Modulo di Statistica

Statistica e Modelli Stocastici
CdL: Ingegneria Informatica

Statistica e Topografia
CdL: Ingegneria delle tecnologie per l'edilizia

Prof. Alessandro Fassò
alessandro.fasso@unibg.it

Tutor Dott. Paolo Maranzano
paolo.maranzano@guest.unibg.it

aa 2020/21

Aspetti organizzativi

Lezioni ed Esami

- Informatici: Lunedì 10.30-12.30, Martedì 10.30-13.30
- Edili: Lunedì 13.30-15.30, Mercoledì 10.30-13.30
- Crash course Matlab/Excel 4 ore, aula computer
- **Ricevimento**: Giovedì 14:30-16:30, in modalità telematica con appuntamento preso via email (alessandro.fasso@unibg.it)
- E-learning:
  www.unibg.it link diretto in corso di definizione

- **E' necessario conoscere Analisi I e Geometria !!**

Modalità d'esame:

Modalità di svolgimento dell'esame

1. Durante il semestre:
   a. una prova su Excel/Matlab
   b. n.4 prove intermedie informatizzate (50%)
   c. Lavoro di gruppo:
      i. Esercizi a casa
      ii. L'ultimo è un piccolo case study sulla regressione
   d. Orale finale su case study e teoria (50%)
2. In appello ordinario:
   con un'unica prova informatizzata ed un orale obbligatorio nello stesso appello.
3. Informatici: occorre aver superato il 1° modulo per fare le prove parziali del 2° modulo.

Elearning ILIAS

- Iscriversi al corso
- PROFILO PUBBLICO: **Abilitare visibilità nome**
- **Caricare foto** nel profilo
- Iscriversi alle prove intermedie indicando se si è in corso o fuori corso
- Iscriversi a un gruppo
- Partecipare ai Forum
- Inviare/ricevere Email per/dai docenti.

Preparazione dell'Esame e Bibliografia

1. Frequenza
2. Lavoro di gruppo
3. Installare Excel/Matlab
   https://www.unibg.it/sites/default/files/avvisi/matlab-tah-student-license.pdf
4. Laboratorio introduttivo a Excel/Matlab
5. *Elearning su ilias*.
   Guida alle lezioni e materiale su internet (elearning/ilias)
6. Walpole et al. (2016) **Analisi Statistica dei Dati per l'Ingegneria**, Pearson
7. S. Ross (2003) **Probabilità e Statistica per l'Ingegneria e le Scienze**, Apogeo
8. Giuliani et al. (2015) **Analisi statistica con Excel**, Maggioli.

Introduzione

Decisioni in condizioni di incertezza

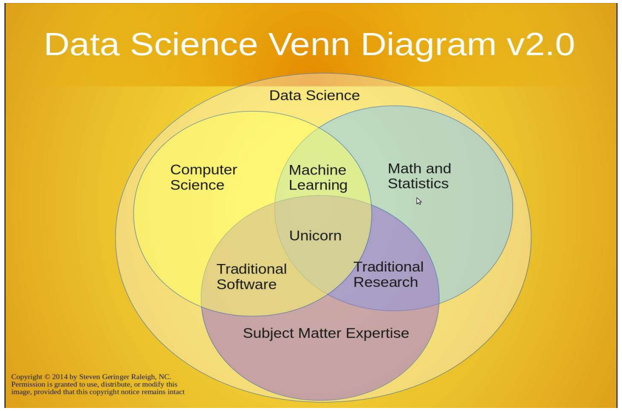
Dati empirici + Teoria specifica + Modello Statistico

$\Downarrow$

Previsioni/Conclusioni operative

- Conoscenza e consapevolezza dell'errore
  ⇒ Identificazione e Quantificazione dell'incertezza
  - sul modello
  - sulla previsione
- Modelli statistici

Data Science & Machine Learning



Articolazione della Materia

- **Statistica Descrittiva**
- **Calcolo delle Probabilità**
- **Inferenza Statistica**

COVID-19: Validità test sierologico

A maggio uno studente di SMS2 mi ha chiesto lumi sulla seguente affermazione riferita all'epidemiologo Vespignani:

"Anche se voi trovate un test che ha sia la **specificità** che la **sensibilità** al 99%, con l'attuale **prevalenza** di casi in Italia, uno è un falso positivo al 50%.

La **validità**, quindi, è come tirare una monetina. Questo per motivi tecnici. Dal punto di vista del singolo il test sierologico è poco importante, è importante invece sui grandi numeri, a livello epidemiologico".

Oggi abbiamo un problema simile anche con i test rapidi.

Validita dei test diagnostici

Esito test = Positivo/Negtivo

Stato individuo = Infettato/Sano

NB: nel caso del test sierolgico per il COVID-19 per "infettato" si intende un individuo che è entrato in contatto col virus e ha sviluppato gli anticorpi (può essere guarito, asintomatico etc.)

1. La prevalenza è la frazione (probabilità) di infetti nella popolazione
$$P(Infettato)$$
2. La specificità è la probabilità che un SANO risulti NEGATIVO al test:
$$Specificità = P(Negativo|Sano)$$
3. La sensibilità è la probabilità che un INFETTATO risulti POSITIVO al test:
$$Sensibilità = P(Positivo|Infettato)$$
4. Validità o valore predittivo del test
$$P(Infettato|Positivo) = ?$$

Calcolo delle Probabilità

argomenti trattati nella parte $1a$ del corso

- Richiami di Insiemistica
- Esperimenti casuali
- Definizione di probabilità
- Proprietà e regole di calcolo
- Probabilità condizionata
- Indipendenza stocastica
- Richiami di calcolo combinatorio
- Teorema di Bayes

Richiami di Insiemistica

- Insiemi $A, B, ..$ contenuti nell'universo Ω
- Elementi $a, b, x, y, \ldots$ appartenenti all'universo Ω

Relazioni fra Insiemi

- Appartenenza $a \in A$ oppure $a \notin A$
- Uguaglianza $A = B$
- Inclusione $A \subset B$ oppure $B \supset A$
- Esempi
  1. $B = \{2, 4, 6, 8, 10\}$ $2 \in B$, $3 \notin B$
  2. $\mathbb{N} = \{\text{numeri naturali}\}$
  3. $B \subset \mathbb{N}$.

Spazio Ω e Insieme vuoto

- Spazio ambiente o *Spazio Campionario*: $A \subset \Omega$
- Insieme vuoto: ϕ
- Insieme di tutti i sottoinsiemi o Insieme delle parti: $\complement = \complement(\Omega)$

$$A \subset \Omega \quad \Rightarrow \quad A \in \complement$$

Operazioni

- Unione

$$A = B \cup C = \{x | x \in B \text{ oppure } x \in C\}$$

- Intersezione

$$A = B \cap C = \{x | x \in B \text{ e } x \in C\}$$

$$A \cap B = A \cdot B = AB$$

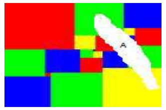
- Differenza

$$A = B - C = \{x | x \in B \text{ e } x \notin C\}$$

- Complemento o Negazione

$$A = nonB = \bar{B} = \Omega - B = \{x | x \notin B\}$$

- Definizione di Insiemi disgiunti: $A \cap B = \phi$
- Partizione di Ω : $B_1, B_2, \ldots, B_k$ *disgiunti* tali che
$$\Omega = B_1 \cup B_2 \cup \ldots \cup B_k = \cup_{j=1}^{k} B_j$$
- Scomposizione di A



$$A = (B_1 \cap A) \cup (B_2 \cap A) \cup \ldots \cup (B_k \cap A)$$

- Regole di De Morgan

$$\boxed{A \cup B = non(\bar{A} \cap \bar{B})} \qquad \boxed{A \cap B = non(\bar{A} \cup \bar{B})}$$

Esperimenti Casuali

- Esperimento deterministico e risultati incerti

Ω = insieme dei possibili risultati sperimentali

- **Esempio 1:** lancio di un dado
$$\Omega = \{f_1, f_2, \ldots, f_6\}$$
  - Eventi elementari $\omega \in \Omega$
  - Eventi $A \subset \Omega$
  - Eventi ed Insiemi

- **Esempio 2:** siamo interessati al **rischio sismico**. In conseguenza di un sisma di una certa intensità ed un certo epicentro si estrae a caso un'unità immobiliare dal catasto e si controlla il suo stato.
  - Siamo quindi interessati all'evento:
  $$B = \text{"la casa estratta a caso dal catasto è danneggiata"}$$
  per cui
  $$\Omega = B \cup \bar{B}.$$
  - Se A indica l'evento "la casa estratta a caso dal catasto è antisismica", abbiamo
  $$B = (B \cap A) \cup (B \cap \bar{A})$$
  cioè, ovviamente: la casa estratta può essere danneggiata essendo di struttura ordinaria o antisismica.

Definizione della Probabilità

Secondo l'**approccio assiomatico** si definisce la probabilità come **la misura** (dell'incertezza) **di un evento.**

Consideriamo Ω finito e $\mathcal{C}$ l'insieme delle parti di Ω. Allora la funzione
$$P : \mathcal{C} \to \mathbb{R}$$
è una probabilità se e solo se per ogni $A, B \subset \Omega$ valgono le segg. 3 proprietà:

1. Nonnegatività: $P(A) \geq 0$
2. Normalizzazione: $P(\Omega) = 1$
3. Additività: Se A, B sono disgiunti allora $P(A \cup B) = P(A) + P(B)$.

Interpretazione della probabilità

1. Approccio soggettivista
   - $P(A)$ misura la fiducia (soggettiva) del verificarsi di A
   - Paradigma della scommessa nei giochi equi

$$P(A) = \frac{posta}{vincita\ netta\ +\ posta}$$

(vincita netta attesa nulla)

2. Approccio classico a priori (campionamento - eventi equiprobabili)
   - $P(A) = \frac{\text{\# casi favorevoli}}{\text{\# casi possibili}}$

3. Approccio modellistico a priori
   - $P(A)$ è conseguenza di una legge nota (fisica ...)
4. Approccio frequentista a posteriori
   - ha un forte significato intuitivo ed "empirico"
   - in assenza di dati non è costruttivo (può ancira essere usato a livello interpretativo) es: sismica o altri eventi rari.

La "scommessa equa"

Si gioca la posta b sul verificarsi di A che ha probabilità $P(A)$ di verificarsi.

Se A si verifica, il banco paga la vincita lorda $v + b$ dove $V = v$ è la vincita netta. Se A non si verifica il banco intasca b e la vincita netta è $V = -b$.

La vincita netta media è allora

$$E(V) = vP(A) + (-b)(1 - P(A))$$

Posto $E(V) = 0$ si ottiene

$$P(A) = \frac{b}{v + b}$$

Naturalmente se $P(A) < \frac{b}{v+b}$ si ha che

$$E(V) < 0$$

che rappresenta il ricavo atteso del banco.

