b) - F_X(a)$$

$$P[X > a] = 1 - P[X \le a] = 1 - F_X(a)$$

- Se X è continua ... ?

Variabile aleatoria funzione di un'altra variabile aleatoria

- Caso discreto

$P[X=-1] = 0,2$, $P[X=0] = 0,3$, $P[X=1] = 0,5$
$P[X=2] = 0,1$

Determinare la densità della variabile aleatoria $Y=X^2$

Supporto $X = \{-1, 0, 1, 2\} \rightarrow$ Supporto $Y = \{0, 1, 4\}$

$P[Y=0] = P[X=0] = 0,3$

$P[Y=1] = P[X=-1] + P[X=1] = 0,6$

$P[Y=4] = P[X=2] = 0,1$

- Caso continuo

Teorema: Sia X variabile aleatoria continua e sia $Y = g(X)$

Se g è una funzione invertibile nel supporto I di X e h è la sua inversa allora:

$$f_Y(y) = f_X(h(y)) \cdot |h'(y)|$$

Dim:

$$\int_a^b f_X(x)\,dx = \int_{g(a)}^{g(b)} f_X(h(y)) \left|\frac{dx}{dy}\right| dy \qquad \boxed{y = g(x) \quad x = g^{-1}(y) = h(y)}$$

$$= \int_{g(a)}^{g(b)} f_X(h(y)) \, |h'(y)| \, dy$$

$y = g(x) \qquad x = g^{-1}(y) = h(y)$

Dim:

$$F_Y(t) = P[Y \le t] = P[g(X) \le t] =$$

$$= \int_{g(X) \le t} f_X(x)\,dx = \int_{y \le t} f_X(h(y)) \left|\frac{dx}{dy}\right| dy =$$

$$= \int_{y \le t} f_X(h(y))\,|h'(y)|\,dy$$

Esempio

Sia X var aleatoria avente densità

$$f_X(x) = \begin{cases} 3/4\,(1-x^2) & -1 \le x \le 1 \\ 0 & \text{se no} \end{cases}$$

Determinare la densità di $Y = e^X$

$$[-1, 1] \xrightarrow{Y=e^X} [e^{-1}, e]$$

$$y = e^x \rightarrow x = \ln y$$

$$f_Y(y) = f_X(\ln y) \cdot |(\ln y)'| =$$

$$= \frac{3}{4}(1-(\ln y)^2)\left|\frac{1}{y}\right| \qquad -1 \le x \le 1$$

1) $f_X(x) = \begin{cases} 2(x-1) & 1 \le x \le 2 \\ 0 & \text{se no} \end{cases}$

$y = x^2$

2) $f_X(x) = \begin{cases} \frac{x^2}{18} & -3 \le x \le +3 \\ 0 & \text{se no} \end{cases}$

Definizione Valore atteso

Data var aleatoria X, data la funzione di ripartizione il suo valore atteso oppure valore medio o speranza matematica, se esiste, è dato da

$$E[X] = \int_0^{+\infty} [1 - F_X(t)]\,dt - \int_{-\infty}^{0} F_X(t)\,dt$$

Si dimostra che

$$E[X] = \sum_i x_i \, P[X = x_i]$$ Caso discreto

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)\,dx$$ Caso continuo

Definizione Moda

Sia X var aleatoria, discreta o continua, si chiama moda di X il valore, se esiste, per cui si raggiunge la densità (nella funz di densità)

Definizione Quantile

Dato $\alpha \in (0,1)$, si dice quantile di ordine α della variabile aleatoria X il più piccolo numero x_α tale che

$$P[X < x_\alpha] \leq \alpha \leq P[X \leq x_\alpha]$$

Definizione mediana

Si dice mediana della variabile aleatoria X il quantile di ordine $0{,}5$ $(x_{0,5})$

Definizione Varianza

Data una var aleatoria X il cui valore atteso vale $E[X] = m$, si chiama varianza di X, e si indica con $Var[X]$ il seguente valore atteso

$$Var[X] = E[(X-m)^2]$$

Proprietà

$$Var[X] = E[X^2] - m^2$$

$m = E[X]$

Dim

$$Var[X] = E[(X-m)^2] =$$
$$= E[X^2 + m^2 - 2mX] =$$
$$= E[X^2] + E[m^2] - E[2mX] =$$
$$= E[X^2] + m^2 - 2m\,E[X] =$$
$$= E[X^2] + m^2 - 2m^2 = E[X^2] - m^2$$

* m è una costante

Disuguaglianza di Markov

Teorema Sia X una var aleatoria e $g: \mathbb{R} \to \mathbb{R}$ tale che $g(x) \geq 0 \quad \forall x \in \mathbb{R}$

Allora, se esiste $E[g(x)]$ si ha:

$$P[g(X) \geq k] \leq \frac{E[g(x)]}{k} \quad \forall k > 0$$

Dim

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)\,dx \quad \text{e} \quad E[g(x)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\,dx$$

$$E[g(x)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\,dx \geq \int_{\{x \mid g(x) \geq k\}} g(x) f_X(x)\,dx \geq$$

$$\geq \int_{\{x \mid g(x) \geq K\}} K\, f_X(x)\, dx = K \int_{\{x \mid g(x) \geq K\}} f_X(x)\, dx = K\, P[g(x) \geq K]$$

$$\downarrow$$

$$E[g(x)] \geq K \cdot P[g(X) \geq K]$$

$$P[g(X) \geq K] \leq \frac{E[g(x)]}{K}$$

Disuguaglianza di Cebicev

$$P[\,|X - E[X]| \geq \varepsilon\,] \leq \frac{Var[X]}{\varepsilon^2} \quad \forall \varepsilon > 0$$

Definizione **Momenti di una variabile aleatoria**

Si dice momento K-esimo di una variabile aleatoria X e si indica con μ_K la seguente quantità

$$\mu_K = E[X^K]$$

$$\mu_K = \sum_i x_i^K P_X(x_i) \quad \text{caso discreto}$$

$$\mu_K = \int_{-\infty}^{+\infty} x^K f_X(x)\, dx \quad \text{caso continuo}$$

Osservazione

La $Var[X] = \mu_2 - (\mu_1)^2 \quad (\mu_1 = E[X])$

Definizione **Funzione generatrice dei momenti**

$$m_X(t) = E[e^{tX}]$$

Proprietà

$$\mu_1 = \left[\frac{d}{dt} m_X(t)\right]_{t=0}$$

$$\mu_1 = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx$$

$$m_X(t) = \int_{-\infty}^{+\infty} e^{tx} f_X(x)\, dx$$

$$\frac{d}{dt} m_X(t) = \int_{-\infty}^{+\infty} x\, e^{tx} f_X(x)\, dx \qquad \frac{d}{dt} e^{tx} = x e^{tx}$$

$$\left[\frac{d}{dt} m_X(t)\right]_{t=0} = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx = \mu_1$$

$$\frac{d^k}{dt^k}\left(e^{tx}\right) = x^k e^{tx} \qquad *$$

$$\mu_k = \left[\frac{d^k}{dt^k} m_X(t)\right]_{t=0}$$

Proprietà

Siano X e Y due var aleatorie, se

$$m_X(t) = m_Y(t) \quad \forall t \in [-t_0, t_0]$$

Allora X e Y hanno la medesima densità

Proprietà

Se X possiede la funzione generatrice dei momenti e $Y = aX + b$ allora esiste anche $m_Y(t)$ e $m_Y(t) = e^{tb} m_X(at)$

~~Proprietà~~ Definizione

Data una variabile aleatoria X, si dice funzione caratteristica di X la funzione

$$H_X(t) = E\left[e^{itx}\right] \quad \text{con } i = \text{unità immaginaria}$$

Schema di Bernoulli

Consideriamo un esperimento che si può condurre solo in 2 modi chiarati convenzionalmente ~~esito~~ "successo" e "insuccesso"

Un esperimento di questo tipo è detto di Bernoulli.

Sia p la probabilità che si verifichi il successo e q quella dell'insuccesso.

Segue $q = 1 - p$

Def: Distribuzione di Bernoulli

Una variabile aleatoria X che vale 1 se la prova di Bernoulli si è conclusa con il successo e 0 se l'insuccesso è detta variabile aleatoria di Bernoulli di parametro p; In simboli $X \sim Bet(p)$

Densità della distribuzione di Bernoulli

$P[X=1] = p \qquad P[X=0] = q$

$P[X=k] = p^k q^{1-k}$ dove $k = 0, 1$

Momenti:

$M_X(t) = pe^t + q$

NB $M_X(t) = E[e^{tX}]$

Dunque $E[X] = pe^t|_{t=0} = p$

$E[X] = \frac{d}{dt} M_X(t)|_{t=0}$

$Var[X] = pq$

NB $E[(X - E[X])^2] =$

$= E[X^2] - E^2[X]$

$F_X(t) = \begin{cases} 0 & \text{se } t < 0 \\ 1-p & \text{se } 0 \leq t < 1 \\ 1 & \text{se } t \geq 1 \end{cases}$

N3 $E[X^2] = \frac{d^2}{dt^2} M_X(t)|_{t=0}$

Definizione

Una sequenza di prove di Bernoulli tra di loro indipendenti per le quali la probabilità del successo è costante è detto schema di Bernoulli

- Distribuzione binomiale

Def (tra n tentativi $\to k$ successi)

La variabile aleatoria X che conta il numero di successi in n prove di uno schema di Bernoulli è detta Binomiale: $X \sim Bin(n, p)$

Proprietà Densità della Distr. Binomiale

$$P[X=K] = \binom{n}{k} p^k q^{n-k}$$ Per ogni k tra 0 e n

Dim

Abbiamo n tentativi con k successi ($n-k$ insuccessi)

Prob di K successi? $\rightarrow p^k$

Prob di $n-k$ insuccessi? $\rightarrow q^{n-k}$

In quanti modi possono avvenire questi k successi?

$\binom{n}{k}$. Dunque $P[X=k] = \binom{n}{k} p^k q^{n-k}$

Momenti della Binomiale (Ricordarsi Calcoli)

$$M_X(t) = (q + pe^t)^n$$

$$E[X] = np \qquad Var[X] = npq$$

$$M_X(t) = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k q^{n-k} =$$

$$= (pe^t + q)^n \quad [\text{Formula di Newton}]$$

$$M_X(t) = (pe^t + q)^n$$

- $E[X] = \frac{d}{dt} M_X(t)\Big|_{t=0} = n(pe^t + q)^{n-1} pe^t\Big|_{t=0} =$

$$= np(p+q) = np \qquad (p+q=1)$$

$E[X^2] \rightarrow \frac{d^2}{dt^2} M_X(t) = np(n-1)(pe^t + q)^{n-2}(pe^t)^2 +$

$$+ np(pe^t + q)^{n-1} e^t$$

Per $t=0$ $= np[(n-1)p + 1] = np[np - p + 1]$

$$Var[X] = E[X^2] - E^2[X] =$$
$$= \cancel{n^2p^2} - np^2 + np - \cancel{n^2p^2} = np(1-p) = npq$$

- **Distribuzione geometrica** (1° successo al k-esimo tentativo)

La variabile aleatoria che conta il numero di prove di uno schema di Bernoulli fino alla comparsa del 1° successo è detta variabile geometrica di parametro p; $X \sim Geo(p)$

Densità della geometrica

$$P[X=k] = p\,q^{k-1}$$

Dim:

$k-1$ tentativi con insuccesso $\rightarrow$ prob? q^{k-1}

al k-esimo tentativo c'è successo $\rightarrow$ prob? p

$$P[X=k] = \underbrace{P[\text{insuccesso}] \cdot P[\text{insuccesso}] \cdots P[\text{insuccesso}]}_{k-1} \cdot p$$
$$= q^{k-1}p$$

Momenti della geometrica

$$M_X(t) = \frac{pe^t}{1-qe^t} \quad \text{per } t < \ln\frac{1}{q}$$

$$E[X] = \frac{1}{p} \qquad Var[X] = \frac{q}{p^2}$$

Dim

$$M_X(t) = E[e^{Xt}] = \sum_{k=1}^{\infty} e^{tk} q^{k-1} p = p\sum_{k=1}^{\infty} e^{tk} q^{k-1} =$$
$$= p\sum_{k=1}^{\infty} e^{tk} e^{-t} e^{t} q^{k-1} = pe^t \sum_{k=1}^{\infty} e^{t(k-1)} q^{k-1} =$$
$$= pe^t \sum_{k=1}^{\infty} (qe^t)^{k-1} = pe^t \sum_{k=0}^{\infty} (qe^t)^k =$$
$$= pe^t \frac{1}{1-qe^t} \quad \text{per } qe^t < 1 \left[\, t < \ln\frac{1}{q} \,\right]$$

$E[X]$ = derivata 1ª $\qquad Var$ = derivata 2ª $- (E[X])^2$

$$E[X] = \frac{d}{dt} m_X(t) \Big|_{t=0} = \frac{pe^t(1-qe^t) - pe^t(-qe^t)}{(1-qe^t)^2}$$

$$= \frac{pe^t(1-qe^t) + pqe^{2t}}{(1-qe^t)^2} \Big|_{t=0} = \frac{p(1-q)+pq}{(1-q)^2} =$$

$$= \frac{p - \cancel{pq} + \cancel{pq}}{p^2} = \frac{1}{p}$$

Proprietà

Sia $X \sim Geom(p)$ allora

$$P[X = i+j \mid X > i] = P[X = j]$$

"Assenza di memoria" Manca una dipendenza dai valori precedenti

Dim (con i e $j \geq 0$)

$$X = i+j \rightarrow X > i \rightarrow (X = i+j) \cap (X > i) = (X = i+j)$$

$$P[X = i+j \mid X > i] = \frac{P[(X=i+j) \cap (X>i)]}{P[X > i]} =$$

(E' la prob condizionata)

$$= \frac{p\,q^{i+j-1}}{q^i} = p\,q^{j-1} = P[X = j]$$

Distribuzione binomiale negativa **Definizione** *

La variabile aleatoria che conta il numero di tentativi necessari al verificarsi dell'n-esimo successo di uno schema di Bernoulli è detta binomiale negativa di parametri n e p; $X \sim Bin(n, p)$

Esercizio

lancio di un dado e gli eventi $A_1 = \{5, 6\}$ $A_2 = \{2, 4, 6\}$ e supponendo di sapere che $P[A_1] = \frac{1}{4}$ e $P[A_2] = \frac{1}{2}$ $P[A_1 \cap A_2] = 1/8$

$B_1 = A_1 \cup A_2$ $B_2 = A_1 \cup (A_2 \cap \bar{A}_1)$ $B_3 = (A_1 \cup A_2) \cap (\bar{A}_1 \cup \bar{A}_2)$

Esercizio fatto 6 pagine dietro riferito B_2 e B_3

$\bar{A}_1 = \{1, 2, 3, 4\}$ $\nearrow 2, 4$

$P[B_2] = P[A_1 \cup (A_2 \cap \bar{A}_1)] = P[\{2, 4\} \cup \{5, 6\}] =$

$= P[\{2, 4, 5, 6\}]$

$\{2, 4, 5, 6\}$ è $A_1 \cup A_2$ la cui prob l'abbiamo già calcolata.