Esempio sulla "scommessa equa"

All'ippodromo scommetto 100/1 sulla vincita di Tornado. Questo significa che

1. la vittoria di Tornado è alquanto inverosimile ma
2. se Tornado vince la posta viene ripagata 100 volte.

Vediamo formalmente il punto 1.

Secondo lo schema della scomessa equa, la propabilità è

$$P(A) = \frac{posta}{vincita\ netta\ +\ posta} = \frac{1}{100}$$

in pratica però la vincita netta media

$$E(V) = 99P(A) + -1(1 - P(A)) = 100p - 1$$

non può essere nulla ma deve essere negativa perché $E(V)$ è proprio il margine del broker.

Perciò ha senso pensare che la probabilità che Tornado vinca sia inferiore a $\frac{1}{100}$

Probabilità e σ –additività

Se Ω è numerabile o continuo allora il 3° assioma si estende come segue.

Consideriamo gli eventi

$$A_1, A_2, \ldots, A_n, \ldots$$

disgiunti a coppie, cioè

$$A_i \cap A_j = \phi \text{ per ogni } i \neq j$$

allora

$$P(A_1 \cup A_2 \ldots) = \sum_{j=1}^{\infty} P(A_j).$$

Proprietà e regole di Calcolo

- Finitezza e normalizzazione, $A \subset \Omega$

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

- Monotonicità, $A \subset B \subset \Omega$,

$$P(A) \leq P(B)$$

- Evento quasi impossibile:

$$P(A) = 0$$

- Evento impossibile:

$$P(\phi) = 0$$

- Evento quasi certo:

$$P(A) = 1$$

- Evento certo:

$$P(\Omega) = 1$$

- Negazione $A \subset \Omega$:

$$P(\bar{A}) = 1 - P(A)$$

- Differenza, $A \subset B \subset \Omega$,

$$P(B - A) = P(B) - P(A).$$

- Unione di eventi qualsiasi, $A, B \subset \Omega$

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

Assegnazione della Probabilità

- $\Omega = \{\omega_1, \ldots, \omega_N\}$

$$0 \underset{(\leq)}{<} P(\omega_i) \underset{(\leq)}{<} 1$$

$$P(\omega_1) + \ldots + P(\omega_N) = 1$$

1. **Esempio** (approccio classico):
   - **a.** esperimento casuale: estrazione di 1 pallina da un'urna contenente 10 palline numerate $1, \ldots, 10$
   - **b.** $\Omega = \{\omega_1, \ldots, \omega_{10}\} = \{1, \ldots, 10\}$
   - **c.** $P(\omega_i) = P(i) = \frac{1}{10}$

2. **Esempio:** lancio ordinato di due monete "distinguibili" m_1 ed m_2
   a. $\Omega = \{\omega_1, \ldots, \omega_4\} = \{t_1t_2, t_1c_2, c_1t_2, c_1c_2\}$
   b. $P(\omega_i) = \frac{1}{4}$
   c. NB: $\sum_{j=1}^{4} P(\omega_j) = 1$

3. **Esempio:** lancio di due monete "indistinguibili" (senza ordine)
   a. $\Omega = \{\omega_1, \ldots, \omega_3\} = \{tt, tc, cc\}$
   b. $P(tt) = P(cc) = \frac{1}{4}$
   c. $P(tc) = \frac{1}{2}$
   d. NB: $\sum_{j=1}^{3} P(\omega_j) = 1$

Probabilità Condizionata

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

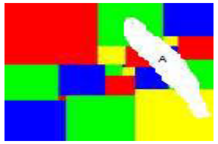
- NB: $P(B) > 0$
- Aggiornamento delle informazioni: $\Omega \to B$

Formula moltiplicativa:

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

Formula delle PROBABILITA' TOTALI

Insegna ad "aggregare" le probablità condizionate ad una serie di eventi che forma una partizione.



Teorema: Se

$B_1, \ldots, B_k$ è una partizione di Ω

(cioè se $B_iB_j = \phi$ e $\cup_{j=1}^{k} B_j = \Omega$) allora

$$P(A) = P(A|B_1)P(B_1) + \ldots + P(A|B_k)P(B_k)$$

Esempio di "Controllo della Qualità"

- D ="pezzo difettoso"
- L_i ="pezzo dalla Linea i"

Linea produttiva	Produzione oraria	Difettosità
1	500	15%
2	1000	1%

$$P(D) = P(D|L_1)P(L_1) + P(D|L_2)P(L_2)$$

$$= 0.15\frac{1}{3} + 0.01\frac{2}{3} \cong 0.057$$

Indipendenza stocastica

Due eventi si dicono indipendenti se

$$P(AB) = P(A)P(B)$$

CONSEGUENZE:

$$P(A|B) = P(A) \quad \text{e} \quad P(B|A) = P(B)$$

Esempio: Lancio di $n = 2$ monete

$$\Omega_1 = \{t_1, c_1\} \quad \text{e} \quad \Omega_2 = \{t_2, c_2\}$$

$$\Omega = \Omega_1 \times \Omega_2 = \{t_1t_2, t_1c_2, c_1t_2, c_1c_2\}$$

$$P(t_2|t_1) = P(t_2)$$

Esempio: $n = 2$ **estrazioni senza reinserimento da**

$$urna = \{3 \; rosse, \; 2 \; verdi\}$$

$$\Omega_1 = \{r_1, v_1\} \quad \text{e} \quad \Omega_2 = \{r_2, v_2\}$$

$$\Omega = \Omega_1 \times \Omega_2 = \{r_1r_2, r_1v_2, v_1r_2, v_1v_2\}$$

$$P(r_1) = \frac{3}{5} \quad \text{e} \quad P(v_1) = \frac{2}{5}$$

$$P(r_2|v_1) = \frac{3}{4} \quad \text{e} \quad P(r_2|r_1) = \frac{2}{4}$$

perciò non c'è indipendenza.

Tuttavia usando le probabilità totali abbiamo:

$$P(r_2) = P(r_2|v_1)P(v_1) + P(r_2|r_1)P(r_1) = \frac{3}{5} = P(r_1) \text{ !!!}$$

Esercizio per casa

Si consideri il caso delle $n = 2$ estrazioni di cui sopra

1. considerando le estrazioni senza rimessa, costruire
   a. la tabella a doppia entrata delle probabilità congiunte
   b. la tabella a doppia entrata delle probabilità condizionate
2. Svolgere l'esercizio 1 nel caso di estrazioni con rimessa

Richiami di Calcolo Combinatorio

Campionamento in blocco

Trattiamo dapprima gli elementi di calcolo combinatorio relativi al "campionamento in blocco" che possiamo definire tramite la seguente analogia formale ed interpretativa:

Estrazioni senza rimessa da un'urna ≅ Campionamento in blocco

Caratteristica è che ogni elemento degli n oggetti dell'urna appare, al più una sola volta nella sequenza dei risultati.

Vediamo ora le formule principali per tale schema di campionamento.

Fattoriale di n

In quanti modi diversi si possono ordinare n oggetti distinguibili ?

Il numero di **Permutazioni** di n oggetti distinguibili è dato dal fattoriale di n, cioè:

$$n! = n(n-1) \times \ldots \times 2 \times 1$$

con $0! = 1$ e $1! = 1$

Proprietà:

$$n! = n(n-1)! \text{ con } 0! = 1$$

Esempi

- Il caso di n piccolo

I simboli A, B, C possono essere ordinati in

$$3! = 3 \times 2 \times 1 = 6$$

modi diversi:

$$ABC, BAC, BCA, CBA, ACB, CAB$$

- Il caso di n medio

Se le $n = 12$ pratiche giacenti nel catasto edilizio urbano cadano a terra e vengono raccolte a caso possono essere riordinate in

$$12! = 12 \cdot 11 \cdot \ldots \cdot 2 \cdot 1 = 479'001'600$$

modi diversi.

- Il caso di n grande

$$100! \cong 9 * 10^{157}$$

formule di Stirling

$$n! \cong \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Matlab:

- **`perms('abc')`**
- **`factorial(n)`**
- **`factorial(n)`**$=171! > \max int$ (=**`inf`**)

Excel:

- $= \textit{fattoriale}(10)$
- $= \textit{fattoriale}(171)$

Disposizioni Semplici (n, r)

Quanti gruppi di r oggetti si possono costruire partendo da n oggetti distinguibili, quando si considerano due gruppi diversi se almeno un elemento è diverso o l'ordine è diverso, essendo gli r oggetti tutti diversi ?

Il numero di Disposizioni Semplici (n, r) di n oggetti a gruppi di r con ordinamento è dato da:

$$D_{n,r} = n(n-1)\ldots(n-r+1) = \prod_{j=n-r+1}^{n} j = \frac{n!}{(n-r)!}$$

Matlab:

$D_{n,r} =$ **`factorial(n)/factorial(n-r)`**

$=$ **`prod([n-r+1:n])`**

Combinazioni Semplici (n,r)

Quanti gruppi di r oggetti si possono costruire partendo da n oggetti distinguibili, quando si considerano due gruppi diversi solo se almeno un elemento è diverso a prescindere dall'ordine ?

Il numero di Combinazioni Semplici (n,r) di n oggetti a gruppi di r senza ordinamento

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

$$= \frac{n(n-1)\ldots(n-r+1)}{r!}$$

NB:

$$C_{n,r} = C_{n,n-r}$$

Esempio (indistinguibili)

Di $n = 10$ tifosi, $r = 3$ sono supporter della squadra A ed hanno la maglia rossa i restanti $n - r = 7$ tengono per la squadra B ed hanno la maglia blu ma sono per il resto indistinguibili. In quanti modi diversi si possono disporre nella fila alle casse dello stadio.?

$$C_{10,3} = \binom{n}{r} = \frac{10!}{7!3!} = 120 = C_{10,7}$$

Domande

- Se $n = 100$ e $r = 30$ in quanti modi si possono disporre ?
- Matlab:
  `nchoosek(n,r)`
- Usando il PC, verificare numericamente le regole $C_{n,r} = C_{n,n-r}$ per $n = 10$, $r = 0,\ldots,10$.

Campionamento con Reinserimento

Trattiamo ora gli elementi di calcolo combinatorio relativi al "campionamento con reinserimento" che possiamo definire tramite l'analogia formale ed interpretativa con le corrispondenti estrazioni da un'urna.

Caratteristica è che ogni elemento degli n oggetti dell'urna può apparire anche più volte.nella sequenza dei risultati.

Vediamo ora le formule principali per tale schema di campionamento.

Disposizioni con Ripetizione (n,r)

Quanti gruppi di r oggetti si possono costruire partendo da n oggetti distinguibili, quando si considerano due gruppi diversi se almeno un elemento è diverso o l'ordine è diverso, essendo gli r oggetti **uguali o diversi** ?

Il numero di Disposizioni con Ripetizione (n,r) di n oggetti a gruppi di r con ordinamento è dato da:

$$D^*_{n,r} = n \times n \times \ldots \times n = n^r$$

Esempio

Prodotto cartesiano. Se il capitolato di una casa prevede $n_1 = 3$ colori per le pareti, $n_2 = 5$ tipi di piastrelle ed $n_3 = 2$ colori di infissi, quanti appartamenti diversi si possono fare ?

$$d^* = 3 \times 5 \times 2 = 30$$

NB: estende il concetto di $D^*_{n,r}$.

Domanda

- Come si calcola la potenza n^r con xls ?

Teorema di Bayes

Problematica delle ”*probabilità inverse”:* o ***"a posteriori"***.

Osservato l'evento A siamo interessati alle ”*probabilità delle cause”* di A

Torniamo all'Esempio di "Controllo della Qualità"

Linea produttiva	Produzione oraria	Difettosità
1	500	15%
2	1000	1%

ci chiediamo se la probabilità che un pezzo provenga dalla linea 1 è diversa quando questo è risultato difettoso o meno

$$P(L_1|D) \gtreqless P(L_1|\bar{D}) \gtreqless P(L_1) \quad ?$$

Teorema di Bayes

- B_j partizione di Ω – Insieme delle possibili cause di A
- $P(B_j)$ Probabilità **a priori** delle cause B_j
- $P(A|B_j)$ Probabilità condizionate o verosimiglianze
- $P(B_j|A)$ Probabilità **a posteriori** della causa B_j noto l'effetto A.

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

Esempio di "Controllo della Qualità" (segue)

Osservato un difettoso D, la probabilità a posteriori che questo venga dalla linea 1 è:

$$P(L_1|D) = \frac{P(D|L_1)P(L_1)}{P(D|L_1)P(L_1) + P(D|L_2)P(L_2)}$$
$$= \frac{0.15\frac{1}{3}}{0.15\frac{1}{3} + 0.01\frac{2}{3}} \cong 0.88$$

che risulta ssai maggiore della corrispondente probabilità marginale o a priori $P(L_1) = \frac{1}{3}$.

- per casa: Calcolare $P(L_2|D)$ con la formula di Bayes e con la regola del complementare.

Esempio: Validita dei test diagnostici

Esito test = Positivo/Negtivo

Stato individuo = Infettato/Sano

NB: nel caso del test sierolgico per il COVID-19 per "infettato" si intende un individuo che è entrato in contatto col virus e ha sviluppato gli anticorpi (può essere guarito, asintomatico etc.)

1. La prevalenza è la frazione (probabilità) di infetti nella popolazione
$$P(Infettato)$$
2. La specificità è la probabilità che un SANO risulti NEGATIVO al test:
$$Specificità = P(Negativo|Sano)$$
3. La sensibilità è la probabilità che un INFETTATO risulti POSITIVO al test:
$$Sensibilità = P(Positivo|Infettato)$$
4. Validità o valore predittivo del test
$$P(Infettato|Positivo) = ?$$

Esempio: Validita dei test diagnostici (segue)

Validità o valore predittivo del test

$$\begin{aligned} P(Infettato|Positivo) &= \frac{P(Positivo|Infettato)P(Infettato)}{P(Positivo)} \\ &= \frac{P(Positivo|Infettato)P(Infettato)}{P(Pos|Sano)P(Sano) + P(Pos|Inf)P(Inf)} \end{aligned}$$

Esempio: Algoritmi anti-spamming

(versione semplificata di popfile: www.paulgraham.com/spam.html, www.paulgraham.com/better.html, getpopfile.org/docs/faq:paulgraham)

- mail = $(parola_1, \ldots, parola_k)$ con k che dipende dalla mail
- Aggiornamento delle informazioni ad ogni nuova mail ricevuta e ri-classificata
- Probabilità di spamming di ciascuna mail $\Rightarrow$ Classificazione automatica via Bayes

B = "la mail è buona"

$\bar{B}$ = "la mail è spamming"

$A = (W_1, \ldots, W_k)$ = "email osservata composta dalle parole $W_1, \ldots, W_k$"

W_j = "$j-ma$ parola presente nella mail A"

Si applica il Teorema di Bayes

Probabilità a posteriori di spamming:

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

Si classifica come spamming se è più probabile: $P(\bar{B}|A) > P(B|A)$

In pratica

1. Apprendimento (stima): si classifica le mail

   classificazione manuale $\Rightarrow$ $\hat{P}(B_i)$ e $\hat{P}(W_j|B_i)$

   Cioè si calcolano le probabilità come frequenze relative per $B_1 = B$ e $B_2 = \bar{B}$ nella casella postale dell'interessato:

$$\hat{P}(B_i) = \frac{\#\left(\text{parole } \in B_i\right)}{\#\left(\text{tutte le parole usate nella casella}\right)}$$

$$\hat{P}(W_j/B_i) = \frac{\#\left(\text{parole } W_j \text{ delle mail classificate } B_i\right)}{\#\left(\text{tutte le parole } \in B_i\right)}$$

Statistica (Ed/Inf) aa20-21 A.Fassò

2. Si usa l'ipotesi semplificativa (molto forte ma funziona in pratica) di indipendenza fra le parole

$$\hat{P}(A|B_i) = \prod_{j=1}^{k} \hat{P}(W_j|B_i)$$

3. Ulteriori dettagli: www.paulgraham.com/spam.html,
4. Estensioni possibili con uso del contesto linguistico e modelli HMM

Variabili Casuali Discrete

(VCD)

Ω finito

Consideriamo dapprima l'esperimento casuale con un numero finito di risultati

$$\Omega = \{\omega_1, \ldots, \omega_N\},\ p(\omega_i)$$

la variabile casuale è una funzione:

$$X = X(\omega_i) : \Omega \to \Re$$

che assume il valore x con probabilità:

$$P(X = x) = \sum_{\iota : X(w_i) = x} p(\omega_i)$$

Statistica-IngInf/Ed 20-21 A.Fassò

Esempio

Consideriamo il lancio di 2 monete

$$X = numero \text{ di teste}$$

Spazio di probabilità:

ω_i	tt	tc	ct	cc
$p(\omega_i)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$X(\omega_i)$	2	1	1	0

Distribuzione di probabilità

x	0	1	2
$p(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Distribuzione di una VCD

Funzione o Distribuzione di probabilità

$$0 < p(x) \leq 1 \text{ per } x = x_1, \ldots, x_k$$

$$p(x_1) + \ldots + p(x_k) = 1$$

NB: vale anche per $k = \infty$.

Funzione di Ripartizione o Distribuzione Cumulata

$$F(x) = P(X \leq x)$$

in pratica

$$F(x_i) = F(x_{i-1}) + p(x_i) \text{ per } i = 1, \ldots, k$$

oppure

$$F(x) = \begin{cases} 0 & x < x_1 \\ F(x_i) \quad \text{per } x_i \leq x < x_{i+1} & i = 1, \ldots, k-1 \\ 1 & x \geq x_k \end{cases}$$

NB: è monotona nondecrescente

Valore Atteso

$$E(X) = x_1 p(x_1) + \ldots + x_k p(x_k) = \sum_{i=1}^{k} x_i p(x_i) = \mu$$

- Linearità 1: $E(a + bX) = a + bE(X)$
- Linearità 2: $E(aX + bY) = aE(X) + bE(Y)$
- Scarti dalla media: $E(X - \mu) = 0$
- Media di trasformate: posto $Y = t(X)$ si ha che

$$E(Y) = E(t(X)) = \sum_{i=1}^{k} t(x_i) p(x_i)$$

Varianza

$$Var(X) = E\left((X - \mu)^2\right) = E(X^2) - \mu^2 = \sigma^2$$

- Proprietà 1: $Var(X) \geq 0$
- Proprietà 2: $Var(a + bX) = b^2 Var(X)$.

Scarto quadratico medio

$$\sigma = std(X) = \sqrt{Var(X)}$$

- Proprietà 2: $std(a + bX) = |b| std(X)$.

Momenti

- Momenti dall'origine, $k \geq 0$

$$\mu_k = E(X^k) = \sum x^k p(x)$$

- Momenti centrati

$$\bar{\mu}_k = E\left((X-\mu)^k\right) = \sum (x-\mu)^k p(x)$$

- Esempi:

$$\mu_0 = 1 \qquad \mu_1 = \mu \qquad \bar{\mu}_2 = \sigma^2$$

- Problema diretto:

$$P(a < X \leq b) = \sum_{a<x\leq b} p(x) = F(b) - F(a)$$

- Problema inverso: Quantili o Percentili:

$$\tilde{x}_p \text{ tale che } F(\tilde{x}_p) = p$$

- Mediana

$$\tilde{x}_{\frac{1}{2}}$$

VCD Uniforme $U(k)$

- Distribuzione

$$p(x) = \frac{1}{k} \qquad \text{per } x = 1, \ldots, k$$

- Ripartizione

$$F(x) = \frac{[x]}{k} \text{ per } 0 < x \leq k$$

- Momenti:

$$E(X) = \frac{k+1}{2}$$

$$Var(X) = \frac{(k+1)(k-1)}{12}$$

Notazione Generale

- ”X ha distribuzione”

$$X \equiv U(k) \qquad X \equiv F \qquad X \equiv p(x) \qquad X \equiv Y$$

- ”X ha distribuzione approssimata”:

$$X \cong U(k) \qquad X \cong F \qquad X \cong p(x) \qquad X \cong Y$$

NOTA: questa approssimazione vale nel senso di un qualche tipo di limite che verrà specificato di volta in volta.