$P[B_3] = ?$

$\bar{A}_1 = \{1, 2, 3, 4\} \cap$

$\bar{A}_2 = \{1, 3, 5\} \rightarrow \{1, 3\}$

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

$P[\{2, 4, 5, 6\} \cap \{1, 2, 3, 4, 5\}] =$

$= P[\{2, 4, 5\}]$

Proprietà *

$P[X = k] = \binom{k-1}{n-1} p^n q^{k-n}$

Momenti

$m_X(t) = \left(\frac{pe^t}{1 - qe^t}\right)^n \quad t < \ln \frac{1}{q}$

$E[X] = \frac{n}{p} \qquad Var[X] = n \frac{q}{p^2}$

Distribuzione di Poisson (caso limite di Bernoulli) Binomiale

NB Bernoulli: $P[X = k] = \binom{n}{k} p^k q^{n-k} \quad E[X] = np$

In questo caso limite $n \to \infty$ $\quad np \to \mu \neq 0$

$p \to 0$

$P[X = k] = \frac{n!}{k!(n-k)!} p^k \frac{q^n}{q^k} =$

$$= \frac{1}{k!} \frac{n!}{(n-k)!} p^k \frac{(1-p)^n}{(1-p)^k} =$$

$$= \frac{1}{k!} \frac{n!}{(n-k)!} \frac{1}{n^k} (np)^k (1-p)^n \frac{1}{(1-p)^k}$$

$n \to \infty \quad p \to 0 \quad np \to \lambda$

$$\frac{n!}{(n-k)!} \frac{1}{n^k} \xrightarrow[n \to +\infty]{} 1 \qquad \text{Segue} \quad \frac{n!}{(n-k)!} \frac{1}{n^k} (np)^k \to \lambda^k$$

$$(1-p)^n \to \left(1 - \frac{np}{n}\right)^n \xrightarrow[np \to \lambda]{} \left(1 - \frac{\lambda}{n}\right)^n \xrightarrow[n \to +\infty]{} e^{-\lambda}$$

$$\frac{1}{(1-p)^k} \xrightarrow{p \to 0} 1$$

k: numero di cui si vuole prendere la probabilità

Perciò il tutto tende a $\frac{1}{k!} \lambda^k e^{-\lambda}$

$$P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

λ: numero successi attesi in un dato intervallo. (freq media accadimento dell'evento considerato)

Def

La variabile aleatoria avente densità

$$P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

è detta variabile di Poisson di parametro λ

$X \sim Poi(\lambda)$

Momenti

$m_X(t) = e^{\lambda(e^t - 1)}$ $\qquad \mathbb{E}[X] = \lambda$

$Var[X] = \lambda$

NB $m_X(t) = E[e^{tX}]$

Dim

$$m_X(t) = \sum_{k=0}^{+\infty} e^{tk} \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{+\infty} \frac{(e^t \lambda)^k}{k!} =$$

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

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

$$\mathbb{E}[X] = \frac{d}{dt} M_X(t) \Big|_{t=0} = \mu e^{\mu(e^t-1)} \mu e^t \Big|_{t=0} = \mu$$

$$E[X^2] \frac{d^2}{dt^2} M_X(t) \Big|_{t=0} = \mu \left[e^{\mu(e^t-1)} \mu e^t e^t + e^{\mu(e^t-1)} e^t \right] \Big|_{t=0} =$$

$$= \mu \left[e^{\mu(e^t-1)} \mu e^{2t} + e^t e^{\mu(e^t-1)} \right] \Big|_{t=0} =$$

$$= \mu(\mu+1)$$

$$Var[X] = E[X^2] - E^2[X] = \mu(\mu+1) - \mu^2 =$$

$$= \mu^2 + \mu - \mu^2 = \mu$$

Esempio

Il numero di automobili che transita per un passaggio a livello è distribuito secondo una variabile di Poisson.
In media transitano 36 automobili in un'ora.
Calcolare la prob. che nell'arco di 5 minuti, transitino 2 automobili, al massimo 2, almeno 2

Sol X = "num. auto che transitano in 5 minuti"

$P[X=2] = ?$ $\quad P[X \leq 2] = ?$ $\quad P[X \geq 2] = ?$

Sappiamo $P[X=k] = \cancel{e^{\mu(e^t-1)}} \dfrac{\mu^k e^{-\mu}}{k!}$ $\qquad \mu = \dfrac{36}{50} \cdot 5 = 3$

$$P[X=2] = \frac{e^{-3} 3^2}{2!} \quad \checkmark$$

$$P[X \leq 2] = P[X=0] + P[X=1] + P[X=2] \quad \checkmark$$

$$P[X \geq 2] = 1 - P[X=0] - P[X=1] \quad \checkmark$$

Schema di Poisson

In un certo intervallo di tempo (o di spazio) si manifestano degli eventi, detti "arrivi", che soddisfano le seguenti condizioni:

1) Non ci sono più arrivi simultanei (non ci sono sovrapposizioni)
2) Il numero medio di arrivi in un'unità di tempo è costante (intensità è costante)
3) Ciascun arrivo avviene in modo indipendente dagli altri

Definizione (Processo di Poisson)

La variabile aleatoria N_t che conta in un certo intervallo di tempo t il numero di arrivi in uno schema di Poisson di intensità λ è detta processo di Poisson

Caratteristiche

1) $N_0 = 0$
2) $N_{t+1} = N_t + N_1$
3) se $\Delta t \to 0$

$$P[N_{t+\Delta t} - N_t = 1] = \lambda \Delta t + o(\Delta t)$$

$$P[N_{t+\Delta t} - N_t \geq 1] = o(\Delta t)$$

$$P[N_{t+\Delta t} - N_t = 0] \simeq 1 - \lambda \Delta t + o(\Delta t)$$

Proprietà

Sia N_t un processo di Poisson di intensità λ, allora $N_t \sim Poi(\mu = \lambda t)$

Distribuzione esponenziale

La variabile aleatoria "tempo di attesa del primo arrivo in un processo di Poisson di intensità λ" è detta esponenziale di variabile λ

Proprietà

$$f_X(x) = \begin{cases} 0 & x \leq 0 \\ \lambda e^{-\lambda x} & x > 0 \end{cases} \qquad F_X(t) = \begin{cases} 0 & t < 0 \\ 1 - e^{-\lambda t} & t \geq 0 \end{cases}$$

Non c'è 0 perché il tempo di attesa $= 0$ non ha senso.

Dim

$F_X(t) = P[X \leq t]$ prob che il primo arrivo sia entro t

$P[X > t] = 1 - P[X \leq t] = 1 - F_X(t)$ ⇙ ? *

Segue $F_X(t) = 1 - P[X > t] = 1 - P[N_t = 0] =$

$= 1 - e^{-\lambda t} \Rightarrow f_X(t) = \frac{d}{dt} F_X(t) = \lambda e^{-\lambda t}$

Momenti della funzione esponenziale

$M_X(t) = \frac{\lambda}{\lambda - t}$ per $t < \lambda$

$\mathbb{E}[X] = 1/\lambda \qquad Var[X] = 1/\lambda^2$

Dim

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f_X(x)\, dx = \int_0^{+\infty} e^{tx} \lambda e^{-\lambda x}\, dx =$$

$$= \lambda \int_0^{+\infty} e^{(t-\lambda)x}\, dx = \lambda \left[\frac{e^{(t-\lambda)x}}{t-\lambda} \right]_0^{\infty} =$$

$$= \lambda \left[\underbrace{\frac{e^{(t-\lambda)\infty}}{t-\lambda}}_{0 \text{ se } t-\lambda < 0 \ (t < \lambda)} - \frac{1}{t-\lambda} \right] = \frac{\lambda}{\lambda - t}$$

Proprietà Assenza di memoria della esponenziale

$$P[X > t+s \mid X > t] = P[X > t+s]$$

Definizione **Distribuzione Gamma**

La variabile aleatoria "tempo di attesa dell'n-esimo arrivo di un processo di Poisson" è detta gamma di parametri n e λ; $X \sim \Gamma(n, \lambda)$

Densità

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{(n-1)!} \lambda^n x^{n-1} e^{-\lambda x} & x > 0 \end{cases}$$

- Densità di Weibul - Funzione rischio
- Distribuzione di Weibul

Sia $Y \sim Exp(\lambda)$ e sia $\lambda > 0$, è detta variabile ~~aleatoria~~ di Weibul la variabile aleatoria $X = Y^{1/\alpha}$, $X \sim Wei(\alpha, \lambda)$

Densità

$$f_X(x) = \begin{cases} 0 & x < 0 \\ \lambda \alpha x^{\alpha-1} e^{-\lambda x^\alpha} & x \geq 0 \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x^\alpha} & x \geq 0 \end{cases}$$

Dim

$Y \geq 0$, $X = Y^{1/\alpha}$

$F_X(x) = P[X \leq x] = P[Y^{1/\alpha} \leq x] =$

$P[Y \leq x^\alpha] = F_Y(x^\alpha) = 1 - e^{-\lambda x^\alpha}$ $\cap$

$f_X(x) = F_X'(x)$

Oss:

$$E[X] = \alpha^{-1/\lambda}\,\Gamma\left(1+\frac{1}{\lambda}\right)$$

$$Var[X] = \alpha^{-1/\lambda}\left[\Gamma\left(1+\frac{1}{\lambda}\right) - \Gamma\left(1+\frac{1}{\lambda}\right)\right]$$

Funzione di rischio

Variabile aleatoria T = "tempo di funzionamento"

Probabilità che, condizionatamente al fatto che, fino all'istante $\bar{t}$ l'apparecchio fosse in funzione esso si guasti immediatamente dopo, ossia all'istante $\bar{t} + \Delta t$ è $P[T \leq \bar{t} + \Delta t \mid T > \bar{t}]$

Distribuzione di Gauss o Normale

$N(\mu, \sigma)$

Def: Una variabile aleatoria X si dice distribuita secondo una normale o gaussiana di parametri $\mu \in \mathbb{R}$ e $\sigma \in \mathbb{R}^+$ ($X \sim N(\mu, \sigma)$) se ha densità di probabilità

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Proprietà

1) $f_X(x)$ è simmetrica rispetto alla retta $x = \mu$
2) ha un ~~minimo~~ massimo assoluto in $x = \mu$ cioè in $\left(\mu, \frac{1}{\sigma\sqrt{2\pi}}\right)$
3) ha asintoto orizzontale $y = 0$
4) ha 2 flessi in $\left(\mu - \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$ e $\left(\mu + \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$
5) $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$
6) All'aumentare di σ il massimo diminuisce (la curva si abbassa)

7) Se $\mu = 0$ è simmetrica rispetto all'asse y

8) $E[X] = \mu$ $\quad Var[X] = \sigma^2$

Funzione di ripartizione è

$$F_X(t) = P[X \leq t] = \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx$$

D1 Distribuzione normale standardizzata

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Ci arriviamo ponendo $z = \frac{x-\mu}{\sigma}$ $\quad x = \sigma z + \mu$

$dx = \sigma\, dz$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Integriamo

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Abbiamo integrato (con sost di dx) per mantenere uguali le prob.

Vettori aleatori

È detto vettore aleatorio o variabile aleatoria multidimensionale, un vettore le cui componenti sono variabili aleatorie.

$$X \equiv (X_1, X_2 \dots X_n)$$

- $P[X_1 \cap X_2 \dots] \equiv P[X_1, X_2 \dots X_n]$

Se abbiamo le variabili aleatorie X e Y

- $\sum_i \sum_j P[X=i, Y=j] = 1$
- $\sum_i P[X=i] = 1$ $\quad$ - $\sum_j P[Y=j] = 1$
- I valori $P[X=i, Y=j]$ formano la densità discreta congiunta del vettore (X, Y)

- Le densità $P[X=i]$, $P[Y=j]$ sono dette densità marginali

Oss: E' sempre possibile ricavare le prob. marginali dalle probabilità congiunte ma non vale il viceversa.

Def Funzione di ripartizione

Data un vettore aleatorio (X, Y) è detta funzione di ripartizione congiunta la fnc

$$F_{X,Y}(x,y) = P[X \leq x, Y \leq y] =$$
$$= P[(X \leq x) \cap (Y \leq y)]$$

Def Funzione di ripartizione marginale

Data una funzione di ripartizione congiunta $F_{X,Y}$ è definita funzione di ripartizione marginale come

$$F_X(x) = P[X \leq x] = P[X \leq x, Y < +\infty] =$$
$$= \lim_{y \to +\infty} F_{X,Y}(x,y)$$

Proprietà

Per ogni $y \in \mathbb{R}$ $\lim_{x \to -\infty} F_{XY}(x,y) = 0$, $\lim_{x \to +\infty} F_{XY}(x,y) = F_Y(y)$

$\forall x \in \mathbb{R}$ $\lim_{y \to \infty} F_{XY}(x,y) = 0$, $\lim_{y \to +\infty} F_{XY}(x,y) = F_X(x)$

- Densità congiunta nel caso di varietà

Si dice densità congiunta del vettore aleatorio di varietà (X, Y) e si indica con P_{XY} l'insieme dei valori $P_{XY}(x_i, y_j) = P[X = x_i, Y = y_j]$

- Densità marginale nel caso di varietà

$$P_X(x_i) = \sum_j P_{XY}(x_i, y_j)$$

$$P_Y(y_j) = \sum_i P_{XY}(x_i, y_j)$$

- Densità congiunta nel caso continuo

Dato un vettore aleatorio continuo (X, Y) è detta funzione di densità di probabilità la funzione $f_{XY}(x,y)$ per la quale, per ogni scelta di $a, b, c, d \in \mathbb{R}$

$$P[a < X < b,\ c < Y < d] =$$

$$= \int_a^b \left(\int_c^d f_{XY}(x,y)\, dy \right) dx = \int_c^d \left(\int_a^b f_{XY}(x,y)\, dx \right) dy$$

Si indica con:

$$\int_a^b dx \int_c^d f_{XY}(x,y)$$

Si dimostra che

$$F_{XY}(s,t) = P[X < s,\ Y < t] =$$

$$= \int_{-\infty}^{s} dx \int_{-\infty}^{t} dy\, f_{XY}(x,y)$$

Sappiamo che $f_X(x) = \dfrac{d F_X(x)}{dx}$

Nel caso bidimensionale

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x\, \partial y} F_{XY}(x,y)$$ derivata seconda mista

<u>Definizione Funzioni densità marginali</u>

Sono dette funzioni densità marginali le funz:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\, dy, \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\, dx$$

Sappiamo che $\int_{-\infty}^{+\infty} f_X(x)\, dx = 1 \left(= \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy\, f_{XY}(x,y) \right)$

Def

Dati n variabili aleatorie X_i con $i = 1 \dots n$, si dicono indipendenti se per ogni scelta gli intervalli A_i si ha che

$$P[X_1 \in A_1, X_2 \in A_2, \dots X_n \in A_n] =$$
$$P[X_1 \in A_1] \cdot P[X_2 \in A_2] \dots P[X_n \in A_n]$$

Si dimostra che (nel caso di variabili indipendenti)

$$p_{XY}(x_i, y_i) = p_X(x_i)\, p_Y(y_i)$$
$$f_{XY}(x_i, y_i) = f_X(x_i) \cdot f_Y(y_i)$$
$$F_{XY}(x, y) = F_X(x)\, F_Y(y)$$

Teorema

Dato un vettore aleatorio (X, Y) (discreto o continuo) la somma $U = X + Y$ delle sue componenti ha densità

$$p_U(u) = \sum_k p_{XY}(k, u-k) \quad \text{discreto}$$
$$f_U(u) = \int_{-\infty}^{+\infty} f_{XY}(t, u-t)\, dt \quad \text{continuo}$$

Def

Sia (X, Y) un vettore aleatorio e $g(X, Y)$ una funzione, $\mathbb{E}[g(X, Y)]$, se esiste, è dato da

$$\mathbb{E}[g(X, Y)] = \sum_i \sum_j g(x_i, y_j)\, p_{XY}(x_i, y_j) \quad \text{c.d.}$$
$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{dy} g(x, y)\, f_{XY}(x, y) \quad \text{c.c.}$$

Proprietà (dim esercizio)

Data coppia di variabili aleatorie X, Y il valore atteso della somma è uguale alla somma dei valori attesi

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Def Covarianza

Si indica con $Cov[X, Y]$ il $\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$

Proprietà (dim esercizio)

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

Proprietà (dim esercizio)

X, Y, V, W variabili aleatorie $\quad \alpha, \beta \in \mathbb{R}$

1) $Cov[X, \alpha] = 0$
2) $Cov[X, Y] = Cov[Y, X]$
3) $Cov[X, X] = Var[X]$
4) $Cov[\alpha X, \beta Y] = \alpha\beta\, Cov[X, Y]$
5) $Cov[X+Y, V+W] =$
$= Cov[X, V] + Cov[X+W] + Cov[Y, V] + Cov[Y+W]$

Def Matrice di covarianza

Dato $X = (X_1, X_2, \dots, X_n)$ la matrice di covarianza è la matrice di elementi

$(C_X)_{ij} = Cov[X_i, X_j]$ $\quad C_X$ è la matrice

Def Coefficiente di correlazione

Date 2 variabili aleatorie X, Y è detto coefficiente di correlazione ρ_{XY}

$$\rho_{XY} = \frac{Cov[X, Y]}{\sqrt{Var[X]\,Var[Y]}}$$

Proprietà

1) $-1 < \rho_{XY} < 1$
2) $|\rho_{XY}| = 1$ se vi è una tra X e Y esiste

un valore deterministico lineare ovvero

$Y = aX + b$ con a e $b \in \mathbb{R}$ con almeno $a \neq 0$

Def **Variabili incorrelate**

Si dicono incorrelate le variabili aleatorie X e Y per cui la $Cov[X, Y] = 0$ (segue che il coeff. di correlazione è nullo)

Proprietà

Se X e Y sono indipendenti allora sono incorrelate

Non è detto il viceversa.