VCD Bernoulliana $B(p)$

Anche detta Binomiale semplice

- Esperimento dicotomico

$$\Omega = \{A, nonA\}$$

- $P(A) = p$
- $X = 1$ se è vero A
- $X = 0$ se è vero $nonA$

$$p(x) = \begin{cases} 1-p & \text{se } x = 0 \\ p & \text{se } x = 1 \end{cases}$$

- Esercizio: calcolare $E(X)$ e $Var(X)$

Indipendenza

Definizione

Due VC X ed Y, definite sullo stesso esperimento casuale (Ω, P) si dicono indipendenti se

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

per ogni x ed y nei rispettivi domini

Teorema

Se X ed Y sono indipendenti con domini D_X e D_Y, medie μ_X e μ_Y e varianze σ_X^2 e σ_Y^2, allora

1. $E(XY) = \mu_X \mu_Y$
2. $E((X-\mu_X)(Y-\mu_Y)) = 0$
3. $V(X+Y) = \sigma_X^2 + \sigma_Y^2$

Dimostrazione:

1. $E(XY) = \sum_{x \in D_X} \sum_{y \in D_Y} xyp(x,y) = \sum_{x \in D_X} xp(x) \sum_{y \in D_Y} yp(y) = \mu_X \mu_Y$

2. $E((X-\mu_X)(Y-\mu_Y)) = \sum_{x \in D_X} \sum_{y \in D_Y} (x-\mu_X)(y-\mu_Y)p(x,y) =$
   $= \sum_{x \in D_X} (x-\mu_X)p(x) \sum_{y \in D_Y} (y-\mu_Y)p(y) = 0$

3. $V(X+Y) = E\left((X+Y-\mu_X-\mu_Y)^2\right) = E\left(((X-\mu_X)+(Y-\mu_Y))^2\right)$
   $= E\left((X-\mu_X)^2\right) + E\left((Y-\mu_Y)^2\right) + 2E((X-\mu_X)(Y-\mu_Y))$
   $= \sigma_X^2 + \sigma_Y^2$

NB:

In 2 abbiamo usato la proprietà fondamentale della media:

$$E(X-\mu_X) = 0$$

VCD Binomiale $Bin(n,p)$

Contatore dei successi in n **Prove Bernoulliane,** cioè in "n **esperimenti casuali dicotomici, indipendenti ed omogenei"**:

- Esperimenti dicotomici

$$\Omega_i = \{A_i, nonA_i\},\ i = 1,\ldots,n$$

- Esperimenti omogenei

$$P(A_i) = p$$

$$\Omega = \Omega_1 \times \ldots \times \Omega_n = \{\omega_1,\ldots,\omega_{2^n}\}$$

- Esperimenti indipendenti $\rightarrow P(A_1 \cap \ldots \cap A_k) = p^k$
- Contatore

$$\begin{aligned} X(\omega_j) &= \text{numero di } A_i \text{ in } \omega_j \\ &= \text{numero di } \textit{successi} \end{aligned}$$

- Distribuzione binomiale

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0,1,\ldots,n$$

- Momenti (media e varianza):

$$E(X) = \sum_{x=0}^{n} x p(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = np$$

$$Var(X) = np(1-p)$$

- Esempi:
  1. estrazioni da un urna con rimessa
  2. Esperimenti Ceteris Paribus
- $Bin(1,p) = B(p)$

Additività della binomiale

Consideriamo n Bernoulliane omogenee e indipendenti:

$$X_1 \equiv B(p),\ X_2 \equiv B(p), \ldots, X_n \equiv B(p)$$

e la loro somma

$$X = \sum_{i=1}^{n} X_i \equiv Bin(n,p)$$

allora

$$E(X) = \sum_{i=1}^{n} E(X_i) = np$$

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p)$$

Additività:

$$X \equiv Bin(n,p) \quad indip \quad Y \equiv Bin(m,p)$$

$$\Downarrow$$

$$Z = X + Y \equiv Bin(n+m,p)$$

VCD Ipergeometrica $IG(n,N,S)$

Consideriamo il contatore dei successi in "n **estrazioni senza reinserimento"** da un **Urna**:

$$U = \{a_1, \ldots, a_S, \bar{a}_1, \ldots, \bar{a}_{N-S}\}$$

$$A_i = \{a_1\} \cup \ldots \cup \{a_S\} \qquad i = 1, \ldots, n$$

$$P(A_1) = \frac{S}{N} = p$$

$X =$ numero di A nelle n estrazioni

- Distribuzione per $n < S$ ed $n < N - S$

$$p(x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}} \qquad x = 0, 1, \ldots, n$$

- Momenti:

$$E(X) = \sum_{x=0}^{n} xp(x) = \sum_{x=0}^{n} x \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}} = np$$

$$Var(X) = np(1-p)\left(1 - \frac{n-1}{N-1}\right)$$

- Note
  - Campionamento da popolazioni finite
  - $Var(X) < np(1-p)$
  - $n \ll N \Rightarrow Bin \cong IG$

Esercizio:

Probabilità di ambo nel lotto: gioco due numeri A e B $\in [1,\ldots,90]$. Ne vengono estratti senza rimessa 5. L'urna è composta di 90 elementi che ripartiamo in 2 favorevoli (A e B) e 98 contrari (gli altri). $X = 0,1,2$ è il numero di favorevoli in $n = 5$ estrazioni.

$$N = 90$$
$$S = 2$$
$$X \equiv Ip(N,S,n)$$

perciò (cfr esercizio Lotto)

$$P(ambo) = P(X = 2) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}} = \frac{\binom{2}{2}\binom{88}{3}}{\binom{90}{5}} = \frac{\binom{88}{3}}{\binom{90}{5}}$$

VCD di Poisson $\wp(\lambda)$

- Distribuzione degli "eventi rari" (cfr anche oltre processo di Poisson)

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!},\ x = 0,1,\ldots$$

- Momenti:

$$E(X) = \lambda$$
$$Var(X) = \lambda$$

Approssimazione della Binomiale

se n grande e p è piccolo

$$Bin(n,p) \cong \wp(\lambda = np)$$

VCD Geometrica $G(\alpha)$

Esempio: Tempi di guasto

- Affidabilità di un sistema semplice.
- Guasti bernoulliani con probabilità p ad ogni istante di tempo $t = 1,2,\ldots$
- Probabilità di avere un guasto al tempo t

$$P(T = t) = (1-p)^{t-1}p, \qquad t = 1,2,\ldots$$

VCD Geometrica

VC del numero di prove per avere un successo

$$p(x) = (1-p)^x p, \qquad x = 0, 1, 2, \dots$$

Momenti

$$E(X) = \frac{1-p}{p} \qquad e \qquad Var(X) = \frac{1-p}{p^2}$$

Variabili Casuali Continue (VCC)

- Premessa: Ω continuo

$$P(\omega) = 0 \qquad \omega \in \Omega$$

- Funzione misurabile

$$X = X(\omega) : \Omega \to \Re \qquad \text{e} \qquad P(X = x) = 0$$

- Funzione di ripartizione

$$F(x) = P(X \le x)$$

- è continua e derivabile q.o.

Densità di probabilità

L'andamento della probabilità sui numeri reali è dato dalla densità

$$f(x) = \frac{d}{dx}F(x)$$

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

$$\int_{-\infty}^{+\infty} f(t)dt = 1$$

Funzione quantile

$$x = Q(p) = F^{-1}(p) \quad se \quad f(x) > 0$$

- Problema diretto: Aree = Probabilità:

$$P(a < X < b) = \int_a^b f(x)dx$$

$$P(a < X < b) = F(b) - F(a)$$

- Problema inverso: Quantili o Percentili:

$$\tilde{x}_p \text{ tale che } F(\tilde{x}_p) = p$$

$$\tilde{x}_p = F^{-1}(p) \text{ se } f > 0$$

- Valore atteso

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \mu$$

- Varianza

$$Var(X) = E\left((X-\mu)^2\right) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$$

MOMENTI

- Momenti dall'origine, $k \geq 0$ (se esistono finiti)

$$\mu_k = E(X^k) = \int_{-\infty}^{+\infty} x^k f(x)dx$$

- Momenti centrati

$$\bar{\mu}_k = E\left((X-\mu)^k\right) = \int_{-\infty}^{+\infty} (x-\mu)^k f(x)dx$$

- Esempi: $\mu_0 = 1 \qquad \mu_1 = \mu \qquad \bar{\mu}_2 = \sigma^2$
- Proprietà $\mu_k < \infty \;\Rightarrow\; \mu_h < \infty$ per ogni $h \leq k$

VCC Rettangolare $R(0,1)$

- Densità costante

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

- Ripartizione

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 < x < 1 \\ 1 & x > 1 \end{cases}$$

- Momenti:

$$E(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^1 xdx = \frac{1}{2}$$

$$Var(X) = E(X^2) - \mu^2 = \int_0^1 x^2dx - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

- ESERCIZIO: Studiare la vcc R(a,b).

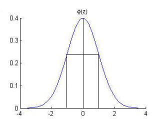
VCC Normale $N(\mu, \sigma^2)$

- Densità di $X \equiv N(\mu, \sigma^2)$

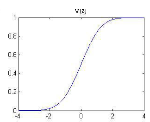
$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$$

- Ripartizione di X

$$F_{\mu,\sigma^2}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

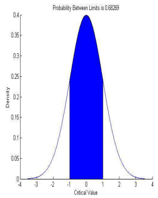
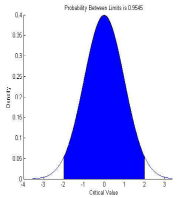
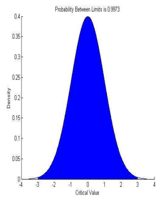


$\phi(z)$



$\Phi(z)$

- Unità di misura della gaussiana $N(\mu, \sigma^2)$ è "σ":

$P(|X-\mu| < \sigma) \cong 0.68$ $\quad$ $P(|X-\mu| < 2\sigma) \cong 0.95$ $\quad$ $P(|X-\mu| < 3\sigma) \cong 0.997$

Inoltre

$$P(|X-\mu| > 4\sigma) < 7 \cdot 10^{-5}$$
$$P(|X-\mu| > 5\sigma) < 6 \cdot 10^{-7}$$

Momenti

$$E(X) = \mu \quad \text{e} \quad Var(X) = \sigma^2$$
$$\bar{\mu}_3 = E\left((X-\mu)^3\right) = 0$$
$$\bar{\mu}_4 = E\left((X-\mu)^4\right) = 3\sigma^4$$

Indici di forma

- **Simmetria**
  La normale è simmetrica:

$$P(X-\mu < a) = P(Z-\mu > -a)$$
$$\Phi\left(\frac{x-\mu}{\sigma}\right) = 1 - \Phi\left(-\frac{x-\mu}{\sigma}\right)$$
$$Sk = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right) = \frac{E\left((X-\mu)^3\right)}{\sigma^3} = 0.$$

- **CURTOSI**

$$k = E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) = \frac{E\left((X-\mu)^4\right)}{\sigma^4} = 3.$$

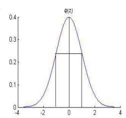
VCC Normale Standard $Z \equiv N(0,1)$

- Densità di Z

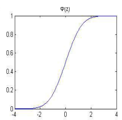
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

- Ripartizione di Z

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$



$\phi(z)$



$\Phi(z)$

- **Momenti** di Z

$$E(Z) = 0$$
$$Var(Z) = E(Z^2) = 1$$
$$Sk(Z) = EZ^3 = 0$$
$$k(Z) = EZ^4 = 3$$

- Problema diretto per Z: Aree = **Probabilità**:

$$P(a < X < b) = \int_a^b \phi(x)dx = \Phi(b) - \Phi(a)$$

- Problema inverso per Z: **Quantili** (Percentili):

$$z_\alpha = \Phi^{-1}(1-\alpha) = \tilde{z}_{1-\alpha}$$

- Standardizzazione di X

$$X \equiv N(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X-\mu}{\sigma} \equiv N(0,1)$$