Dim

$$Cov[X, Y] = E[(x - E[X])(y - E[Y])] =$$
$$= E[XY] - E[X]\,E[Y]$$

$$E[XY] = \sum_i \sum_j x_i y_j P(X_i, Y_j) =$$
$$= \sum_i \sum_j x_i y_j P(x_i) P(y_j) =$$
$$= \sum_i x_i P(x_i) \cdot \sum_j y_j P(y_j) = E[X]\,E[Y]$$

Proprietà

1) $Var[X+Y] = Var[X] + Var[Y] + 2\,Cov[X, Y]$
2) $Var[X-Y] = Var[X] + Var[Y] - 2\,Cov[X, Y]$
3) Se X_i sono indipendenti segue

$$Var[\sum_i X_i] = \sum_i Var[X_i]$$

Trasformazione di vettori aleatori

$(X, Y) \to (U, V)$ $\quad u = g_1(x, y) \quad v = g_2(x, y)$

$$\Updownarrow$$

h_1 e h_2 sono funzioni inverse. $\quad x = h_1(u, v) \quad y = h_2(u, v)$

Teorema

Sia (X, Y) un vettore aleatorio continuo e sia $(U, V) = g(X, Y)$ una ~~funzione~~ trasformazione invertibile con trasformazione inversa $h(U, V)$

Allora $f_{U,V}(U,V) = f_{X,Y}(h_1(U,V), h_2(U,V)) \cdot |\det(J)|$

$$J = \begin{bmatrix} \frac{dh_1}{dU} & \frac{dh_1}{V} \\ \frac{dh_2}{dU} & \frac{dh_2}{dV} \end{bmatrix}$$

Distr. Uniforme

Prob → valore costante in $[a, b]$

Def

X ha distr. uniforme in $[a, b] \subset \mathbb{R}$ se è assolutamente continua con densità di probabilità

$$f_X(t) = \begin{cases} \frac{1}{b-a} & \text{se } t \in [a, b] \\ 0 & \text{se no} \end{cases}$$

$$F_X(t) = \begin{cases} 0 & \text{se } t < a \\ \frac{t-a}{b-a} & \text{se } t \in [a, b] \\ 1 & \text{se } t > b \end{cases}$$

$E[X] = \frac{a+b}{2}$ $\quad$ $Var[X] = \frac{(b-a)^2}{12}$

Ottenere la Normale

$$B(n, p) \xrightarrow{n \to +\infty} N(\mu, \sigma) \xleftarrow{\lambda \to +\infty} Poi(\lambda)$$

Distribuzione del Chi-quadro χ^2

Siano X_i, $i = 1, \dots n$ variabili aleatorie con distribuzione normale $N(0,1)$ indipendenti tra loro

Sia X una variabile aleatoria

$$X = X_1^2 + X_2^2 + \dots X_n^2$$

Allora si prova che questa variabile aleatoria è distribuita secondo una distribuzione chi-quadro

Def Diremo che la variabile aleatoria X è distribuita secondo una $\chi^2(n)$ con gradi di libertà n se è assolutamente continua con densità di probabilità

$$f_X(t) = \begin{cases} 0 & \text{se } t < 0 \\ \frac{1}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{2}\right)^{n/2} t^{\frac{n}{2}-1} e^{-\frac{1}{2}t} & t \geq 0 \end{cases}$$

Dove

$$\Gamma(\alpha) = \int_0^{+\infty} e^{-x} x^{\alpha-1} dx \quad \text{Gamma di Eulero}$$

$E[X] = n$

$Var[X] = 2n$

Distribuzione T di Student

Siano $Z \sim N(0,1)$ e $Y \sim \chi^2(n)$ due variabili aleatorie indipendenti

Definiamo $X = \frac{Z}{\sqrt{Y/n}}$ e si prova che X è una variabile aleatoria distribuita secondo una T di Student con n gradi di libertà

Def

X è distribuita secondo una T di Student con n gradi di libertà se è assolutamente continua con densità di probabilità

$$f_X(t) = \Gamma\left(\frac{n+1}{2}\right) / \Gamma\left(\frac{n}{2}\right) \cdot \frac{1}{\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}(n+1)}$$

Si prova che

- $E[X] = 0$ se $n > 1$ $\quad Var[X] = \frac{n}{n-1}$ se $n > 2$
- Per $n \geq 30$ la T di Student approssima $N(0,1)$

Consideriamo $X_1, X_2 \dots X_n$ una n-upla di variabili indipendenti e identicamente distribuite (iid)

STATISTICA

Campione Casuale

Data una variabile aleatoria X chiamiamo campione casuale (o brevemente campione) di ampiezza n una n-upla di variabili aleatorie tra loro indipendenti ciascuna delle quali distribuita come la popolazione X

NB

$X_1 \dots X_n \longrightarrow x_1, x_2 \dots x_n$

n-upla di variabili $\qquad$ n-upla di valori (realizzazione)

Media Campionaria

La media campionaria di un campione casuale di ampiezza n $X_1 \dots X_n$ è data da

$$\overline{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Momento Campionario di ordine k

$$M_n^k = \frac{X_1^k + X_2^k + \dots + X_n^k}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Varianza Campionaria

1) Valore atteso noto $E[X] = \mu$

$$S_0^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

2) Valore atteso ignoto

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

Il -1 deriva dal fatto che abbiamo usato già 1 volta gli n dati

Proposizione

Se è nota la media della popolazione X segue

$$S_0^2 = M_n^{(2)} - 2m\bar{X}_n + m^2$$

Se la media della popolazione è incognita, allora

$$S_n^2 = \frac{1}{n-1}\left(n M_n^{(2)} - n\bar{X}_n^2\right)$$

Proposizione Valore atteso e varianza della media campionaria

Se la popolazione X è stata di media e varianza $E[X] = m$ e $Var[X] = V$. Dato un campione iid $X_1, X_2, \dots, X_n$ si ha:

$$E[\bar{X}_n] = m \quad , \quad Var[\bar{X}_n] = \frac{1}{n}V$$

Dim

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

$$= \frac{1}{n} n E[X] = m$$ ← Sappiamo tutte le n la stessa distr (di X)

$$Var[\bar{X}_n] = Var\left[\frac{X_1 + X_2 + \dots X_n}{n}\right] = \frac{1}{n^2}Var[X_1] + \dots Var[X_n]$$

↑ NB: $V(aX) = a^2 V(X)$, a = costante

$$= \frac{1}{n^2} n\, Var[X] = \frac{V}{n}$$

Proposizione Valore atteso della varianza campionaria

Se la popolazione X è stata di media e varianza $E[X] = m$ $Var[X] = V$ allora dato un campione $X_1, X_2 \dots X_n$ si ha

$$E[S_0^2] = V \qquad E[S_n^2] = V$$

Facciamo caso che la popolazione segue la legge normale

$$X \sim Nor(\mu = m, \sigma^2 = V)$$

Proprietà Distr media campionaria.

Sia $X_1, X_2 \dots X_n$ un campione estratto da $X \sim N(m, V)$

Detta $\bar{X}_n$ la media campionaria si ha

$$\bar{X}_n \sim N\left(\mu = m, \sigma^2 = \frac{V}{n}\right) \left(\frac{\bar{X}_n - m}{\sqrt{\frac{V}{n}}} \sim Nor(\mu=0, \sigma \right.$$

Proprietà

Sia $X_1 \dots X_n$ un campione estratto da una popolazione $X \sim N(\mu = m, \sigma^2 = V)$ allora detta S_0^2 la varianza campionaria (a media nota)

$$S_0^2 \sim \frac{V}{n} \chi^2 \ (V = n)$$

Invece con media incognita

$$S_n^2 \sim \frac{V}{n-1} \chi^2 \ (V = n-1)$$

Proprietà

Sia $X_1, X_2 \dots X_n$ un campione estratto da una popolazione $X \sim N(\mu = m, \sigma^2 = V)$

$\bar{X}_n$ è indipendente da S_n^2

Proprietà

Sia $X_1 \dots X_n$ un campione estratto da una popolazione $X \sim N(\mu = m, \sigma^2 = V)$ allora

$$\frac{\bar{X}_n - m}{\sqrt{\frac{S_n^2}{n}}} \sim T \ (V = n-1)$$

← Radice nella Normale

Statistica descrittiva: Appunti Muscio CAP 1

Capitolo 1

Statistica descrittiva

Con il termine statistica descrittiva si intende una raccolta di metodi, strumenti matematici atti ad organizzare serie di dati per evidenziare in forma sintetica simmetrie, periodicità, leggi di altro genere ovvero di descrivere in maniera immediatamente comprensibile le informazioni dagli stessi dati. Sia $E = \{x_1, x_2, \ldots, x_n\}$ un insieme di dati di numerosità n diamo alcune semplici definizioni.

i) modalità del carattere

I valori assumibili dai dati discreti sono di solito un numero limitato: per questo motivo conviene raggrupparli considerando l'insieme dei valori assumibili detti modalità del carattere. Sia N il numero dei valori assumibili allora denoto con $S = \{s_1, s_2, \ldots, s_N\}$ l'insieme delle modalità.

ii) frequenza assoluta e relativa

$$f_j = n^\circ \text{ elementi di } E \text{ con valore } s_j \qquad j = 1, 2, \ldots, N$$

$$p_j = \frac{f_j}{n} \qquad n = \text{ numerosità di } E$$

iii) Frequenza cumulata assoluta e relativa

$$F_j = \sum_k f_k \quad \text{con } k \text{ tale che } s_k \leq s_j \qquad j = 1, 2, \ldots, N$$

$$P_j = \sum_k p_k \quad \text{con } k \text{ tale che } s_k \leq s_j \qquad j = 1, 2, \ldots, N$$

Tabella 1.1: Numero stanze in un campione di appartamenti.

3	4	2	6	5	2	4	4
2	5	4	4	5	7	5	4
5	7	8	4	3	6	2	3
5	2	7	2	4	8	4	2
6	5	4	4	6	5	3	3
8	5	2	5	6	5	5	4
2	6	4	5	5	7	3	4
3	3	3	4	4	3	4	6
4	3	7	4	4	6	4	2
4	4	6	3	2	3	5	4

Esempio 1.1 *Nella tabella 1.1 sono elencati i dati relativi al numero di stanze possedute da 80 appartamenti scelti a caso tra tutti quelli che si trovano in un determinato quartiere di una data città.*

Esercizio 1.1 *Dalla tabella 1.1 creare una tabella come la seguente*

n° stanze	freq. ass.	freq. rel	freq. cum. ass.	freq. cum. rel.
1				
2				
3				
4				
5				
6				
7				
8				

Utilizzare Excel (funzione conta.se*) e fare gli istogrammi.*

Tabella 1.2: Costo mq di un campione di appartamenti.

2,11	3,08	2,35	3,54	0,44	2,24	4,60	1,88
2,08	1,90	2,15	5,11	3,69	0,88	2,56	4,00
3,15	3,67	3,15	4,09	4,57	1,06	2,05	2,34
4,17	4,10	4,75	1,90	2,36	0,90	2,07	3,23
4,21	2,12	1,21	2,10	4,05	5,42	0,85	4,80
2,11	5,08	2,78	4,88	1,11	1,83	1,85	2,87
2,23	3,20	2,80	2,19	1,88	2,16	2,74	2,45
1,19	3,79	1,24	3,06	2,11	3,70	2,91	1,80
3,48	4,10	3,13	0,90	3,07	4,10	1,66	2,88
2,11	1,90	1,18	0,75	1,60	3,85	1,45	2,00

Può succedere che le modalità siano elevate, allora conviene considerare dei sottoinsiemi di S. Chiamo classe C un sottoinsieme di S.

Chiamo partizione di S ogni famiglia di classi tra loro disgiunte la cui unione sia S ovvero

$$C_i \subseteq S \quad i = 1, \ldots, k \in N$$

$$C_i \cap C_j = \phi \quad \forall i \neq j$$

$$\bigcup_{i=1}^{k} C_i = S$$

Esempio 1.2 *Nella tabella 1.2 sono elencati i dati relativi al costo al metro quadro (in migliaia di Euro) di 80 appartamenti scelti a caso tra quelli che si trovano in un quartiere di una città italiana.*

Volendo raggruppare in classi i dati riportati in questa tabella, la prima cosa che occorre fare è osservare quali sono i valori minimo e massimo in essa riportati. Essendo questi 0,44 e 5,42, possiamo arbitrariamente pensare all'insieme S dei valori assumibili come al sottoinsieme $[0.40, 5.50] \subset \mathbb{R}$. Se vogliamo suddividere S in 5 classi, potremmo ad esempio scegliere le

seguenti:

$$C_1 = (0.40, 1.50],$$

$$C_2 = (1.50, 2.30],$$

$$C_3 = (2.30, 3.00],$$

$$C_4 = (3.00, 4.00],$$

$$C_5 = (4.00, 5.50].$$

Esercizio 1.2 *Riferendosi alla tabella 1.2 creare una matrice con*

costo mq	*freq. ass.*	*freq. rel*	*freq. cum. ass.*	*freq. cum. rel.*
C_1				
C_2				
C_3				
C_4				
C_5				

Creare istogrammi con Excel.

1.1 Indici di tendenza centrale

Gli indici di tendenza centrale sono delle misure in grado di sintetizzare con un solo valore numerico, i valori assunti dai dati.

i) Media e Media pesata

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

$$\bar{x}_p = \frac{1}{s}\sum_{i=1}^{n} s_i x_i \qquad , s_i \text{ sono i pesi} \quad s = \sum_{i=1}^{n} s_i \tag{1.2}$$

Proposizione 1.1 *La media rende minima la funzione*

$$f(x) = \sum_{i=1}^{n} (x - x_i)^2$$

ovvero la media e' quel punto che dista di meno da tutti i punti della serie di dati $\{x_i\}$.

Dim.

$$f'(x) = 2\sum_{i=1}^{n} (x - x_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} x - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad n\,x = \sum_{i=1}^{n} x_i$$

da cui si ricava

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

Bisogna adesso dimostrare che in $\bar{x}$ $f(x)$ è minima.

$$f''(x) = 2\sum_{i=1}^{n} 1 = 2\,n > 0$$

allora $\bar{x}$ è l'unico punto di minimo di $f(x)$. ■

ii) Mediana ($\hat{x}$)

È quel numero che "sta nel mezzo". Ovvero, ordinati i dati $\{x_i\}$ in ordine crescente allora se n è dispari la mediana è l'elemento di posto

$$\frac{n+1}{2} \quad .$$

Se n è pari la mediana non è univocamente determinata. Una possibile approsimazione si ottiene prendendo la **media aritmetica dei due valori centrali**, ovvero quelli di posto

$$\frac{n}{2}, \qquad \frac{n}{2}+1.$$

Un'altra approssimazione si ottiene utilizzando un'interpolazione lineare a partire dai due valori centrali (vedi esempio).

svantaggio: risistemare i dati in ordine crescente (non nella media)

vantaggio: non dipende dai valori estremi

iii) Moda

È quel valore che si ripete più volte nella serie di dati.

Dà un andamento qualitativo dei dati e non è garantito che sia un unico numero (distr. bi o multi-modale).

iv) Quantili, percentili, Quartili

Supponiamo di avere un insieme di dati ordinati in modo crescente. Abbiamo gia' visto che la mediana e' il *valore che sta nel mezzo*. In analogia possiamo definire

Definizione 1.1

Si chiama quantile di ordine $\alpha \in [0,1]$, e lo si indica con q_α, un valore per cui alla sua sinistra compare almeno il 100 $\alpha\%$ delle osservazioni e alla sua destra almeno il 100 $(1-\alpha)\%$.

Alle volte, si usa il termine percentile, al posto di quantile, in questo caso α e' indicata come percentuale. Per esempio $q_{0.95}$ e' il novantacinquesimo percentile.

Definizione 1.2

Si dicono primo quartile, secondo quartile e terzo quartile, e si indicano con Q_1, Q_2, Q_3 i quantili, rispettivamente, di ordine 0.25, 0.5, 0.75 e quindi

$$Q_1 = q_{0.25} \quad , \quad Q_2 = q_{0.50} \quad , \quad Q_3 = q_{0.75}$$

Il secondo quartile coincide con la mediana.

Calcolo dei quartili

Come già detto per la mediana, anche i quartili ed i percentili non sono univocamente determinati. Vediamo due modi per calcolarli:

1. *media aritmetica*

   In analogia con quanto già visto per il calcolo della mediana

   - fissiamo $\alpha = 0.25, 0.5, 0.75$ e calcoliamo $\alpha(n+1)$
   - se $\alpha(n+1) = m \in \mathbb{N}$ allora $Q_\alpha = x_m$
   - se $\alpha(n+1) \notin \mathbb{N}$ allora si prende la sua parte intera [1], che è quel numero $m \in \mathbb{N}$ tale che $m < \alpha(n+1) < m+1$ e quindi, la media aritmetica

   $$Q_\alpha = \frac{x_m + x_{m+1}}{2} \tag{1.3}$$

2. *interpolazione lineare*

   Supponiamo di avere due coppie di dati $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ e di voler stimare il valore $y^\star$ che corrisponde a $x^\star \in]x_1, x_2[$. Con il metodo dell'interpolazione lineare si stima il valore $y^\star$ tramite la retta passante per i due punti P_1 e P_2 (detta retta interpolante), ovvero

   $$y^\star = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x^\star - x_1) \quad . \tag{1.4}$$

   Utilizziamo questo metodo per stimare la mediana:

[1] la parte intera di un numero x si indica con $[x]$, mentre la sua parte frazionaria è $x - [x]$.

- fissiamo $\alpha = 0.25, 0.5, 0.75$ e calcoliamo $\alpha(n+1)$
- se $\alpha(n+1) = m \in \mathbb{N}$ allora $Q_\alpha = x_m$
- se $x^\star = \alpha(n+1) \notin \mathbb{N}$ allora utilizziamo la retta interpolante (1.4).

  Sia $m = [x^\star]$ (parte intera) e $\beta = x^\star - [x^\star] = x^\star - m$ (parte frazionaria), allora $P_1(m, x_m)$, $P_2(m+1, x_{m+1})$ e quindi avremo

$$y^\star = Q_\alpha = x_m + \frac{x_{m+1} - x_m}{m+1-m}(x^\star - m) = x_m + (x_{m+1} - x_m)\beta \tag{1.5}$$

In EXCEL la funzione QUARTILE, viene calcolata usando un' opportuna interpolazione lineare.

Esempio 1.3 *Siano assegnati i seguenti dati, che rappresentano le eta' di un campione di persone*

$$E = \{x_i, i = 1 : 18\} = \{16, 18, 18, 19, 20, 20, 20, 20, 21, 21, 21, 22, 23, 25, 28, 30, 31, 37\}$$

La numerosita' del campione e' n=18. Si ottiene come media 18, mediana 21 e moda 20 (calcolare con EXCEL ed anche la distribuzione delle frequenze).

Per il calcolo dei quartili si ha:

1. Per $\alpha = 0.25$ (primo quartile) 0.25(18+1)= 4.75 quindi $m = [4.75]$=4 e la parte frazionaria $\beta = 0.75$. Usando la media aritmetica (1.3)

$$Q_{0.25} = \frac{x_4 + x_5}{2} = \frac{19 + 20}{2} = 19.5$$

mentre con l'interpolazione lineare (1.5)

$$Q_{0.25} = 19 + 0.75\,(20 - 19) = 19.75$$

2. Per $\alpha = 0.5$ (secondo quartile o mediana) 0.50(18+1)= 9.5 quindi $m = 9$ e β=0.5 . Usando la (1.3)

$$Q_{0.5} = \frac{21 + 21}{2} = 21$$

che coincide con il valore ottenuto utilizzando la (1.5).

3. Per $\alpha = 0.75$ (terzo quartile) 0.75(18+1)= 14.25 quindi $m = 14$ e β=0.25 . Usando la (1.3)

$$Q_{0.75} = \frac{25 + 28}{2} = 26,5$$

mentre con la (1.5) si ottiene

$$Q_{0.75} = 25 + 0.25\,(28 - 25) = 25,75$$

Da questa analisi possiamo concludere che il 25 % delle persone del campione hanno un'eta minore o uguale a 19,5 anni, il 50 % minore o uguale a 21 e il 75 % minore o uguale a 26,5 .

1.2 Indici di variabilità

Può accadere che 2 serie di dati abbiano stessa media e/o mediana, ma le 2 serie sono molto diverse.

Esempio 1.4

$$E_1 = \{0.5, 0.8, 2.0, 2.7, 4.0\} \qquad \bar{x}_1 = \hat{x}_1 = 2$$

$$E_2 = \{1.4, 1.7, 2.0, 2.1, 2.8\} \qquad \bar{x}_2 = \hat{x}_2 = 2$$

Però i dati di E_2 sono più omogenei (cioè vicini tra loro).

Occorre pertanto definire indici che misurino il grado di variabilità o dispersione.

a) Varianza (s^2)

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

ovviamente s^2 è tanto più grande, tanto più i dati sono "distanti" dalla media.

Esempio 1.5

$$s^2_{E_1} = 1.64, \qquad s^2_{E_2} = 0.22 \quad \textit{(omogeneo!)}$$

Proposizione 1.2

$$s^2 = \overline{(x^2)} - (\bar{x})^2$$

Dim.

$$\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = && (1.6)\\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = && (1.7)\\
&= \frac{1}{n}\left[\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}\bar{x}^2 - 2\sum_{i=1}^{n}x_i\bar{x}\right] = && (1.8)\\
&= \frac{1}{n}\left[\sum_{i=1}^{n}x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n}x_i\right] = && (1.9)\\
&= \overline{x^2} + (\bar{x}^2) - 2(\bar{x}^2) = \overline{x^2} - (\bar{x}^2) \qquad \blacksquare && (1.10)
\end{aligned}$$

b) Scarto quadratico medio (Deviazione standard)

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

c) Scarto medio assoluto

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

Cosa fare se 2 serie di dati hanno la stessa varianza?

Esempio 1.6

$$E_3 = \{3, 4, 5, 6, 7\} \qquad \bar{x}_3 = 5 \quad s_3^2 = 2.5$$

$$E_4 = \{13, 14, 15, 16, 17\} \qquad \bar{x}_4 = 15 \quad s_4^2 = 2.5$$

In entrambi i casi un dato rispetto al precedente varia di 1. Ma questa variazione è più importante nelle serie di dati E_3 (che sono numeri più piccoli rispetto ad E_4) che nella serie di dati E_4.

È logico pensare che i dati di E_3 siano più dispersi di quelli di E_4, anche perchè $\bar{x}_4 > \bar{x}_3$. Per questo motivo si definisce un **coefficiente di variazione** (c.v.)

$$c.v. = \frac{s}{\bar{x}}$$

Per E_3 $c.v. = \frac{\sqrt{2.5}}{5}$ per E_4 $c.v. = \frac{\sqrt{2.5}}{15}$. A valori maggiori del $c.v.$ corrisponde una maggiore variabilità dei dati.

Esempio 1.7 *Nella seguente tabella vengono riportati il prezzo (in EURO/LITRO) di un particolare combustile e la frequenza con cui esso viene venduto giornalmente in un distributore.*

prezzi	*14.5*	*16.8*	*12.3*	*10.7*	*11.4*	*18.1*	*20.6*	*13.8*
frequenze	*7*	*5*	*8*	*12*	*10*	*6*	*4*	*11*

Calcolare media, varianza, moda e mediana.

In questo caso:

- la media e' quella pesata (1.2) con x_i i *prezzi* ed i pesi s_i dati dalle $frequenze$.
- La varianza si calcola
$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_p)^2 f_i}{\sum_{i=1}^{k} f_i}$$
dove $k = 8$.
- Per prima cosa si devono ordinare i dati in modo crescente nel prezzo. Sia N la somma delle frequenze. Essendo $N = 63$ (dispari) allora la mediana e' elemento di posto 64/2=32 e quindi pari a 13.8

1.2.1 Dati raggruppati per classi

Supponiamo che gli n dati sono raggruppati in classi C_i $(i = 1, .., k)$, come nell'esempio 1.2. Detto m_i il valore centrale della classe C_i, allora la media e la varianza sono cosi' definite

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} m_i \, f_i \quad , \quad s^2 = \frac{1}{n} \sum_{i=1}^{k} (m_i - \bar{x})^2 \, f_i \tag{1.11}$$

dove f_i e' la frequenza assoluta della classe.

Esempio 1.8 *Calcolare la media e la varianza dell'esempio 1.2.*

In questo caso il numero di dati e' $n = 80$, il numero di classi $k = 5$ e

$$C_1 = (0.40, 1.50] \quad , \quad m_1 = \frac{0.4 + 1.5}{2} = ...$$

$$C_2 = (1.50, 2.30] \quad , \quad m_2 = \frac{1.5 + 2.3}{2} = ...$$

$$C_3 = (2.30, 3.00] \quad , \quad m_3 = \frac{2.3 + 3.0}{2} = ...$$

$$C_4 = (3.00, 4.00] \quad , \quad m_4 = \frac{3 + 4}{2} = ...$$

$$C_5 = (4.00, 5.50] \quad , \quad m_5 = \frac{4 + 5.5}{2} = ...$$

applicando la (1.11) si ottiene ....

Per il calcolo dei quartili si applica l'iterpolazione lineare come nel seguente esempio.

Esempio 1.9 *La tabella che segue mostra la distribuzione di frequenza degli stipendi settimanali di 65 impiegati di una ditta. Calcolare la media, la varianza ed i quartili.*

Tabella

stipendi (classe)	Numero impiegati (frequenza)
50 - 59.99 €	8
60 - 69.99 €	10
70 - 79.99 €	16
80 - 89.99 €	14
90 - 99.99 €	10
100 - 109.99 €	5
110 - 119.99 €	2

Per il calcolo del valore medio e della varianza applichiamo la formula (1.11), dove, per esempio, il valore centrale della prima classe di stipendi e'

$$\frac{50 + 59.99}{2} = 54.995 \quad .$$

Nel nostro esempio, ho delle classi di stipendio con la relativa frequenza, che da' la posizione dei dati. Poiche' le classi sono intervalli numerici, utilizzaremo per il calcolo dei quartili l'interpolazione lineare. Il primo quartile Q_1 e' lo stipendio che occupa la posizione $n \times 0.25 = 65/4 = 16.25$. Per capire qual'e' lo stipendio che corrisponde a questa posizione, bisogna prendere la colonna delle frequenze. La prima classe corrisponde alla posizione 8 e quindi ne servono 8.25 della seconda classe (infatti $8 + 8.25 = 16.25$). Ora l'incremento di stipendio della seconda classe e' 10 €con frequenza di 10 e quindi se x e' l'incremento di stipendio con posizione 8.25 sara' valida la seguente proporzione

$$10 : 10 = x : 8.25 \quad \rightarrow \quad x = 8.25$$

pertanto

$$Q_1 = 59.99 + 8.25 = 68.24 \text{ €} \quad .$$

Analogamente Q_2 e' lo stipendio che occupa la posizione $n \times 0.5 = 65/2 = 32.5$. Dato che le prime due classi hanno frequenze totali pari a 18, dobbiamo prendere 14.5 della terza classe (infatti $18 + 14.5 = 32.5$) . L'incremento di stipendio della terza classe e' 10 con frequenza 16 e quindi se x e' l'incremento di stipendio con posizione 14.5 sara'

$$10 : 16 = x : 14.5 \quad \rightarrow \quad x = 9.06$$

pertanto

$$Q_2 = 69.99 + 9.06 = 79.05 \text{ €} \quad .$$

Analogamente Q_3 e' lo stipendio che occupa la posizione 65 x 0.75 = 48.75. Le prime 4 classi hanno frequenza totale 48, dobbiamo prendere 0.75 della quinta classe, cioe'

$$10 : 10 = x : 0.75 \quad \rightarrow \quad x = 0.75$$

pertanto

$$Q_3 = 89.99 + 0.75 = 90.74 \text{ €} \quad .$$

In conclusione, il 25% degli impiegati guadagna fino a 68.24 €, il 50% fino a 79.05 €, il 75% fino a 90.74 €

1.3 Indici di forma

Due indici statistici numerici che tengono conto della forma di una distribuzione di una serie di dati sono:

a) Asimmetria: E' una misura dello scostamento di una distribuzione dalla simmetria. Se la curva di frequenza di una distribuzione ha una coda piu' lunga a destra del massimo centrale, piuttosto che a sinistra, la distribuzione si dice *positivamente asimmetrica*. Se e' vero il contrario si dice *negativamente asimmetrica*.

Definizione 1.3

Date n osservazioni $x_1, x_2, ..., x_n$ e' detto <u>indice di asimmetria (o skewness)</u>, la quantita'

$$sk = \frac{m_3}{\sqrt{m_2^3}} \quad , \quad m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \quad k = 2, 3, 4, ...$$

dove m_k sono i momenti centrati di ordine k.

Questo indice indica se la distribuzione del campione e' simmetrica rispetto alla media (sk=0): se sk >0 la distribuzione sara' piu' concentrata a sinistra, con una coda piu' lunga a destra, il contrario se sk <0.

Figura 1.1: Figura sinistra: indice di asimmetria: sk > 0 (asimmetria a destra). Figura destra: indice di asimmetria: sk <0 (asimmetria a sinistra)

a) Curtosi: E' una misura dell'appiattimento di una distribuzione di dati rispetto alla distribuzione normale (gaussiana).