- Riscalatura di Z

$$Z \equiv N(0,1) \quad \Rightarrow \quad X = \mu + \sigma Z \equiv N(\mu, \sigma^2)$$

- Problema diretto per X: Aree = Probabilità:

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
$$= \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \phi(x)dx = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

- Problema inverso per X: Quantili (Percentili):

$$x_\alpha = \mu + \sigma\Phi^{-1}(1-\alpha)$$
$$= \mu + \sigma z_\alpha = \tilde{x}_{1-\alpha}$$

VCC Esponenziale Negativa $Exp^{-}(\lambda)$

Distribuzione delle durate

$$f_\lambda(x) = \lambda e^{-\lambda x}, \qquad x > 0, \lambda > 0$$

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

Momenti

$$E_\lambda(X) = \frac{1}{\lambda} \qquad Var_\lambda(X) = \frac{1}{\lambda^2}$$

Processo di Poisson

Esempi: Chiamate ad un server/Eventi sismici nel tempo

Si considerano gli intertempi fra successive chiamate/eventi indipendenti

$$X_i \equiv Exp^{-}(\lambda)$$

ed il numero di chiamate/eventi nell'intervallo $(a,b]$

$$N_{(a,b]} = \#\left\{ j \,|\, a < \sum_{i=1}^{j} X_i \leq b \right\}$$

è una VC di Poisson:

$$N_{(a,b]} \equiv \wp(\lambda(b-a)).$$

Posto $a = 0$ e $b = t$, N_t è detto "processo di Poisson".di intensità λ.

Così si ha che

$$E(N_t) = \lambda t$$

Perciò

$$\frac{E(N_t)}{t} = \lambda = \text{intensità}$$

= numero medio di chiamate/sismi nell'unità di tempo

Inoltre tempi ed intensità medi sono reciproci

$$E(X) = \frac{1}{E(N)}.$$

Si richiama la funzione Gamma completa

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

Proprietà

$$\Gamma(1) = 1$$

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1),\ \alpha > 1$$

$$\Gamma(n) = (n-1)!,\ n \text{ intero}$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

VCC Gamma $\Gamma(r,\lambda)$

Generalizza la $Exp^{-}(\lambda)$:

$$f(x;r,\lambda) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1} e^{-\lambda x}, \qquad x > 0, \lambda > 0, r = 1,2,..$$

$$P(X > 0) = 1$$

Momenti

$$E(X) = \frac{r}{\lambda} \qquad Var(X) = \frac{r}{\lambda^2}$$

La famiglia Gamma

- $\Gamma(1,\lambda) \equiv Exp^{-}(\lambda)$
- X_1 ed X_2 $\Gamma(r_i,\lambda)$ indipendenti $\Rightarrow X_1 + X_2 \equiv \Gamma(r_1 + r_2, \lambda)$

VC Doppie

Opzionale

- Distribuzione congiunta (superficie)

$$f(x,y)$$

- Ripartizione (volume)

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(t,u)dtdu$$

- Distribuzione Marginale della X (vcd)

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

$$E(X) = \mu_X, \qquad Var(X) = \sigma_X^2$$

- Distribuzione Marginale della Y (vcd)

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

$$E(Y) = \mu_Y, \qquad Var(Y) = \sigma_Y^2$$

- Momenti misti:

$$E(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy$$

Covarianza

$$Cov(X,Y) = E((X-\mu_X)(Y-\mu_Y)) = E(XY) - \mu_X\mu_Y$$

Ceofficiente di **correlazione**

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

- Indipendenza: X ed Y sono indipendenti sse

$$f(x,y) = f_X(x)f_Y(y)$$

  - Indipendenza $\Rightarrow E(XY) = E(X)E(Y)$
  - Indipendenza $\Rightarrow$ Incorrelazione: $Cov = \rho = 0$.

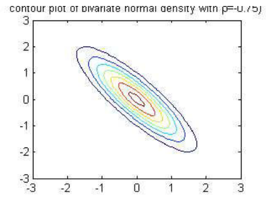
Esempio: la Normale doppia N_2

Opzionale

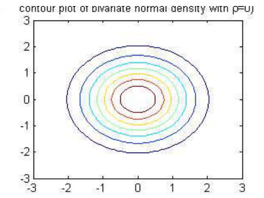
Consideriamo il caso di marginali standardizzate

$$Z_i \equiv N(0,1)$$

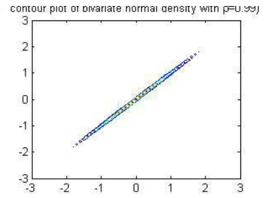
$$f(u,v;\rho)\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}(u^2+v^2-2\rho uv)\right)$$



$N_2(\rho = -0.75)$



$N_2(\rho = 0)$



$N_2(\rho = +0.99)$

Curve di livello di $f(u,v;\rho)$

VC Multiple

$$X = (X_1,\ldots,X_k)$$

- Distribuzione congiunta

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

- Indipendenza

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

- *VC iid*: se $X_1,\ldots,X_k$ sono indipendenti e identicamente distribuite

$$f_{X_1}(\quad) \equiv\ldots\equiv f_{X_k}(\quad) \equiv f(\quad)$$

Somma di VC

Dato $X_1,\ldots,X_n$, con

$$E(X_i) = \mu_i, \qquad Var(X_i) = \sigma_i^2$$

- Se $E(X_i) = \mu_i$

$$E(X_1+\ldots+X_n) = \mu_1+\ldots+\mu_n$$

- Se $Var(X_i) = \sigma_i^2$ e $X_1,\ldots,X_n$ sono indipendenti

$$Var(X_1+\ldots+X_n) = \sigma_1^2+\ldots+\sigma_n^2$$

- NB: se manca indipedendenza non vale l'additività della *Var*:

$$Var(X+Y) = \sigma_X^2+\sigma_Y^2+2Cov(X,Y).$$

Additività della Normale

Se

$$X_1, \ldots, X_n \; indip \; N(\mu_i, \sigma_i^2)$$

$$\sum X_i \equiv N\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$$

Se

$$X_1, \ldots, X_n \; iid \; N(\mu, \sigma^2)$$

$$\sum X_i \equiv N(n\mu, n\sigma^2)$$

In assenza di normalità:

- Qual è la distribuzione di $X_1 + \ldots + X_n$?

Teorema Limite Centrale

Ovvero:

"E' la somma che fa il totale (Totò)"

Teorema Limite Centrale

(per la somma)

Se

$X_1, \ldots, X_n$ sono $iid\,F$ con (μ, σ^2) finiti

allora, per $n \to \infty$

$$T_n = X_1 + \ldots + X_n \cong N(n\mu, n\sigma^2)$$

qualsiasi sia F.

VCC Campionarie

Abbiamo visto VCD Campionarie come la Bin e la IG.

Campione casuale da F

$$X_1, \ldots, X_n \quad iid \quad F$$

$$E(X_i) = \mu \qquad Var(X_i) = \sigma^2$$

Media Campionaria

$$\bar{X} = \frac{1}{n} \sum_j X_j$$

- Teroema delle $3M$

$$E\bar{X} = \mu$$

- Varianza di $\bar{X}$

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

Distribuzione di $\bar{X}$

Se

$$X_1, \ldots, X_n \, iid \, N(\mu, \sigma^2)$$

allora, per ogni $n \geq 1$,

$$\bar{X}_n \equiv N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \bar{X}_n - \mu \equiv N\left(0, \frac{\sigma^2}{n}\right) \qquad \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \equiv N(0,1).$$

In assenza di normalità:

- Qual è la distribuzione di $\bar{X}$?

Teorema Limite Centrale

(per la media)

Se

$$X_1, \ldots, X_n \text{ sono } iid \, (\mu, \sigma^2)$$

allora, per $n \to \infty$,

$$\bar{X}_n \cong N\left(\mu, \frac{\sigma^2}{n}\right)$$

Esempi:

- $Bin(n,p) \cong N(np, np(1-p))$.
- $t_n \cong N(0,1)$
- Velocità di convergenza: in pratica $n > 30$
- Quale F ?
- $U_1, \ldots, U_n \, iid \, R(0,1)$

$$X = \left(\sum_{i=1}^{12} U_i - 6 \right)$$

Natura approssimazione:

$$|f_X(t) - \phi(t)| < 0.0050 \quad \forall t$$

$$|F_X(t) - \Phi(t)| < 0.0023 \quad \forall t$$

Legge dei Grandi Numeri

Se $X_1, \ldots, X_n \; iid \; \mu$

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

Inoltre se $\sigma^2 < \infty$

$$E\left((\bar{X}_n - \mu)^2\right) = \frac{\sigma^2}{n} \to 0.$$

Esempio:

Prove bernoulliane $B(p)$

$$\bar{X} = Frazione\,di\,favorevoli \to p$$

Varianza Campionaria

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})^2$$

- Media di S^2

$$E(S^2) = \sigma^2$$

Traccia DIM:

$$E\left(\sum_{j=1}^{n} (X_j - \bar{X})^2 \right) = E\left(\sum_{j=1}^{n} (X_j \mp \mu - \bar{X})^2 \right)$$

$$= E\left(\sum_{j=1}^{n} (X_j - \mu)^2 + \sum_{j=1}^{n} (\bar{X} - \mu)^2 - 2 \sum_{j=1}^{n} (X_j - \mu)(\bar{X} - \mu) \right)$$

poiché nel doppio prodotto si ha

$$\sum_{j=1}^{n} (X_j - \bar{X})(\bar{X} - \mu) = (\bar{X} - \mu) \sum_{j=1}^{n} (X_j - \mu) = n(\bar{X} - \mu)^2$$

si conclude che

$$E\left(\sum_{j=1}^{n} (X_j - \bar{X})^2 \right) = (n-1)\sigma^2$$

- Qual è la distribuzione di S^2 ?

VCC legate alla Normale
Distribuzione Chi Quadrato: χ^2_n

Motivazione:

- $Z_1, \ldots, Z_n \; iid \, N(0,1) \quad \Rightarrow \quad \sum Z_j^2 \equiv \chi^2_n$
- $S_n^2 \equiv \frac{\sigma^2}{n-1} \chi^2_{n-1}$

Proprietà del χ^2

- Asimmetria: $P(\chi^2_n > 0) = 1$ $\qquad f(x)$ è asimmetrica
- il parametro n è detto **gradi di libertà**
- Gamma: $\chi^2_n \equiv \Gamma(\frac{n}{2}, \frac{1}{2})$
- Momenti: $E(\chi^2_n) = n \qquad Var(\chi^2_n) = 2n$
- Normalità asintotica: per $n \to \infty, \qquad \chi^2_n \cong N(n, 2n)$

Distribuzione t di Student: t_n

Motivazioni

$$t = \frac{N(0,1)}{\sqrt{\chi^2_n / n}} \equiv t_n$$

se il numeratore ed il denominatore sono indipendenti.