Definizione 1.4

Date n osservazioni $x_1, x_2, ..., x_n$ e' detta <u>Curtosi</u>, la quantita'

$$\kappa = \frac{m_4}{m_2^2}$$

Si prova che, se $\kappa > 3$ allora la distribuzione (detta leptocurtica) e' piu' appuntita rispetto alla normale (con code piu' grandi), se $\kappa < 3$ (platicurtica) e' piu' appiattita (con code piu' piccole), infine se $\kappa = 3$ ha la stessa altezza di una normale.

Figura 1.2: La curtosi

1.4 La correlazione tra due serie di dati

Supponiamo di avere due set di dati $\{x_i\}$, $\{y_i\}$ di numerosità n. Ci chiediamo se esiste una certa relazione tra questi dati ovvero se sono tra loro indipendenti. Per rispondere a questa domanda si puo' confrontare le variazioni delle coppie di dati rispetto ai rispettivi valori medi.

$$x_i - \bar{x} \qquad y_i - \bar{y}.$$

È ovvio supporre che esista una relazione di dipendenza tra $\{x_i\}$ e $\{y_i\}$ se $x_i - \bar{x}$, $y_i - \bar{y}$ hanno lo stesso segno.

Quindi, tanto più i prodotti $(x_i - \bar{x})(y_i - \bar{y})$ hanno concordanza di segno, tanto più i dati considerati hanno forte dipendenza. Anche nel caso in cui a valori positivi di $(x_i - \bar{x})$ corrispondono valori negativi di $(y_i - \bar{y})$ o viceversa, denota una forte dipendenza tra i dati considerati.

Invece, se tutti i prodotti $(x_i - \bar{x})(y_i - \bar{y})$ hanno segni diversi, la loro somma risulta piccola in valore assoluto (ovvero tende a zero) e potrebbe esserci indipendenza tra le due serie di dati.

Definizione 1.5 Si definisce *covarianza*

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

Teorema 1.1 *Si prova facilmente che* $c_{x\,y} = \overline{xy} - \overline{x}\,\overline{y}$

Definizione 1.6 Due serie di dati x_i, y_i sono *statisticamente incorrelati* se:

$$c_{x\,y} = 0 \quad .$$

Da quanto detto questo indice è tale che:

i) $c_{x\,y} \in \mathbb{R}$

ii) se $\{x_i\}$, $\{y_i\}$ sono fortemente correlati $c_{x\,y}$ è grande in valore assoluto;

iii) Se $c_{x\,y} > 0$ i due set di dati si dicono *correlati positivamente*, e questo significa che si muovono nella stessa direzione (all'aumentare dell'uno l'altro aumenta e viceversa). Viceversa se $c_{x\,y} < 0$ i due set di dati si dicono *correlati negativamente*.

iv) se $\{x_i\}$, $\{y_i\}$ sono *statisticamente incorrelati*, dal teorema precedente, si ha che $\overline{xy} \simeq \bar{x}\,\bar{y}$ ovvero la media del prodotto delle due serie di dati ($\overline{xy}$) e' uguale al prodotto delle medie delle singole serie di dati ($\bar{x}\bar{y}$). Ma questo non ci assicura che le due serie di dati siano tra loro indipendenti. Vedremo piu' avanti un esempio in cui le due serie di dati hanno covarianza piccola senza essere per questo indipendenti.

1.5 Analisi di regressione per una serie di dati

Assegnato un insieme E di coppie di dati ci domandiamo se esiste un legame funzionale del tipo

$$y = f(x)$$

che descriva bene la relazione tra i dati.

Un'analisi di questo tipo si chiama **analisi di regressione**.

A questo punto come si fa a determinare la f, che al suo interno contiene dei paramentri in modo che questo legame sia buono?

1.5.1 Metodo dei minimi quadrati

Si cerca f tale che sia minima la funzione *residuo*

$$g(f) = \sum_{i=1}^{n} [f(x_i) - y_i]^2$$

1.5.2 Regressione lineare

In questo caso la funzione f e' una retta

$$f(x) = mx + q$$

$$g(m, q) = \sum_{i=1}^{n} [mx_i + q - y_i]^2$$

(incognite m,q!)

Poiché $g(m, q)$ è una funzione di due variabili:

condizione necessaria e sufficiente affinché $P_* = (m_*, q_*)$ sia minimo relativo è che

i)

$$\left.\frac{\partial g}{\partial m}\right|_{P_*} = 0 \qquad , \qquad \left.\frac{\partial g}{\partial q}\right|_{P_*} = 0$$

ii)

$$H(m_*, q_*) = g_{mm} g_{qq} - (g_{mq})^2 > 0$$

iii)

$$g_{qq}(m_*, q_*) > 0$$

dove

$$g_m = \frac{\partial g}{\partial m} = \lim_{h \to 0} \frac{g(m+h,q) - g(m,q)}{h}$$

$$g_{mm} = \frac{\partial^2 g}{\partial m^2} \qquad , \qquad g_{mq} = \frac{\partial^2 g}{\partial m \partial q}$$

$$\frac{\partial g}{\partial m} = 2 \sum_{i=1}^{n} (m x_i + q - y_i) x_i$$

$$\frac{\partial g}{\partial q} = 2 \sum_{i=1}^{n} (m x_i + q - y_i)$$

$$\frac{\partial^2 g}{\partial m^2} = 2 \sum_{i=1}^{n} x_i^2 > 0$$

$$\frac{\partial^2 g}{\partial q^2} = 2n > 0$$

$$\frac{\partial^2 g}{\partial m \partial q} = 2 \sum_{i=1}^{n} x_i > 0 \qquad \text{Provare che } H(m,q) > 0!$$

Risolviamo ora il sistema:

$$\begin{cases} \sum_i (m x_i + q - y_i) x_i = 0 \\ \sum_i (m x_i + q - y_i) = 0 \end{cases}$$

$$\begin{cases} \sum_i m x_i^2 + \sum_i q x_i - \sum_i x_i y_i = 0 \\ \sum_i m x_i + nq - \sum_i y_i = 0 \end{cases}$$

dividendo per n si ottiene:

$$\begin{cases} \dfrac{m}{n}\sum_i x_i^2 + q\overline{x} - \dfrac{1}{n}\sum_i x_i y_i = 0 \\ m\overline{x} + q - \overline{y} = 0 \Rightarrow q = \overline{y} - m\overline{x} \end{cases}$$

$$m\frac{\sum_i x_i^2}{n} + \overline{x}\,\overline{y} - m(\overline{x})^2 - \frac{1}{n}\sum_i x_i y_i = 0 \qquad , \qquad m = \frac{\frac{1}{n}\sum_i x_i y_i - \overline{xy}}{\frac{\sum_i x_i^2}{n} - (\overline{x})^2}$$

Da cui ricordando la covarianza

$$c_{xy} = \frac{1}{n}\sum_i x_i y_i - \overline{x}\,\overline{y}$$

e la varianza per la variabile x

$$s_x^2 = \frac{\sum_i x_i^2}{n} - (\overline{x})^2$$

da cui

$$m = \frac{c_{xy}}{s_x^2} \qquad , \qquad q = \overline{y} - \frac{c_{xy}}{s_x^2}\overline{x}$$

Il metodo applicato fornisce la retta che meglio approssima i dati, ma non il grado di approssimazione. Per questo motivo si introduce il **coefficiente di correlazione lineare (o di Pearson)**

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

Proposizione 1.3

i) $r_{xy} \in [-1, 1]$

ii) se $r_{xy} = \pm 1$ i dati (x_i, y_i) sono perfettamente allineati con la retta di regressione

ii) se $r_{xy} > 0$ la retta è ascendente

NB: nella pratica, se $\mid r_{xy} \mid < 0.9$, i dati si allontanano dall'andamento rettilineo.

Figura 1.3: Indice di correlazione $r = r_{xy}$

Esempio 1.10 *Tabella carichi di rottura* Si deve controllare la resistenza di un campione di 15 travi di cemento, tutte ottenute dalla stessa gittata, misurando sia i carichi di prima lesione X_i che quelli di rottura finale Y_i (in Kg), come in tabella 1.3.

Con EXCEL calcolare: s_x , s_y , c_{xy} , r_{xy} $\overline{x}$, $\overline{y}$, retta di regressione e grafico dei dati x,y con retta di regressione.

Esempio 1.11 *Assegnati i set di dati in tabella 1.4, calcolare la covarienza ed il coefficiente di Pearson. Si osservi che la covarianza tende a zero ma non il coefficiente di Pearson !*

Tabella 1.3: Carichi di rottura.

I^a lesione	rottura
2550	4650
2900	4650
3000	4700
3000	4750
3000	4775
3000	4775
3250	4800
3250	4950
3250	5050
3600	5100
4225	5100
4650	5150
4750	5175
5175	5250
5300	5300

Tabella 1.4: Esempio sulla covarianza ed indipendenza

x	0.185	0.22	0.233	0.247	0.255	0.2745
y	0.049	0.053	0.054	0.0565	0.058	0.0605

1.5.3 Parabola dei minimi quadrati

Regressione non lineare

In questo caso

$$f(x) = a + bx + cx^2$$

a questo punto devo rendere minimo

$$g(a,b,c) = \sum_{i=1}^{n} [a + bx_i + cx_i^2 - y_i]^2$$

allora

$$\frac{\partial g}{\partial a} = 0 \qquad , \qquad \frac{\partial g}{\partial b} = 0 \qquad , \qquad \frac{\partial g}{\partial c} = 0$$

$$\begin{cases} \sum_i y_i = an + b\sum_i x_i + c\sum_i x_i^2 \\ \sum_i x_i y_i = a\sum_i x_i + b\sum_i x_i^2 + c\sum_i x_i^3 \\ \sum_i x_i^2 y_i = a\sum_i x_i^2 + b\sum_i x_i^3 + c\sum_i x_i^4 \end{cases}$$

Sistema di 3 equazioni in 3 incognite

Esempio 1.12 Trovare la soluzione del sistema lineare 3x3 con EXCEL

$$AX = B$$

Se $\det A \neq 0$ (teorema di Cramer) allora

$$\exists\ A^{-1} \quad t.c. \quad A^{-1}AX = A^{-1}B$$

quindi la soluzione e'

$$X = A^{-1}B$$

NB:

A→ n x n , X→ n x 1 , B→ n x 1

Esempio 1.13 *Utilizzando la tabella del carichi di rottura 1.3, calcolare la parabola dei minimi quadrati.*

Un modo grossolano per vedere la bontà dell'approssimazione è calcolare il *residuo*:

$$g(\overline{a}, \overline{b}, \overline{c}) = \sum_i = [\overline{a} + \overline{b}x_i + \overline{c}x_i^2 - y_i]^2$$

- Confrontare il valore ottenuto con la retta di regressione
- Mostrare il grafico dei dati e la parabola di regressione

Esempio 1.14 *Popolazione USA: in tabella 1.5 sono assegnati per alcuni anni la popolazione degli Stati Uniti d'America.*

Si vogliono fare delle stime di crescita di questa popolazione. In particolare:

- Calcolare la retta di regressione e r_{xy};
- Calcolare la parabola di regressione;
- Approssimare i dati con la curva esponenziale

$$y = ae^{bx}$$

In questo caso ci possimao ricondurre al caso della regressione lineare :

$$\ln y = \ln(ae^{bx})$$

$$\ln y = \ln a + \ln e^{bx} = \ln a + bx$$

con un cambiamento di variabili, si ha

$$\tilde{y} = \alpha + bx$$

- Approssimare i dati con una legge esponenziale $y = b\,m^x$ (utilizzare REGR.LOG)
- Grafico dei dati sperimentali con quelli di regressione.

Tabella 1.5: Popolazione USA

Anno	Popol(mln)
1840	17.1
1850	23.2
1860	31.4
1870	39.8
1880	50.2
1890	62.9
1900	76.0
1910	92.0
1920	105.7
1930	122.8
1940	131.7
1950	151.1
1960	179.3

Suggerimento: scegliere x in modo che:

. . . . . .

anno 1890 $\rightarrow$ x=-1

anno 1900 $\rightarrow$ x=0

anno 1910 $\rightarrow$ x=1

anno 1920 $\rightarrow$ x=2

. . . . . .

Figura 3.13: Distribuzione T di Student

3.11 Legge dei grandi numeri

Hp:

Supponiamo di avere una successione $\{X_i\}$ di variabili aleatorie statisticamente indipendenti con identica funzione di ripartizione. Definiamo la nuova variabile aleatoria:

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \qquad (media \quad campionaria)$$

supponiamo inoltre che

$$E[X_i] = \mu \quad , \quad V[X_i] = \sigma^2 \qquad i = 1, \ldots, n \tag{3.15}$$

Ts:

$$\overline{X}_n \xrightarrow{d} M$$

dove

a) M è una variabile aleatoria che assume il valore μ con probabilità 1

b) $\xrightarrow{d}$ è un particolare tipo di convergenza

Sia $\begin{cases} X_n & \text{variabile aleatoria con funzione di ripartizione } F_n \\ X & \text{variabile aleatoria con funzione di ripartizione F} \end{cases}$

allora $X_n \xrightarrow{d} X \;\Leftrightarrow\; \lim_{n\to\infty} F_n = F$

La legge dice che medie molto grandi di variabili aleatorie tendono alla media vera, ovvero

$$E[\overline{X}_n] \longrightarrow \mu$$

3.12 Teorema del limite centrale

La legge dei grandi numeri ci dice che $\overline{X}_n \longrightarrow M$, ma non ci dice con quale rapidità (cioè a partire da quale valore di n).

Sotto le stesse ipotesi della legge dei grandi numeri, allora:

$$\overline{X}_n \xrightarrow{d} X \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{3.16}$$

cioè per n grande $\overline{X}_n$ si distribuisce come una normale $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, i cui parametri dipendono da quelli delle variabili X_i (cioè da μ e σ).

Dim. Dalla definizione di variabili aleatorie statisticamente indipendenti (2.11), dalle (2.8) e (3.15) si ha

$$V[\bar{X}_n] = V\left[\sum_{i=1}^{N} \frac{X_i}{n}\right] = \frac{1}{n^2} V\left[\sum_{i=1}^{N} X_i\right] = \frac{1}{n^2} \sum_{i=1}^{N} V\left[X_i\right] = \frac{1}{n^2} \sum_{i=1}^{N} \sigma^2 = \frac{\sigma^2}{n}$$

e quindi la deviazione standard della media campionaria e' pari a $\sigma/\sqrt{n}$. Per dimostrare che la media campionaria si distribuisce come una normale di parametri μ e $\sigma/\sqrt{n}$, bisogna sfruttare la disequazione di Chebychev, che esula dai nostri scopi.

Nella pratica il teorema del Limite Centrale ci dice che per *n sufficientemente grande*, la variabile aleatoria media campionaria ha come funzione di distribuzione una normale di parametri μ e $\sigma/\sqrt{n}$, indipendentemente dalla distribuzione della popolazione. Cio' accade per **n $\geq$ 30**.

Capitolo 4

Stime di parametri

La statistica *inferenziale* consente di dedurre caratteristiche particolari di una popolazione analizzando un numero finito e piccolo (preferibilmente) di suoi individui detto *campione*. Quando le caratteristiche che si vogliono individuare sono esprimibili numericamente, allora prendono il nome di *parametri*.

4.1 Problema del campionamento

Devo determinare le caratteristiche di una popolazione con un numero limitato di individui. Infatti, se la popolazione è vasta, per risparmiare tempo e denaro conviene analizzare un piccolo campione.

4.1.1 Strategie di campionamento

i) Casuale: Associo ad ogni individuo un numero e con un generatore di numeri casuali ne estraggo un certo numero.

Attenzione: se la statistica va fatta sugli abitanti di una città, non si deve fare il campionamento ad esempio tra abbonati al telefono o quelli che si incontrano per strada. Non si prenderebbero in considerazione chi non ha telefono e chi esce poco.

ii) Stratificato: La popolazione è suddivisa in gruppi con stesse caratteristiche (es. età, sesso, etc..)

ii) A grappoli: Si suddivide la popolazione in gruppi *eterogenei*, in modo che ogni singolo gruppo rappresenti l'intera popolazione.

Sia X il carattere della popolazione su cui si è interessati a fare delle inferenze (es. peso, altezza, etc... su una popolazione di persone). Il valore assunto da questo carattere varia a seconda dell'individuo considerato e viene indicato con x. Quindi X è una variabile aleatoria con *distribuzione sconosciuta*, che corrisponde a quella che si otterrebbe facendo ricorso alle tecniche della statistica descrittiva e potendo quindi utilizzare l'intera popolazione.

Definizione 4.1 - *Campione casuale di numerosità n*

É una n-upla $(X_1, X_2, \ldots, X_n)$ di variabili aleatorie indipendenti (estratte da una popolazione) aventi ognuna la stessa distribuzione del carattere X della popolazione. I valori assunti da questa n-upla

$$(x_1, x_2, \ldots, x_n)$$

sono le misure fatte e sono dette *realizzazioni* di $(X_1, X_2, \ldots, X_n)$.

Definizione 4.2 - *Parametro e stima*

Un *parametro* è un valore numerico che descrive una caratteristica di una popolazione, ed è una grandezza associata ad una sua distribuzione (quale il valore atteso e la varianza).

Una *stima* del parametro è una misura fatta sul campione.

Esempio 4.1

X="costo al mq degli appartamenti della città"

Sia $(X_1, X_2, \ldots, X_{80})$ un campione di numerosità 80. Si considera il parametro valore atteso:

μ="costo medio al mq degli appartamenti della città"

Ovviamente non si conosce μ, perchè non si hanno i dati relativi a tutti gli appartamenti della città.

Una stima del parametro μ puo' essere fatta con la media sugli 80 valori

$$\overline{X}_n = \frac{1}{80}\sum_i^n x_i$$

presumibilmente il valore vero di μ sarà diverso da $\overline{X}_n$.

Allora le stime sono anch'esse delle variabili aleatorie definite in funzione del campione

$$H_n = f(X_1, \ldots, X_n)$$

Esse prendono il nome di *statistiche campionarie* e le loro distribuzioni sono chiamate *distribuzioni campionarie.*

4.2 Principali distribuzioni campionarie

Sia X il carattere di una popolazione con distribuzione cumulativa F, valore atteso μ, varianza σ sconosciuti. Vediamo come stimare questi due parametri incogniti.

Definizione 4.3 - *Media campionaria n-esima di un campione casuale*

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} \tag{4.1}$$

Proposizione 4.1

Se le variabili aleatorie X_i del campione sono tutte indipendenti e

$$E[X_i] = \mu, \qquad V[X_i] = \sigma^2 \qquad i = 1, n$$

allora

$$E[\overline{X}_n] = \mu, \qquad V[\overline{X}_n] = \frac{\sigma^2}{n} \qquad (dimostrare) \tag{4.2}$$

In tal caso il valore atteso $\overline{X}_n$ non dipende dalla numerosità del campione, mentre la sua varianza ne dipende. Questo significa che la media campionaria sara' piu' vicina al valore incognito μ quanto piu' e' grande la numerosita' del campione.

Osservazione:

Per la legge dei grandi numeri ed il teorema del limite centrale, per $n \geq 30$, $\overline{X}_n$ si può approssimare con una variabile aleatoria avente distribuzione normale di parametri μ e $\dfrac{\sigma}{\sqrt{n}}$:

$$\overline{X}_n \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Definizione 4.4 - *Varianza campionaria n-esima*

Sia $(X_1, \ldots, X_n)$ *un campione estratto da una popolazione avente distribuzione* F*, media* μ *e deviazione standard* σ*. Si definisce* varianza campionaria n-esima*:*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad . \tag{4.3}$$

La distribuzione di S_n^2 si chiama *distribuzione della varianza campionaria n-esima.*

Si prova che

$$E[S_n^2] = \frac{n-1}{n}\ \sigma^2 \tag{4.4}$$

$$V[S_n^2] = \frac{1}{n}\left(E[X^4] - \frac{n-3}{n-1}\ \sigma^4\right)$$

Ancora per il teorema del limite centrale si prova che per $n \geq 30$, la sua distribuzione si può approssimare con

$$N(\overline{\mu}, \overline{\sigma})$$

$$\overline{\mu} = \frac{n-1}{n}\ \sigma^2$$

$$\overline{\sigma} = V[S_n^2] \quad .$$

A causa del fattore $\dfrac{n-1}{n}$ in (4.4), si preferisce considerare una nuova statistica

Definizione 4.5 - *Varianza campionaria n-esima corretta*

$$\widehat{S}_n^2 \ = \ \frac{n}{n-1}\ S_n^2 \ = \ \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2 \tag{4.5}$$

da cui ovviamente

$$E[\widehat{S}_n^2] = \sigma^2 \quad . \tag{4.6}$$

Abbiamo visto che una stima del valore atteso μ di una popolazione e' data dalla media campionaria $\overline{X}_n$, mentre una stima di σ e' data dalla varianza campionaria S_n^2 o quella corretta $\widehat{S}_n^2$.

FUNZIONI EXCEL 4.1

- DEV.ST $= \widehat{S}_n = \sqrt{\dfrac{n}{n-1}}\, S_n$ da usare per un campione
- DEV.ST.POP $= S_n = \sqrt{\dfrac{1}{n}\sum_i (X_i - \overline{X}_n)^2}$ da usare su tutta la popolazione

4.3 Stimatori puntuali

Sia θ un parametro incognito di una popolazione X. Estratto dalla popolazione un campione di numerosita' n, chiamo **estimatore puntuale** $\hat{\theta}$ un numero costruito a partire dalle realizzazioni $(x_1, x_2, ..., x_n)$ del campione casuale. Per esempio la media campionaria e' un estimatore puntuale del parametro valore di aspettazione, come anche la varianza campionaria.

Definizione 4.6 - *Estimatore corretto o non-distorto (unbiased)*
Un estimatore θ di una variabile aleatoria si dice ***corretto o non-distorto (unbiased)****, se il suo valore di aspettazione coincide con il valore vero.*

Dalle proprieta' $(4.2)_1$, (4.6) ne segue che la media campionaria e la varianza campionaria corretta sono estimatori non-distorti, mentre la varianza campionaria e' un estimatore distorto.

Osserviamo che per stimare un parametro θ possiamo definire diversi estimatori corretti. Un criterio per stabilire quale sia preferibile e' il seguente

Definizione 4.7 *Siano $H_{1,n}, H_{2,n}$ due estimatori corretti del parametro θ di una popolazione. Allora diremo che H_1 e' piu' efficiente di H_2 se vale*

$$V[H_{1,n}] \leq V[H_{2,n}]$$

Esempio 4.2 *Si considerino i seguenti estimatori della media μ di una popolazione*

$$H_{1,n} = \frac{X_1 + X_2 + ... + X_n}{n} \quad , \quad H_{2,n} = \frac{X_1}{n} + \frac{X_2 + X_3 + ... + X_n}{2(n-1)}$$

proviamo che il primo e' piu' efficiente del secondo.

Si osservi che il secondo estimatore da' piu' importanza alla prima componente del campione, a cui e' assegnato un peso di valore $\frac{1}{2}$ invece di $\frac{1}{2(n-1)}$. Supposto che tutte le variabili aleatorie del campione siano indipendenti (si veda (4.2)), e' facile vedere che i due estimatori sono corretti, cioe'

$$E[H_{1,n}] = E[H_{2,n}] = \mu$$

mentre

$$V[H_{1,n}] = \frac{\sigma^2}{n} \quad \leq \quad V[H_{2,n}] = \left[\frac{1}{4} + \frac{1}{4(n-1)}\right]\sigma^2 \quad \square$$

4.3.1 Altri metodi

Per effettuare delle stime puntali di parametri esistono altri metodi. Quelli principali sono il *metodo dei momenti* e di *massima verosimiglianza.* Nel seguito considereremo solo il secondo metodo con il seguente esempio

Esempio 4.3 *Supponiamo di estrarre da una popolazione X distribuita secondo un'Esponenziale (3.4) di parametro incognita λ, il campione*

$$(3.0, 4.1, 2.8, 5.5, 1.5, 2.2, 6.01.2, 3.2, 0.9) \quad .$$

Vogliamo stimare λ con il metodo di massima verosimiglianza.

Siccome i campioni sono indipendenti, allora

$$\mathcal{L}(x_1, ..., x_{10}, \lambda) = \prod_{i=1}^{10} \lambda e^{-\lambda x_i} = \lambda^{10} e^{-30.4\lambda}$$

La stima del parametro si ottiene determinando quel valore di λ per cui risulta massima la funzione $\mathcal{L}$, ovvero

$$\mathcal{L}' = \lambda^9 e^{-30.4\lambda}(10 - 30.4\lambda) = 0 \quad \rightarrow \quad \lambda = 0.329$$

Osservazione

Nel nostro caso la media campionaria $\bar{x} = 3.04$. Per una variabile aleatoria con distribuzione esponenziale sappiamo che

$$E[X] = \frac{1}{\lambda}$$

e se poniamo $E[x] \simeq \bar{x}$ otteniamo ancora $\lambda = 0.329$. Questa osservazione pero' non puo' essere generalizzata a tutte le stime ottenute con il metodo di massima verosimiglianza.

Esempio 4.4 *Supponiamo di effettuare N misure della stessa grandezza fisica x_i, che siano tra loro statisticamente indipendenti ed inoltre affette da errori casuali distribuiti secondo la legge di Gauss. La densità di probabilità corrispondente all'evento casuale costituito dall'osservazione degli N valori (applicando il teorema della probabilità composta)*

$$\prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{x^\star - x_i}{2\sigma_i^2}\right)$$

dove $x^\star$ e' il valore vero (incognito) e σ_i gli errori quadratici medi (noti). Detta funzione di verosimiglianza

$$\mathcal{L}(x_1, ..., x_N, x) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{x - x_i}{2\sigma_i^2}\right)$$

la stima più verosimile di $x^\star$ è quella che rende massima L, rispetto alla variabile incognita x

Si prova che questo massimo esiste ed e' unico e vale

$$\bar{x} = \frac{1}{K}\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \quad , \quad K = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{4.7}$$

4.4 Campionamento da una distribuzione normale

Consideriamo una popolazione normalmente distribuita $N(\mu, \sigma)$ ed estraiamo da essa un campione di numerosita' n. Valgono i seguenti teoremi:

Teorema 4.1 *Sia $(X_1, .., X_n)$ un campione estratto da $N(\mu, \sigma)$, allora la media campionaria $\overline{X}_n$ e' ancora distribuita secondo una normale ma con parametri μ e $\sigma/\sqrt{n}$, ovvero*

$$\overline{X}_n \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad . \tag{4.8}$$

Osservazione

Notiamo che il risultato (4.8) vale per qualsiasi n. Un simile risultato e' ottenuto con il teorema del limite centrale (3.16), dove pero' l'unica ipotesi e' che $n \geq 30$.

Teorema 4.2 *Sia $(X_1, .., X_n)$ un campione estratto da $N(\mu, \sigma)$ di cui supponiamo di non conoscere μ. Allora la varianza campionaria S_n^2 e' tale che*

$$S_n^2 \simeq \frac{\sigma^2}{n-1}\chi^2_{n-1} \quad . \tag{4.9}$$

Teorema 4.3 *Sia $(X_1, .., X_n)$ un campione estratto da $N(\mu, \sigma)$. Allora la variabile aleatoria*

$$T_n = \frac{\overline{X}_n - \mu}{\frac{\widehat{S}_n}{\sqrt{n}}} \simeq t_{n-1} \tag{4.10}$$

Osservazione

Per calcolare T_n occorre conoscere μ (media su tutta la popolazione) ed anche la media e la varianza campionaria, ma non la σ. Inoltre T_n e' distribuita secondo una t di Student che ha come solo parametro n (cioè il grado di libertà),e che quindi non dipende più dai parametri μ e σ.

4.5 Stime intervallari

La stima di un parametro di una popolazione data da un solo numero è detta *stima puntuale* (es. valore medio), mentre se è data da 2 numeri si dice *stima intervallare*.

Ad esempio: "La misura di una distanza è 5,28 mt (stima puntuale)"; "La misura di una distanza è compresa tra gli estremi 5,28±0,03 mt (stima intervallare)".

Sia θ un parametro puntuale di una popolazione (sconosciuta) e $\widehat{\theta}$ una sua realizzazione. Un intervallo del tipo

$$I = [\widehat{\theta} - e_1, \widehat{\theta} + e_2] \subseteq \mathbb{R}$$

conterrà il valore θ con maggiore o minore probabilità a seconda dell'ampiezza.

se e_1, e_2 sono grandi, $P(\theta \in I) \simeq 1$

se $e_1 \simeq e_2 \simeq 0$, $P(\theta \in I) \simeq 0$

Definizione 4.8 - *Intervallo di confidenza per il parametro θ*

Fissato $\alpha \in [0, 1]$, si chiama *intervallo di confidenza con livello di fiducia α*, quell'intervallo:

$$[\widehat{\theta} - e_1, \widehat{\theta} + e_2]$$

tale che

$$P(\theta \in [\widehat{\theta} - e_1, \widehat{\theta} + e_2]) = 1 - \alpha$$

solitamente

$$\alpha = 0,1 \ (90\%) \qquad \alpha = 0,05 \ (95\%) \qquad \alpha = 0,01 \ (99\%)$$

A questo punto si costruiscono intervalli di confidenza per i parametri *media* e *varianza* di una popolazione

Convergenza di una distribuzione

Si dice che la successione di variabili aleatorie $\{X_n\}$ converge in distribuzione alla variabile aleatoria X. $X_n \xrightarrow{d} X$ se

$$\lim_{n \to +\infty} F_{X_n}(x) = F_X(x)$$

per tutti i valori di x dove $F_X(x)$ è continua

Def Convergenza in probabilità)

Si dice che la successione di variabili aleatorie $\{X_n\}$ converge in probabilità alla variabile aleatoria X se

$$\forall \varepsilon > 0 \quad \lim_{n \to +\infty} P[\,|X_n - X| < \varepsilon\,] = 1$$

$X_n \xrightarrow{P} X$ in simboli

Convergenza quasi certa

Si dice che la successione di variabili aleatorie $\{X_n\}$ converge quasi certamente alla var. X se

$$\forall \varepsilon > 0 \quad P[\,\lim_{n \to \infty} |X_n - X| < \varepsilon\,] = 1$$

$X_n \xrightarrow{q.c} X$ in simboli

Convergenza in media quadratica

$X_n \xrightarrow{m.q.} X$ se $\lim_{n \to +\infty} E[(X_n - X)^2] = 0$

Si dimostra che

$X_n \xrightarrow{q.c.} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$

$X_n \xrightarrow{m.q.} X$

Teorema **Legge debole dei grandi numeri**

Sia X_n una successione di variabili aleatorie iid provenienti da una popolazione X data di media $E[X] = m$ allora:

$\forall \varepsilon > 0 \quad \lim_{n \to \infty} P[|\bar{X}_n - m| < \varepsilon] = 1$ cioè la successione delle medie campionarie converge in probabilità ad m, il valore atteso di tutte le X_n