Se $X_1, \ldots, X_n \; iid \, N(\mu, \sigma^2)$ allora

$$t = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \equiv t_{n-1}$$

Proprietà

$$f_n(t) = k_n \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

- $f_n(t)$ è simmetrica e campanulare.
- Momenti: $\mu_{k,n} = E(t_n^k)$ esiste finito solo per $k < n$.
- $E(t_n) = 0$ per $n > 1$.
- $V(t_n) = \frac{n}{n-2}$ per $n > 2$. $\qquad$ Inoltre $V(t_n) \to 1$ per $n \to \infty$.
- $k_n = E\left(\frac{t_n^4}{\sigma(t_n)^4}\right) = 3\frac{n-2}{n-4}$ per $n > 4$.
- $k_n > 3$. $\qquad$ Inoltre $k_n \to 3$ per $n \to \infty$.
- $t_n \to N(0,1)$ per $n \to \infty$.

F di Snedecor - $F_{n,m}$ (opzionale)

Motivazione

$$F = \frac{\chi_n^2/n}{\chi_m^2/m} \equiv F_{n,m}$$

se il numeratore ed il denominatore sono indipendenti.

Se $X_1,\ldots,X_n \, iid N(\mu_1,\sigma^2)$ sono indipendente da $Y_1,\ldots,Y_m \, iid N(\mu_2,\sigma^2)$ allora

$$F = \frac{S_X^2}{S_Y^2} \equiv F_{n-1,m-1}$$

Proprietà

- $P(F > 0) = 1$
- $f(x)$ è asimmetrica
- n sono i **gradi di libertà del numeratore** m quelli **del denominatore**
- Se $m \to \infty$, $F_{n,m} \cong \chi_n^2$
- Tavole: $F_{n,m} \equiv \frac{1}{F_{m,n}}$ perciò

$$F_{\alpha,n,m} = \frac{1}{F_{1-\alpha,m,n}}$$

Uso delle VCC campionarie

In generale, per queste VC, una volta nota la distribuzione

$$P(X \leq x) = F(x)$$

interessano i percentili

$$\tilde{x}_p : F(\tilde{x}_p) = p$$

$$\tilde{x}_p = F^{-1}(p)$$

che si calcolano non per via analitica ma usando le tavole, excel, R o **Matlab** (o simili).

Esercizio Chi quadrato

Dato il seguente campione casuale Gaussiano

$$X_1,\ldots,X_n \; IID \; N(\mu,\sigma^2).$$

Si considerino le variabili standardizzate:

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

e la somma dei quadrati delle variabili standardizzate:

$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2.$$

Calcolare la media e la varianza di χ^2 :

$$E(\chi_n^2) = ?$$

$$V(\chi_n^2) = ?$$

Statistica

Prof. Alessandro Fassò
ingegneria.unibg.it/fasso

CdL: Ing.Informatica e Ing.Edile aa 2018/2019

2^a parte
Inferenza Statistica

Inferenza e Campionamento

- Popolazione: finita/infinita, reale/virtuale
- Campione: sottoinsieme della popolazione
- Inferenza: Campione → Popolazione
  - Stima puntuale
  - Intervalli di confidenza
  - Verifica di ipotesi

Stima

Popolazione: X grandezza di interesse con distribuzione $f_\theta(x)$, θ parametro ignoto.

$$\theta = \theta(f)$$

- Campione casuale semplice da X (oppure da f, oppure da F):

$$X_1, \ldots, X_n \quad \text{iid} \quad f_\theta(x)$$

- *Stima* di θ :

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

⇒ $\hat{\theta}$ è una particolare V.C. detta *statistica*
⇒ Incertezza sull'errore di stima

$$\hat{\theta} - \theta$$

Principio del campionamento ripetuto

Si valutano le proprietà di $\hat{\theta}$ nell'ipotesi di ripetere il processo di campionamento un gran numero di volte.

Sono rilevanti in quest'ottica l'interpretazione frequentista della probabilità, la legge dei grandi numeri ed il metodo Monte Carlo.

Problemi di stima

1. Indagini demoscopiche
   - percentuale di "favorevoli"
2. Misura di una grandezza fisica
   - valutazione dell'errore e correzione (calibration)
3. Qualità di un processo produttivo,
   - controllo in accettazione
4. Stima di un segnale (a gradino)
   - dominio delle frequenze e stima parametrica di un segnale
5. Probabililtà di "aspettare troppo" in una coda.

Stima della Media

Dato $X_1,\ldots,X_n$ iid F con $E(X) = \mu$ e $Var(X) = \sigma^2$, la media campionaria

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

è una *stima* di μ.

- **Teorema delle** $3M$: $E(\bar{X}) = \mu$
- Varianza della media campionaria $Var(\bar{X}) = \frac{\sigma^2}{n}$
- Distribuzione di $\bar{X}$

$$\bar{X} \cong N\left(\mu, \frac{\sigma^2}{n}\right)$$

Se $F(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$ questa distribuzione vale per ogni n

In generale vale *per* $n \to \infty$ (Teorema limite centrale).

Stima della Varianza

Dato $X_1,\ldots,X_n$ iid F con $E(X) = \mu$ e $Var(X) = \sigma^2$,

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

è una *stima* di σ^2.

- **Distribuzione Chi-Quadrato** con $n-1$ gradi di libertà: se X_i *iid* $N(\mu,\sigma^2)$ allora

$$S^2\frac{n-1}{\sigma^2} \text{ è } \chi^2_{n-1}$$

Usando le proprietà del χ^2 si ottiene facilmente che:

- $E(S^2) = \sigma^2$
- $Var(S^2) \cong \frac{2\sigma^4}{n-1}$

Stima di una percentuale

Caso 1

Schema di campionamento: "n **estrazioni con reinserimento"** da un **Urna binaria** con composizione

$$\pi = \frac{\#(A)}{N}.$$

All'iesima estrazione si pone

$$X_i = \begin{cases} 1 & \text{se evento } A \\ 0 & \text{se evento } \bar{A} \end{cases}$$

da cui

$$X_1,\ldots,X_n \text{ iid } Bin(1,\pi)$$

allora, il numero di eventi "a" nel campione,

$$S = X_1 + \ldots + X_n \qquad \text{è} \qquad Bin(n, \pi)$$

inoltre, la percentuale campionaria,

$$\hat{\pi} = \bar{X} = \frac{S}{n}$$

è stima di π :

$$E(\hat{\pi}) = \pi \qquad \text{e} \qquad Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n}.$$

Caso 2

Schema di campionamento: "n **estrazioni senza reinserimento.**

Allora

$$S = X_1 + \ldots + X_n \qquad \text{è} \qquad IG(n, N, N\pi)$$

e

$$\hat{\pi} = \frac{S}{n}$$

è stima di π :

$$E(\hat{\pi}) = \pi$$

e

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Stima nonparametrica di F

Avendo a disposizione un campione $X_1, \ldots, X_n \quad iid \quad F,$ ci interessa stimare

$$\theta = P(X \leq t) = F(t)$$

supponendo, per ora, t prefissato.

A tal fine consideriamo la funzione di ripartizione empirica in t, detta anche frequenza cumulata

$$\hat{F}_n(t) = \frac{\#(X_i \leq t, \quad i = 1, \ldots, n)}{n} = \frac{\sum_{i=1}^{n} I(X_i \leq t)}{n}.$$

Si nota che

$$E(I(X_i \leq t)) = P(X \leq t) = F(t)$$

e

$$I(X_i \leq t) \quad iid \quad Bin(1, F(t))$$

da quanto visto per la stima di π, si ha che

$$n\hat{F}_n(t) \sim Bin(n, F(t))$$

e (con probabilità uno):

$$\hat{F}_n(t) \to F(t) \quad per \quad n \to \infty$$

NB: In realtà la stima fatta per un prefissato t può essere estesa a tutto il funzionale, infatti

la convergenza di $\hat{F}_n$ è **uniforme** in t:

$$Var\left(\hat{F}(t)\right) = \frac{F(t)(1-F(t))}{n} \leq \frac{0.25}{n}$$

Perciò, usando la f.r. empirica, possiamo stimare il parametro *funzionale*

$$\theta(t) = F(t) \quad \forall t.$$

Teoria Generale della stima

Consideriamo un campione

$$X_1, \ldots, X_n \text{ iid } f_\theta(x)$$

ed uno stimatore

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$$

Correttezza o non distorsione:

$$E_\theta\left(\hat{\theta}_n\right) = \theta$$

Bias o distorsione

$$b\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$$

Correttezza asintotica

$$\lim_{n\to\infty} E_\theta\left(\hat{\theta}_n\right) = \theta$$

Esercizio:

Dimostrare che

$$\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$$

è astinoticamente non-distorto.

Errore quadratico medio

Sia o meno presente l'errore **sistematico** di uno stimatore dato dal bias, l'incertezza, in termini di campionamento ripetuto è data dalla probabilità di avere "errori di stima" o, in sintesi quadratica:

$$\begin{aligned} MSE\left(\hat{\theta}\right) &= E\left(\hat{\theta} - \theta\right)^2 \\ &= Var\left(\hat{\theta}\right) + b\left(\hat{\theta}\right)^2 \end{aligned}$$

Efficienza

- Efficienza: dati due stimatori $\hat{\theta}_A$ e $\hat{\theta}_B$ il confronto fra i due stimatori si basa su

$$e(A,B) = \frac{MSE\left(\hat{\theta}_B\right)}{MSE\left(\hat{\theta}_A\right)}$$

se

$$e(A,B) \quad \grave{e} \quad \begin{cases} > 1 & \hat{\theta}_A \text{ è più efficiente} \\ = 1 & \hat{\theta}_A \text{ e } \hat{\theta}_A \text{ sono equivalenti} \\ < 1 & \hat{\theta}_A \text{ è meno efficiente} \end{cases}$$

Consistenza

Si dice che $\hat{\theta}_n$ è una stima consistente se "l'incertezza su θ scompare per $n \to \infty$", cioè se

$$\hat{\theta}_n \to \theta \qquad per\ n \to \infty$$

Questo limite è da intendersi "in probabilità" cioè occorre che, $\forall \varepsilon > 0$, valga il limite

$$P\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \to 0$$

Condizione sufficiente per la consistenza

$$E_\theta\left(\hat{\theta}_n\right) \to \theta \qquad per\ n \to \infty$$

$$Var_\theta\left(\hat{\theta}_n\right) \to 0 \qquad per\ n \to \infty$$

Corollario:

Se $MSE\left(\hat{\theta}_n\right) \to 0$ allora $\hat{\theta}_n$ è consistente.