Teorema **Legge forte dei grandi numeri**

Stesse Hyp del precedente. Segue

$P[\lim_{n \to \infty} |\bar{X}_n - m| < \varepsilon] = 1$

Teorema centrale del limite

Sia X_n una successione di variabili aleatorie c.i.d provenienti da una popolazione X data di media e di varianza $(E[X] = m, Var[X] = v)$ allora per ogni $x \in \mathbb{R}$ segue

$$\lim_{n \to \infty} P\left[\frac{X_n - m}{\sqrt{\frac{v}{n}}} \leq x\right] = \phi(x)$$ ← Distr normale standardizzata

cioè la successione delle medie campionarie standardizzate converge in legge ad una $Nor(\mu = 0, \sigma^2 = 1)$

Dal teorema si capisce che $\bar{X}_n \sim Nor(\mu = m, \sigma^2 = \frac{V}{n})$

Osservazioni:

- $Bin(n, p) \approx Nor(\mu = np, \sigma^2 = np)$
  se $n \geq 30$, $np \geq 5$, $nq \geq 5$
- $Bin(n, p) \approx Nor(\mu = n/p, \sigma^2 = \frac{n(1-p)}{p^2})$
  se $n \geq 30$
- $Poi(\lambda) \approx Nor(m = \lambda, \sigma^2 = \lambda)$ se $\lambda \geq 30$

~~Definizione~~

Definizione di statistica

Si dice statistica di una variabile aleatoria T, funzione nota delle variabili aleatorie $X_1, \dots, X_n$

~~tale che~~ $T(X_1, \dots, X_n) = T(X)$ $X \equiv$ popolazione

Definizione di Stimatore

Data una popolazione dipendente da un parametro θ e un campione casuale $X_1 \dots X_n$, si dice stimatore una statistica T che non dipende da θ e che serva per stimare θ

Definizione **Correttezza**, **correttezza asintotica**

Uno stimatore si dice corretto (o non distorto) se il suo valore atteso coincide con la grandezza da stimare

$$E[T_n] = \theta$$

Si dirà asintoticamente corretto se

$$E[T_n] \longrightarrow \theta \quad \text{per } n \to \infty$$

Definizione di **consistenza**

Uno stimatore si dice consistente in media quadratica se

$$E[(T_n - \theta)^2] \xrightarrow{n\to\infty} 0$$

Si dirà consistente in probabilità se

$$P[|T_n - \theta| < \varepsilon] \xrightarrow{n\to\infty} 0 \quad \forall \varepsilon > 0$$

Definizione di **Asintotica normalità**

$$\frac{T_n - E[T_n]}{\sqrt{Var[T_n]}} \overset{n\to\infty}{\sim} Nor(0,1)$$

Esempio

$\theta > 0$ Supponiamo X abbia densità

$$f_X = \begin{cases} \theta x^{\theta-1} & 0 \le x \le 1 \\ 0 & \text{se no} \end{cases} \quad \text{con } \theta > 0$$

Vogliamo stimare θ (metodo dei momenti)

$$E[X] = \int_0^1 x\,\theta x^{\theta-1}\,dx = \frac{\theta}{\theta+1}$$

Da cui $\theta = \dfrac{E[X]}{1-E[X]}$

Lo stimatore di θ è $\hat{\theta} = \dfrac{\bar{X}_n}{1-\bar{X}_n}$

Esempio

$$f_X(x) = \begin{cases} \frac{2\theta+1}{2}(x^2)^{\theta} & -1 \le x \le 1 \\ 0 & \text{se no} \end{cases} \quad \theta \ge 0$$

$E[X] = 0$ per l'integrale

$$E[X^2] = \int_{-1}^{1} x^2 \frac{2\theta+1}{2}(x^2)^{\theta}\,dx = \frac{2\theta+1}{2\theta+3}$$

$$\theta = \frac{3E[X^2]-1}{2(1-E[X^2])}$$

NB $M_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

$$\hat{\theta} = \frac{3M_n^2-1}{2(1-M_n^2)}$$

Esercizi

1) $X \sim Ber(p)$
2) $X \sim Poi(\mu)$

Stimare i parametri p e μ

Metodo della massima verosimiglianza

$X_1, X_2 \dots X_n$

$$L(X_1, X_2 \dots X_n, \theta) = P_X(X_1, \theta)\, P_X(X_2, \theta) \dots P(X_n, \theta)$$
$$= \prod_{i=1}^{n} P_X(x_i, \theta)$$

Esempio

$$f_X(x) = \begin{cases} \theta x^{\theta-1} & 0 \le x \le 1 \\ 0 & \text{se no} \end{cases} \qquad \theta > 0$$

$$L(\theta) = (\theta x_1^{\theta-1})(\theta x_2^{\theta-1}) \dots (\theta x_n^{\theta-1}) =$$
$$= \theta^n \left(\prod_i x_i\right)^{\theta-1}$$

$$\frac{d}{d\theta} L(\theta) = \frac{n}{\theta} + \ln\left[\prod_i x_i\right] = 0$$

0 per essere punto stazionario

$$\theta = -\frac{n}{\ln(\prod_i x_i)} \qquad \theta = -\frac{n}{\sum_i \ln x_i}$$

Stimatore

Con la derivata seconda si capisce se è un massimo

Esercizi

1) $P_X(x) = \theta^x (1-\theta)^{1-x}$
2) $Poi(\mu)$
3) $Exp(\lambda)$
4) $Norm(\mu, \sigma^2)$

Stimare i parametri

Verifica di ipotesi

Struttura di un test di ipotesi

- Ipotesi: quella da considerare sulla base del campione
  L'ipotesi di partenza è detta ipotesi nulla ed è indicata con H_0

  Esempio:

  $H_0: \mu = 5$ questa è detta ipotesi semplice

  $H_0: \mu \leq 5$ OR $H_0: \mu \geq 5$ ipotesi composta

  Il contrario dell'ipotesi nulla è l'ipotesi alternativa e sarà indicata con H_1 o H_a

  Si tratta di ciò che dovremo accettare per vero nel caso in cui si rifiutasse l'ipotesi nulla

- Statistica test (U)
  Si tratta di una variabile aleatoria della quale almeno approssimativamente possiamo specificare la distribuzione.

  Indicheremo con u il valore di U

- regione ~~della~~ critica (C)
  Si rifiuta H_0 se $u \in C$
  Non si rifiuta H_0 se $u \notin C$

- Errori possibili
  Gli errori che si possono commettere sono i seguenti:

  Se H_0 è vera $\begin{cases} \text{Se non si rifiuta } H_0 \text{ non si sbaglia} \\ \text{Se si rifiuta } H_0 \text{ commettiamo un errore detto del 1° tipo} \end{cases}$

  Se H_0 è falsa $\begin{cases} \text{Se non si rifiuta} \Rightarrow \text{errore 2° tipo} \\ \text{Se si rifiuta non si sbaglia} \end{cases}$

- Livello di significatività (α) — margine di errore

Livello di soglia α tale che la probabilità di commettere l'errore del 1° tipo sia non superiore ad α.

Solitamente $\alpha = 10\%$, 5% o 1%

- Funzione potenza

Se il test è un test parametrico, cioè riguarda un parametro, la probabilità di rifiutare l'ipotesi dipende dal valore incognito del parametro (θ). Questa probabilità è detta funzione potenza del test

$$\pi(\theta) = P_\theta[\text{rifiutare } H_0]$$

oppure $\pi(\theta) = P_\theta[u \in C]$

- p-value — Valore sulla frontiera di C

Ci si può chiedere a quale livello di significatività corrisponderebbe una regione critica la cui frontiera fosse costituita proprio da u, cioè per quale livello il valore osservato di u valica la frontiera di C. Questo valore prende il nome di p-value

$$\text{p-value} = \sup\{\alpha : u \in C\} = \inf\{\alpha : u \in C\}$$

Dato un certo test, rifiuteremo H_0 per tutti i livelli maggiori del p-value e non lo rifiuteremo per tutti i livelli inferiori

Test riguardanti una sola popolazione

- Test riguardanti la media di una normale

Supponiamo $X \sim Nor(\mu, \sigma^2)$ e sia n il numero di osservazioni di tale variabile aleatoria

Tipologie di ipotesi che si possono fare su μ:

Tipo I $\begin{cases} H_0: \mu = \mu_0 & \text{Neutro} \\ H_1: \mu \neq \mu_0 \end{cases}$

Tipo II $\begin{cases} H_0 : \mu = \mu_0 \text{ o } \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$ favore del produttore

Tipo III $\begin{cases} H_0 : \mu = \mu_0 \text{ o } \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$ favore del consumatore

La statistica test U vista è diversa se σ^2 è nota o incognita.

Nei 2 casi troviamo anche la regione critica e il p-value

Caso 1: σ^2 nota, z-test

$$\bar{X}_n \sim Nor(\mu, \sigma^2 = V)$$

Si può scrivere

$$\frac{\bar{X}_n - \mu}{\sqrt{V/n}} \sim Nor(0,1)$$

Scegliamo come statistica la standardizzazione che sarebbe corretta eseguire se fosse vera H_0 ossia scegliamo per μ il valore μ_0

$$U = \frac{\bar{X}_n - \mu_0}{\sqrt{V/n}}$$

Scriviamo $U \overset{H_0}{\sim} Nor(0,1)$ U segue la Nor(0,1) se H_0 è vera

1) Tipo I, dobbiamo decidere tra $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$

Propenderemo per H_0 se il valore osservato di U sarà vicino a 0, mentre propenderemo per H_1 se lo si troverà in una delle 2 code della distribuzione

Si deve stabilire il valore di $\bar{z}$ in corrispondenza del quale si inizia a rifiutare H_0

Poiché il nostro test ha livello di significatività α vogliamo che

$$\alpha = \max P[\text{errore di I tipo}] =$$
$$= \max P[|U| > \bar{z} \mid H_0 \text{ vera}]$$

cioè α deve essere il valore massimo per cui U che nell'ipotesi è una normale standard, finisca in una delle 2 code

$$\begin{cases} U \sim Nor(0,1) \\ P[|U| > \bar{z}, \geq \alpha/2] \end{cases} \Rightarrow \bar{z} = z_{1-\alpha/2}$$

quindi rifiutiamo H_0 se $|u| > z_{1-\alpha/2}$

2) tipo II, dobbiamo decidere tra $H_0: \mu = \mu_0$ e $H_1: \mu > \mu_0$

Propenderemo per H_0 se il valore osservato u di U sarà "vicino a 0", mentre il fatto di avere utilizzato nella creazione della statistica test un valore "troppo piccolo" per la media ci farà propendere per H_1, se u si troverà nella coda a destra

Con conti analoghi ai precedenti si trova $\bar{z} = z_{1-\alpha}$ e pertanto ~~H_0 se~~ rifiuteremo H_0 se $u > z_{1-\alpha}$

3) tipo III

Analogo al tipo II con $H_1: \mu < \mu_0$

Può qui rifiuteremo H_0 se $u < -z_{1-\alpha}$ e qui allora u si troverebbe nella coda a sx se H_0 è rifiutata.

Tipo	rifiuto di H_0 se
I	$\lvert u \rvert > z_{1-\alpha/2}$
II	$u > z_{1-\alpha}$
III	$u < -z_{1-\alpha}$

Tabella riassuntiva per la normale a varianza nota

Esempio

Le batterie di una certa marca hanno durata minima di 22 ore.

Sappiamo che σ del loro tempo di vita è 3,5 ore.

Esaminando 20 batterie è stata riscontrata una media di 20,7 ore.

Supponendo che le ~~variabili~~ durate delle batterie siano variabili aleatorie normalmente distribuite, al livello di significatività del 5% possiamo concludere che la durata media delle batterie sia inferiore a quanto dichiarato?

Soluzione

X = "durata batteria", $X \sim Nor(\mu, \sigma^2 = 3{,}5^2)$

$H_0: \mu = 22 \qquad H_1: \mu < 22$

Con $\alpha = 5\%$, H_0 è rifiutata se $u < -z_{0,95}$

Si ricava che $z_{0,95} = +1{,}645$

$$u = \frac{20{,}7 - 22}{\sqrt{\sigma^2/20}} = -1{,}661 < -z_{0,95}$$

Rifiutiamo H_0, quindi la durata data di 22 ore non è vera

Caso 2: σ^2 incognita — t-test

$$U = \frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}} \sim T(\nu = n-1)$$

Tipo	Rifiuto H_0 se
I	$\lvert u \rvert > t_{1-\alpha/2}$
II	$u > t_{1-\alpha}$
III	$u < -t_{1-\alpha}$

Tabella riassuntiva per la normale a varianza incognita

Esempio

Stesso caso precedente ma con varianza incognita

$$u = \frac{20,7 - 22}{\sqrt{5_n^2/20}} = -1,661$$

$t_{0,95} = 1,7291$

$u \nless -t_{0,95}$

Non rifiutiamo H_0, la accettiamo

Test riguardanti la varianza di una normale

Tipi di ipotesi: $(\sigma^2 \leq \sigma_0^2)$

Tipo I $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases}$ Tipo II $\begin{cases} H_0 \quad \sigma^2 = \sigma_0^2 \\ H_1 \quad \sigma^2 > \sigma_0^2 \end{cases}$

Tipo III $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \quad (\text{o } \sigma^2 \geq \sigma_0^2) \\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$

Caso 1: μ nota

$$S_0^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

Sappiamo che $S_0^2 \sim \frac{\sigma^2}{n} \chi^2(\nu = n)$

$$\Updownarrow$$

$$U = \frac{n}{\sigma^2} S_0^2 \overset{H_0}{\sim} \chi^2(\nu = n) \quad \Big| \quad \frac{n}{\sigma^2} S_0^2 \sim \chi^2(\nu = n)$$

Se siamo nel Tipo I dobbiamo decidere tra
$H_0: \sigma^2 = \sigma_0^2$ e $H_1: \sigma^2 \neq \sigma_0^2$

- H_0 se valore osservato di U si trova in corrispondenza della regione dove la densità della distribuzione χ^2 è maggiore.
- H_1 se u si trova in una delle code della distribuzione

NB: χ^2

Non è simmetrica

$\bar{C}_1 = \chi^2_{\alpha/2;\, n}$ $\qquad$ $\bar{C}_2 = \chi^2_{1-\alpha/2;\, n}$

Rifiutiamo H_0 in uno dei 2 casi

- $u < \chi^2_{\alpha/2}$
- $u > \chi^2_{1-\alpha/2}$

Se siamo nel tipo II si deve decidere tra
$H_0: \sigma^2 = \sigma_0^2$ e $H_1: \sigma^2 > \sigma_0^2$
$(\sigma^2 \leq \sigma_0^2)$

Rifiutiamo H_0 se $u > \chi^2_{1-\alpha}$

Se siamo nel Tipo III dobbiamo decidere tra
$H_0: \sigma^2 = \sigma_0^2$ e $H_1: \sigma^2 < \sigma_0^2$
$(\sigma^2 \geq \sigma_0^2)$

Rifiutiamo H_0 se $u < \chi^2_{\alpha}$

Caso 2; μ incognita

$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2 (\nu = n-1)$$

$$U = \frac{n-1}{\sigma^2} S_n^2 \overset{H_0}{\sim} \chi^2 (\nu = n-1)$$

III

Rifiuto H_0 se

Tipo I: $u < \chi^2_{\alpha/2}$ o $u > \chi^2_{1-\alpha/2}$

II: $u > \chi^2_{1-\alpha}$

III: $u < \chi^2_{\alpha}$

Esempio

Una produzione di dischi di ferro il cui diametro X è distribuito con una normale e non presenta problemi si sa che $\alpha = 5\%$

$$X \sim Nor(\mu = 7,\ \sigma^2 = 0{,}0001)$$

Su un campione si misurano 10 dischi e si ottengono i seguenti valori

7,065 6,975 7,036 7,002 6,976 6,972
7,051 7,026 7,021 6,978

È lecito considerare che la produzione funzioni correttamente

$$\begin{cases} H_0: \mu = 7 \\ H_1: \mu \neq 7 \end{cases} \qquad \begin{cases} H_0: \sigma^2 = 0{,}0001 \\ H_1: \sigma_0^2 \neq 0{,}0001 \end{cases}$$

$\bar{X}_{10} = 7{,}009$

$S^2_{10} = 0{,}0011$

$$\rightarrow u_x = \frac{\bar{X}_n - \mu_0}{\sqrt{S_{10}^2/n}} = 0{,}8581$$

$$\rightarrow u_1 = \frac{n-1}{\sigma_0^2} S_n^2 = 99$$

$T(\nu = 9)$

$\chi^2_{0{,}99} = 81{,}6660 \rightarrow$ rifiuto H_0

Ho non rifiuta

Test asintotici su p probabilità di successo di un processo di Bernoulli

Se $X \sim Ber(p)$ per n grande

$$\bar{X}_n \approx Nor\left(\mu = p,\ \sigma^2 = \frac{p(1-p)}{n}\right)$$

Tipo I $\begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases}$

Tipo II $\begin{cases} H_0: p = p_0 \text{ oppure } p \leq p_0 \\ H_1: p > p_0 \end{cases}$

Tipo III $\begin{cases} H_0: p = p_0 \text{ oppure } p \geq p_0 \\ H_1: p < p_0 \end{cases}$

$$U = \frac{\bar{X}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\sim} Nor(0,1)$$

Z-test

Tipo	Rifiuto H_0 se
I	$\lvert u \rvert > z_{1-\alpha/2}$
II	$u > z_{1-\alpha}$
III	$u < -z_{1-\alpha}$

Test parametrici di confronto fra 2 popolazioni

Siano $X_1, X_2, \dots X_n$ e $Y_1, Y_2 \dots Y_m$ 2 popolazioni con $n, m \in \mathbb{N}$ e non per forza uguali.