Problemi:

Vedi $\left(MRH - \text{inglese}\right)$ p.142 e 143.

esercizi_stima_MRH_p142.pdf

Intervalli di Confidenza

IC per la media noto σ^2

Consideriamo la stima di μ noto σ in campioni normali. Dall'identitià

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

costruiamo l'intervallo di confidenza (IC) con livello di confidenza $1 - \alpha$ dato da

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Alternativamente l'intervallo di confidenza per la media si può scrivere:

$$IC_\alpha(\bar{X}) = \left\{\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right\}$$
$$= \left[\bar{X} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Interpretazione

- Abbiamo una confidenza del $100(1-\alpha)\%$ che $IC(\bar{X}) \ni \mu$
- Esempi $\alpha = 0.05$ o 0.10 in piccoli campioni, $\alpha = 0.01$ o $\alpha = 0.001$ in medi e grandi campioni
- Interpretazione in termini di campionamento ripetuto
- Esercizio MRH, 4-27.c) e d), (esercizi_IC_MRH_p174.pdf)

Limiti inferiori e Limiti superiori

$$LIC(\bar{X}) = \left\{\mu : \bar{X} - z_\alpha\frac{\sigma}{\sqrt{n}} \leq \mu\right\}$$

$$LSC(\bar{X}) = \left\{\mu : \mu \leq \bar{X} + z_\alpha\frac{\sigma}{\sqrt{n}}\right\}$$

Ampiezza Campionaria

La lunghezza dell'intervallo IC_α è

$$a = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Se vogliamo avere un confidenza del $100(1-\alpha)\%$ che la stima abbia un errore

$$|\bar{X} - \mu| < \varepsilon_0$$

chiederemo che

$$\frac{a}{2} < \varepsilon_0$$

e questo vale se

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{\varepsilon_0}\right)^2$$

IC per la media, ignota σ^2

Se la varianza è ignota si usa S al posto di σ e si ottiene l'intervallo:

$$\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

Commenti

- L'ampiezza dell'intervallo è stocastica:
$$A = 2t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$
- Per $\sigma = S$
$$A > a$$
- Esercizio 4.c) e d) tema d'esame del 11-11-02.

Intervallo di confidenza per una percentuale

Grandi campioni

Dalla normalità approssimata di $\hat{\pi}$:

$$\frac{\hat{\pi}-\pi}{\sqrt{\pi(1-\pi)/n}} \cong N(0,1)$$

si ha, l'intervallo di approssimata confidenza $1-\alpha$:

$$\hat{\pi}-z_{\alpha/2}\frac{\sqrt{\hat{\pi}(1-\hat{\pi})}}{\sqrt{n}} \leq \pi \leq \hat{\pi}+z_{\alpha/2}\frac{\sqrt{\hat{\pi}(1-\hat{\pi})}}{\sqrt{n}}.$$

Piccoli Campioni

Consideriamo i percentili di ordine $\frac{\alpha}{2}$ ed $1-\frac{\alpha}{2}$ della $v.c.\ X,\ Bin(n,\pi)$. In particolare indichiamo con $x_1(\pi)$ il valore che meglio soddisfa l'approssimazione

$$P(X \leq x_1) \stackrel{\leq}{\cong} \frac{\alpha}{2}$$

e $x_2(\pi)$ analogamente

$$P(X \geq x_2) \stackrel{\leq}{\cong} \frac{\alpha}{2}.$$

Abbiamo allora la relazione

$$P_\pi(x_1(\pi) \leq n\hat{p} \leq x_2(\pi)) \geq 1-\alpha.$$

Siano L_i le soluzioni delle 2 equazione in π date da

$$x_i(\pi) = n\hat{p}, \quad i = 1,2$$

si ottiene che

$$P_\pi(L_1 \leq \pi \leq L_2) \geq 1-\alpha,$$

da cui l'intervallo di confidenza

$$L_1 \leq \pi \leq L_2.$$

NB:

1. il calcolo di x_i ed L_i si basa sulle tavole della Bin o con l'uso di software
2. **Esercizio Matlab**: che differenza c'e' fra gli IC di binofit e normfit ?
3. Se π **è piccolo**, l'approssimazione con la normale non vale e si usano i percentili della **Poisson** con la stessa logica vista per la Binomiale.

Intervallo di confidenza per σ^2

Ricordiamo che la distribuzione della varianza campionaria è di tipo χ^2 :

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}.$$

Perciò possiamo scrivere

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} \leq \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha.$$

o, equivalentemente,

$$P\left(S^2\frac{(n-1)}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq S^2\frac{(n-1)}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) = 1-\alpha$$

da cui l'IC

$$S^2\frac{(n-1)}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le S^2\frac{(n-1)}{\chi^2_{1-\frac{\alpha}{2},n-1}}$$

o

$$IC(S^2) = \left[S^2\frac{(n-1)}{\chi^2_{\frac{\alpha}{2},n-1}}, S^2\frac{(n-1)}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right].$$

Intervalli di confidenza asintotici

Se abbiamo uno stimatore di θ asintoticamente normale

$$\hat\theta \cong N\left(\theta, \frac{\tau^2}{n}\right)$$

allora, per n grande, l'IC approssimato per θ è dato da

$$\hat\theta \mp z_{\frac{\alpha}{2}}\frac{\tau}{\sqrt{n}}.$$

Esempio: $\hat\theta = \hat\theta_{ML}$.

Verifica di ipotesi

Esempio: Verifica della correttezza di una bilancia commerciale

Siano

$$X_1,\dots,X_{16} \quad iid \quad N(\mu,\sigma^2)$$

$n = 16$ pesate in g di un peso campione corrispondente a $\mu_0 = 30g$.

Le pesate sono effettuate tramite una bilancia la cui precisione è data da $\sigma = 3mg$.

Errori di misura: $Y_i = X_i - \mu_0$

Errore medio $\bar Y = \frac{1}{n}\sum_i Y_i$

- La bilancia è corretta se vale $H_0 : \mu_y = 0$
- La bilancia è distorta (contro il cliente) se $H_1 : \mu_y > 0$

Se H_0 è vera

$$\bar Y \text{ è } N\left(0, \frac{\sigma^2}{n}\right)$$

Supponiamo che l'esperimento di cui sopra abbia dato

$$\bar y = 1mg.$$

Ci chiediamo se, alla luce di questo risultato possiamo considerare

La bilancia corretta ?

o dobbiamo concludere che

la bilancia è distorta

Equivalentemente: dobbiamo concludere che

H_0 è vera ?

oppure

H_0 è falsa ?

Significatività Osservata

Per valutare se l'ipotesi H_0 è **credibile** alla luce dei dati, cioè se è **compatibile** con il dato osservato ($\bar{y} = 1mg$)

confrontiamo $\bar{y} = 1mg$ con la sua distribuzione nell'ipotesi che H_0 sia vera cioè $N\left(0, \frac{3}{\sqrt{16}}\right)$ e calcoliamo la Significatività Osservata o p-value. Nel nostro caso

$$\hat{\alpha} = P(\bar{Y} > 1|H_0) = 1 - \Phi\left(\frac{1}{3/\sqrt{16}}\right) = 1 - \Phi(1.33) = 9.1\%$$

in generale se $H_1 : \mu_y > 0$, osservato un certo valore di $\bar{y}$ la significatività osservata è data da:

$$\hat{\alpha} = P(\bar{Y} > \bar{y}|H_0) = P\left(N\left(0, \frac{\sigma^2}{n}\right) > \bar{y}\right) = 1 - \Phi\left(\frac{\bar{y}}{\sigma/\sqrt{n}}\right)$$

NB: Più è piccolo $\hat{\alpha}$, meno è credibile l'ipotesi H_0

Test unilaterale sulla media

Sia $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$ e consideriamo un sistema di ipotesi **unilaterale** (destro)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

Come nell'esempio della bilancia valori alti di $\bar{X}$ sono contro H_0 e la regola di decisione o test sarà del tipo **ad una coda**

$\bar{X} > x_c \Rightarrow$ rifiuto H_0.

Decisioni in condizioni di incertezza

Errori di decisione

	Stato Effettivo:	
Decisione:	è vera H_0	è falsa H_0
Accetto H_0		II tipo
Rifiuto H_0	I tipo	

Decisioni in condizioni di incertezza

Errori di decisione

Decisione:	Stato Effettivo: è vera H_0	è falsa H_0
Accetto H_0		II tipo
Rifiuto H_0	I tipo	

associati ai 2 errori abbiamo i rischi detti, rispettivamente, del fornitore e del cliente

$$\alpha \geq P(\text{errore I tipo}) = \text{rischio di I tipo}$$

$$\beta \geq P(\text{errore II tipo}) = \text{rischio di II tipo.}$$

Test unilaterale sulla media

In questo caso il rischio di I tipo è

$$\alpha = P(\bar{X} > x_c | H_0) = 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma/\sqrt{n}}\right)$$

mentre, per un particolare valore $\mu_1 > \mu_0$, quello di II tipo è

$$\beta(\mu_1) = P(\bar{X} \leq x_c | \mu_1) = \Phi\left(\frac{x_c - \mu_1}{\sigma/\sqrt{n}}\right).$$

Come scegliere x_c?

Approccio decisionale:

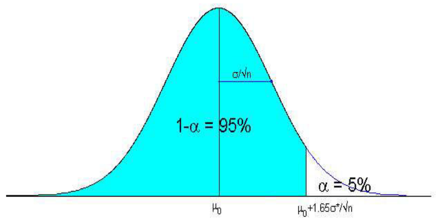
Fissiamo un livello massimo del rischio di I tipo, α, che chiamiamo **SIGNIFICATIVITA' NOMINALE** del test.

Vogliamo una regola decisionale in cui il rischio di I tipo non superi α (per esempio $\alpha = 5\%$ o 0.1%)

Il risultato è la **regola decisionale** data dalle 2 regioni:

Regione di Accettazione (A) : $\bar{X} \leq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$

Regione di Rifiuto $(R = \bar{A})$: $\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$



Il test di livello α è quindi caratterizzato dal **valore critico**

$$x_c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

dove

$$\alpha = 1 - \Phi(z_\alpha)$$

Significatività osservata e significatività nominale

Equivalente è la seg. regola decisionale data sulla significatività osservata:

$$\begin{aligned} &accetto\ H_0 \quad se\ \hat{\alpha} \geq \alpha \\ &rifiuto\ H_0 \quad se\ \hat{\alpha} < \alpha \end{aligned}$$

NB: R è anche detta **regione critica** del test.