$X \sim Nor(\mu = \mu_X,\ \sigma^2 = \sigma_X^2)$

$Y \sim Nor(\mu = \mu_Y,\ \sigma^2 = \sigma_Y^2)$

Test

Tipo I : $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$

Tipo II : $\begin{cases} H_0 : \mu_X = \mu_Y \text{ o } \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$

Tipo III : $\begin{cases} H_0 : \mu_X = \mu_Y \text{ o } \mu_X \leq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$

La nostra variabile aleatoria sarà la differenza tra le 2 variabili aleatorie

Si osserva
$$\bar{X}_n \sim Nor\left(\mu = \mu_X,\ \sigma^2 = \frac{\sigma_X^2}{n}\right)$$
$$\bar{Y}_m \sim Nor\left(\mu = \mu_Y,\ \sigma^2 = \frac{\sigma_Y^2}{m}\right)$$

Cioè
$$\bar{X}_n - \bar{Y}_m \sim Nor\left(\mu = \mu_X - \mu_Y,\ \sigma^2 = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

Caso 1 ; varianza nota

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim Nor(0,1)$$

Allora
$$U = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \overset{H_0}{\sim} Nor(0,1)$$

Tipo	Rifiuto H_0 se
I	$\lvert u \rvert > z_{1-\alpha/2}$
II	$u > z_{1-\alpha}$
III	$u < -z_{1-\alpha}$

Esempio :

Prima delle vacanze : 5 corse con tempo medio 58,82 sec

Dopo : 6 corse con tempo medio 54,41 sec

tempo X per fare 400m è distribuito normalmente

e che la varianza sia uguale $\sigma^2 = 0,1$

Le vacanze hanno contribuito a peggiorare le prestazioni?

X = "tempo prima delle vacanze"

Y = "tempo dopo le vacanze"

$H_0: \mu_X = \mu_Y$ le vacanze non hanno influito

$H_1: \mu_X < \mu_Y$ le vacanze hanno influito

$$u = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \overset{*}{=} -3,0812 \qquad z_{1-\alpha} = 1,6449 \text{ con } \alpha = 5\%$$

$u < -z_{1-\alpha} \Rightarrow$ Rifiuto H_0

* Dato che devono essere uguali è 0.

Esercizio Cons. in Matlab

Supponiamo H_0 ($\mu_X = \mu_Y$) (vero nel tipo III)

Caso 2: varianze incognite, uguali

$$\sigma_X^2 = \sigma_Y^2 = V$$

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

$$S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \bar{Y}_m)^2$$

$$\hat{V} = S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

$\hat{V}$ prende il nome di varianza combinata

La statistica test è data da

$$U = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \overset{H_0}{\sim} T(\nu = n + m - 2)$$

Tipo	Rifiuto H_0 se
I	$\lvert u \rvert > t_{1-\alpha/2;\, n+m-2}$
II	$u > t_{1-\alpha;\, n+m-2}$
III	$u < -t_{1-\alpha;\, n+m-2}$

Caso 3: Varianze incognite, non uguali

$$U = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \overset{H_0}{\sim} T(\nu = \nu^*)$$

dove $\nu^* = P.I.\left(\frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2}\right)$

Dove P.I. indica parte intera

Confronto di due Normali da campioni accoppiati

C'è un campione di lunghezza n, formato da n coppie $(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)$

X e Y sono normali e la loro differenza $W = X - Y$ sarà anch'essa normalmente distribuita

$W \sim Nor(\mu_W, \sigma_W^2)$ dove $\mu_W = \mu_X - \mu_Y$, σ_W^2 incognita

Tipo I: $\begin{cases} H_0: \mu_W = 0 \\ H_1: \mu_W \neq 0 \end{cases}$

Tipo II: $\begin{cases} H_0: \mu_W = 0 \quad o \quad \mu_W \leq 0 \\ H_1: \mu_W > 0 \end{cases}$

Tipo III: $\begin{cases} H_0: \mu_W = 0 \quad o \quad \mu_W \geq 0 \\ H_1: \mu_W < 0 \end{cases}$

Segue

$$S_W^2 = \frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W}_n)^2$$

$$U = \frac{\bar{W}_n}{\sqrt{\frac{S_W^2}{n}}} \overset{H_0}{\sim} T(\nu = n-1)$$

Tipo	Rifiuto H_0 se
I	$\lvert u \rvert > t_{1-\alpha/2}$
II	$u > t_{1-\alpha}$
III	$u < -t_{1-\alpha}$

; $n-1$

Esercizio su Matlab

i	1	2	3	4	5	6
tempo calcolato prima esperimento	170,3	162,2	210,3	158,3	180,1	192,3
tempo calcolato dopo esperimento	166,5	165,8	171,1	163,2	152,2	161,5

Stabilire se il farmaco ha fatto effetto ai livelli di significatività 1%, 5%, 10%

Data una osservazione

Regressione lineare

Y, X : capire se Y è correlata in maniera lineare da X

Y: Variabile risposta X: variabile esplicativa o predittiva

Vogliamo capire se esiste $Y = \beta_0 + \beta X + E$ ← errore

- Vengono raccolte n osservazioni (x_i, y_i)
- Stimare i coefficienti β_0 e β, i cui stimatori vengono indicati con $\hat{\beta}_0$ e $\hat{\beta}$
- β_0 e β li cerchiamo col metodo dei minimi quadrati

Cerchiamo $y = \hat{\beta}_0 + \beta x$

Data (x_i, y_i) segue $y_i \rightarrow \hat{y}_i$

$\cancel{y_i = \beta_0 + \beta x_i}$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta} x_i$

Si osserva che

$$y_i - \hat{y}_i = y_i - (\beta_0 + \beta x_i)$$

Si scelgono come stimatori quei particolari valori che minimizzano la somma dei quadrati delle differenze

$$f(\beta_0, \beta) = \sum_j \left(y_j - (\beta_0 + \beta x_j) \right)^2$$

Quindi vogliamo quei β_0 e β che minimizzano questa funzione

La coppia di valori $(\hat{\beta}_0, \hat{\beta})$ che minimizza la funzione $f(\beta_0, \beta)$ è data da

$$\hat{\beta} = \frac{\sum_{ij} (x_j - \bar{x})(y_j - \bar{y})}{\sum_j (x_j - \bar{x})^2}$$

dove $\bar{x}$ sono i $\bar{y}$ valori medi

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$$

Dimostrazione

$$f(\beta_0, \beta) = \sum_j \left(y_j - (\beta_0 + \beta x_j) \right)^2$$

① $$\frac{\partial f}{\partial \beta_0} = -2 \sum_j \left(y_j - (\beta_0 + \beta x_j) \right)$$

② $$\frac{\partial f}{\partial \beta} = -2 \sum_j x_j \left(y_j - (\beta_0 + \beta x_j) \right)$$

Poniamo $\frac{\partial f}{\partial \beta_0} = 0$ e $\frac{\partial f}{\partial \beta} = 0$

① $$\sum_j y_j - \sum_j \beta_0 - \beta \sum_j x_j = 0$$

$$\bar{y} = \frac{1}{n} \sum_j y_j \qquad \bar{x} = \frac{1}{n} \sum_j x_j$$

$$n\bar{y} - n\beta_0 - n\beta\bar{x} = 0 \qquad \beta_0 = \bar{y} - \beta\bar{x} = \hat{\beta}_0$$

② $$\sum_j x_j y_j - \beta_0 \sum_j x_j - \beta \sum_j x_j^2 = 0$$

$$\sum_j x_j y_j - (\bar{y} - \beta\bar{x}) \sum_j x_j - \beta \sum_j x_j^2 = 0$$

$$\sum_j x_j y_j - \bar{y} \sum_j x_j + \beta\bar{x} \sum_j x_j - \beta \sum_j x_j^2 = 0$$

$$\sum_j (y_j - \bar{y}) x_j - \beta \sum_j (x_j - \bar{x}) x_j = 0$$

$$\beta = \frac{\sum_j (y_j - \bar{y})\, x_j}{\sum_j (x_j - \bar{x})\, x_j}$$

Osservo che $\sum_j \bar{x}(x_j - \bar{x}) = 0$

$\sum_j \bar{x} x_j - \sum_j \bar{x}^2 = \bar{x} \sum_j x_j - n\bar{x}^2 = \bar{x} n \bar{x} - n\bar{x}^2 = 0$

Vale anche per $\sum_j \bar{x}(y_j - \bar{y})$

$$\beta = \frac{\sum_j (y_j - \bar{y}) x_j - \sum_j (y_j - \bar{y})\bar{x}}{\sum_j (x_j - \bar{x}) x_j - \sum_j (x_j - \bar{x})\bar{x}} =$$

$$= \frac{\sum_j (y_j - \bar{y})(x_j - \bar{x})}{\sum_j (x_j - \bar{x})^2} = \hat{\beta}$$

Abbiamo $\hat{y}_j = \hat{\beta}_0 + \hat{\beta} x_j$

$r_j = y_j - \hat{y}_j = y_j - (\hat{\beta}_0 + \hat{\beta} x_j)$

Introduciamo una varianza

$$\hat{v} = \frac{1}{n-2} \sum_j (y_j - \hat{y}_j)^2 = \frac{1}{n-2} \sum_j r_j^2$$

Abbiamo usato i dati già 2 volte

Notazione

$S_{xx} = \sum_j (x_j - \bar{x})^2 = \sum_j x_j^2 - n\bar{x}$

$S_{yy} = \sum_j (y_j - \hat{y})^2$

$S_{xy} = \sum_j (x_j - \bar{x})(y_j - \bar{y}) = \sum_j x_j y_j - n\bar{x}\bar{y}$

Definizioni

$SS_{Reg} = \sum_j ((\hat{\beta}_0 + \hat{\beta} x_j) - \bar{y})^2$ Varianza spiegata

$SS_{Res} = \sum_j (y_j - (\hat{\beta}_0 + \hat{\beta} x_j))^2$ Varianza residua

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ $\hat{\beta}_0 = \bar{y} - \beta\bar{x}$

$\hat{v} = \frac{SS_{Res}}{n-2}$

$$SS_{Reg} = \left(\frac{S_{xy}^2}{S_{xx}}\right)^{*} \qquad S_{xx}$$

Osservazione

La media campionaria y_j e $\hat{y}_j$ coincidono

$$\sum_j \hat{y}_j = \sum_j (\hat{\beta}_0 + \hat{\beta} x_j) = \sum_j (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_j)$$

$$= \sum_j \bar{y} = n\bar{y} \qquad \bar{y} = \frac{1}{n}\sum_j y_j$$

Varianza y e y non coincidono

$$S_y^2 = \frac{1}{n-1}\sum_j (y_j - \bar{y}) = \frac{1}{n-1} S_{yy}$$

$$S_{\hat{y}}^2 = \frac{1}{n-1}\sum_j ((\hat{\beta}_0 + \hat{\beta} x_j) - \bar{y}) = \frac{1}{n-1} SS_{Reg}$$

Teorema

La varianza totale è la somma della varianza spiegata più la varianza residua

$$S_{yy} = SS_{Reg} + SS_{Res}$$

Dim

$$SS_{Res} = S_{yy} - SS_{Reg} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} =$$

$$= \frac{S_{xx} S_{yy} - (S_{xy})^2}{S_{xx}}$$

Coefficiente di determinazione

$$R^2 = \frac{SS_{Reg}}{SS_{yy}} = \frac{S_{yy} - SS_{Res}}{S_{yy}} = 1 - \frac{SS_{Res}}{SS_{yy}}$$

Deve essere 0 per avere una buona stima (cioè SS_{Res})

Si dimostra che

$$R^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}\right)^2$$

$$r = \hat{\beta} \sqrt{\frac{S_{xx}}{S_{yy}}}$$

Coefficiente di condizione corretto

$$R^2_{corr} = 1 - \frac{SS_{Res}/(n-2)}{S_{yy}/(n-1)}$$

Test relativi ai parametri

Ipotesi di normalità NB: $y = \beta_0 + \beta x + E$

$\Downarrow$

$$E \sim Nor(\mu = 0, \sigma^2 = V)$$

$\Downarrow$

Proprietà

$$\hat{\beta} \sim Nor\left(\mu = \beta, \sigma^2 = \frac{V}{S_{xx}}\right)$$

$$\hat{\beta}_0 \sim Nor\left(\mu = \beta_0, \sigma^2 = V\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

Si ha $\left(\frac{n-2}{V}\right)\hat{V} = \frac{1}{V}\sum_j (y_j - \hat{y}_j) \sim \chi^2 (\nu = n-$

Test su β

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{V}{S_{xx}}}} \sim Nor(\mu = 0, \sigma^2 = 1)$$

V è incognita e consideriamo la stima $\hat{V}$

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{\hat{V}}{S_{xx}}}} \sim T(\nu = n-2)$$

Test d'ipotesi

tipo I: $\begin{cases} H_0: \beta = \beta^* \\ H_1: \beta \neq \beta^* \end{cases}$

tipo II: $\begin{cases} H_0: \beta = \beta^* \text{ o } \beta \leq \beta^* \\ H_1: \beta > \beta^* \end{cases}$

tip III $\begin{cases} H_0: \beta = \beta^* \text{ oh } \beta \geq \beta^* \\ H_1: \beta < \beta^* \end{cases}$

Statistica test

$$U = \frac{\hat{\beta} - \beta^*}{\sqrt{\frac{\hat{V}}{S_{xx}}}} \overset{H_0}{\sim} T(v = n-2)$$

Dove β^* è il valore "vero"

A volte $\hat{\beta} - \beta^*$ si dice $\tilde{\beta}$
si fa perché a volte non si conosce β^*