- $\hat{\alpha}$ è **la più piccola** significatività nominale per **rifiutare** H_0
- $\hat{\alpha}$ è **la più grande** significatività nominale per **accettare** H_0

Rischi

Se $\mu \leq \mu_0$:

$$\begin{aligned} \alpha(\mu) &= P\big(\text{errore di I tipo}\big) \\ &= P\left(\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} | \mu\right) = 1 - \Phi\left(z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \leq \alpha \end{aligned}$$

Se è vera H_1 e $\mu > \mu_0$:

$$\begin{aligned} \beta(\mu) &= P\big(\text{errore di II tipo}\big) \\ &= P\left(\bar{X} \leq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} | \mu\right) = \Phi\left(z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \leq 1 - \alpha \end{aligned}$$

Potenza di un test

In generale si definisce la potenza:

$$\pi(\mu) = 1 - \beta(\mu).$$

In particolare, per il test unilaterale sulla media,

$$\begin{aligned} \pi(\mu) &= P\left(\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} | \mu\right) = 1 - \Phi\left(z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi(z_\alpha - \sqrt{n}\,\delta) \\ &= \Phi(\sqrt{n}\,\delta - z_\alpha) \end{aligned}$$

dove

$$\delta = \frac{\mu - \mu_0}{\sigma}.$$

La potenza misura la capacità del test di riconoscere la falsità di H_0.

Esempio della Bilancia

Se facciamo un test a livello $\alpha = 5\%$ e ci interesse una distorsione pari a $\mu_1 = 2mg$ la potenza del test basato su $n = 16$ prove è

$$\begin{aligned} \pi(\mu_1) &= 1 - \Phi\left(1.65 - \sqrt{16}\,\frac{2-0}{3}\right) \\ &= 1 - \Phi(-1.02) \cong 84.5\% \end{aligned}$$

Note:

Fissato μ ed α

1. In generale: un test si dice *Corretto* o **nondistorto** se
$$\pi(\mu) \geq \alpha$$
nel ns. caso vale infatti $\alpha \leq \pi = \Phi(\sqrt{n}\,\delta - z_\alpha) \leq 1$
2. fissato n, $\pi = \pi(\delta) \uparrow 1$ per $\delta \to \infty$
3. fissato n, $\pi = \pi(\delta) \downarrow \alpha$ per $\delta \to 0$
4. fissato δ, $\pi = \pi_n \uparrow 1$ per $n \to \infty$
5. Un test si dice **consistente** se, fissato α, $\forall \delta > 0$, si ha
$$\pi_n \to 1 \quad per \quad n \to \infty.$$

Determinazione di n

Si consideri un sistemi di ipotesi semplici
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_1 > \mu_0$$

Fissato α è possibile imporre una certa potenza per esempio $\pi \geq \pi^*$?

Ricordando la definizione di $\pi = 1 - \beta$ si ha
$$\pi^* \leq \pi(\mu_1) = \Phi(\sqrt{n}\,\delta - z_\alpha)$$
e risolvendo la disuguaglianza rispetto ad n si ottiene
$$\begin{aligned} n &\geq \left((z_\alpha + \Phi^{-1}(\pi^*)) \frac{\sigma}{\mu_1 - \mu_0} \right)^2 \\ &\geq \left(\frac{z_\alpha - \Phi^{-1}(\beta^*)}{\delta} \right)^2 \\ &\geq \left(\frac{z_\alpha + z_{\beta^*}}{\delta} \right)^2 \end{aligned}$$

Problemi e osservazioni

1. $\pi(\mu) > \pi^* \quad \forall \mu > \mu_1$
2. Discutere $n = n(\alpha, \pi, \sigma, \mu_1 - \mu_0)$.
3. Discutere il "Contratto Cliente-Fornitore".
4. Calcolare n quando $\alpha = \beta$.

Problema della Bilancia

Determinare n in modo che il test a livello $\alpha = 5\%$ abbia una potenza almeno del 95% per distorsioni pari o superiori a $\mu_1 = 2mg$

Ricorda: sotto H_0 la $\bar{Y}$ è $N\left(\mu_0 = 0, \frac{\sigma^2}{n}\right)$

Test bilaterale sulla media

Esempio

Verifica della correttezza di una bilancia scientifica

come l'esempio della bilancia commerciale ma ora

- La bilancia è corretta se vale $H_0 : \mu_y = 0$
- La bilancia è distorta se $H_1 : \mu_y \neq 0$

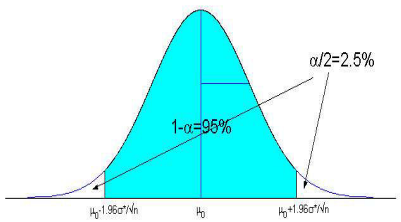
In generale

Il sistema di ipotesi è bilaterale

$$H_0 : \mu = \mu_0 \quad contro \quad H_1 : \mu \neq \mu_0$$

Ora il test è detto **a due code e** la **regione di accettazione,** è data da

$$\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



Regione di Rifiuto

Se $\bar{X}$ è esterno all'intervallo di accettazione si rifiuta l'ipotesi nulla:

$$\bar{X} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad oppure \quad \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}$$

in corrispondenza il rischio di I tipo è la somma delle due code

$$\alpha = P(R|H_0) = P\left(\bar{X} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\mu_0\right) + P\left(\bar{X} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\mu_0\right).$$

Potenza

$$\pi(\mu) = P(R|\mu) = P\left(\bar{X} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\mu\right) + P\left(\bar{X} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\mu\right)$$

Se $\mu > \mu_0$

$$\pi(\mu) \cong P\left(\bar{X} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}|\mu\right) = 1 - \Phi\left(z_{\frac{\alpha}{2}} - \sqrt{n}\,\delta\right)$$

dove

$$\delta = \frac{\mu - \mu_0}{\sigma}.$$

Problemi

1. Calcolare $\pi(\mu)$ per $\mu < \mu_0$
2. Riflettere se e quando è meglio un test unilaterale o bilaterale:
   1. in termini di ipotesi
   2. in termini di potenza

Test t sulla media ignota la varianza

$X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$, σ^2 ignoto

$$H_0 : \mu = \mu_0$$

Si usa la statistica t

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

che sotto H_0 ha distribuzione t_{n-1}.

Test unilaterale

$$H_1 : \mu > \mu_0$$

La regione critica è data da

$$t > t_{n-1,\alpha}$$

Significatività osservata o p-value, osservato $t = t^\circ = \frac{\bar{x}^\circ - \mu_0}{S/\sqrt{n}}$:

$$\hat{\alpha} = P(t_{n-1} > t^\circ)$$

Problema:
Costruire il test unilaterale sinistro

Test bilaterale

Se l'ipotesi alternativa è di tipo bilaterale

$$H_1 : \mu \neq \mu_0$$

abbiamo un test a due code e la regione critica è data da

$$|t| > t_{n-1,\frac{\alpha}{2}}$$

Test e IC

Richiamiamo l'intervallo di confidenza per la media μ a livello α :

$$\mathbf{IC}_\alpha(\bar{x}) = \left\{ \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}$$

che ha forma simile alla regione di accettazione $\mathbf{A}$ del test bilaterale per $H_0 : \mu = \mu_0$

$$\mathbf{A}_\alpha(\mu_0)=\left\{ \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}.$$

Si verifica immediatamente che

$$\mu_0 \in \mathbf{IC}_\alpha(\bar{x}) \Leftrightarrow \bar{x} \in \mathbf{A}_\alpha(\mu_0).$$

Cioè: un campione con una media $\bar{x}$ porterebbe ad accettare tutte le ipotesi nulle μ_0 che stanno in $IC(\bar{x})$.

Test unilaterale sulla varianza

Consideriamo un campione normale, $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$ ed il sistema di ipotesi sulla varianza dato da

$$H_0 : \sigma^2 = \sigma_0^2 \qquad contro \qquad H_1 : \sigma^2 > \sigma_0^2$$

La regione di rifiuto del test a livello α è data da valori alti della varianza campionaria

$$S^2 > \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\alpha}$$

osservato un certo valore s^2 la significatività osservata è data da

$$\hat{\alpha} = P\left(\chi^2_{n-1} > s^2 \frac{n-1}{\sigma_0^2} \right)$$

Test bilaterale sulla varianza

Se il sistema di ipotesi è

$$H_0 : \sigma^2 = \sigma_0^2 \qquad contro \qquad H_1 : \sigma^2 \neq \sigma_0^2$$

allora la **regione di accettazione** è

$$\frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\frac{\alpha}{2}} \leq S^2 \leq \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\frac{\alpha}{2}}$$

Problema:

Il test non è simmetrico come quello sulla media, in particolare riflettere sulla differenza fra test unilaterale destro e sinistro.

Test su una percentuale

Consideriamo un campione $X_1, \ldots, X_n$ *iid* $Bin(\pi)$ e il sistema di ipotesi

$$H_0 : \pi = \pi_0 \qquad contro \qquad H_1 : \pi > \pi_0$$

Caso 1

Se $0 << \pi_0 << 1$ e n non è piccolo, usando il limite centrale troviamo la regione di accettazione

$$\hat{\pi} \leq \pi_0 + z_\alpha \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

Problemi

- Qual è la regione di rifiuto del test a due code ?
- Discutere la differenza fra test ad una e a due code nel caso che π sia la percentuale di difettosi in un processo produttivo.

Caso 2

Se π_0 è piccolo ed è appropriata l'approssimazione con la distribuzione di Poisson,

si pone

$$\lambda_0 = n\pi_0$$

e la regione di accettazione è data da

$$\hat{\pi} \leq \frac{x_\alpha}{n}$$

dove x_α è il quantile di ordine $1-\alpha$ della distribuzione $Poisson(\lambda_0)$, è cioè il più piccolo x tale per cui $F_\lambda(x) \geq 1-\alpha$

Caso 3

n non è grande a sufficienza per usare le approssimazioni di cui sopra. Si usa allora direttamente la distribuzione binomiale in modo analogo al Caso 2.

Per esempio per il test bilaterale si ha la regione di accettazione

$$\frac{x_1(\pi_0)}{n} \leq \hat{\pi} \leq \frac{x_2(\pi_0)}{n}$$

dove $x_1(\pi)$ ed $x_2(\pi)$ si ottengono come nell'IC per la binomiale.

Test a due campioni

(opzionale)

I test ad un campione emergono spesso nella sperimentazione, nella ricerca e sviluppo e nel miglioramento della qualità per vedere se un'innovazione è migliorativa rispetto ad una situazione nota caratterizzata per esempio da una variabilità del tipo $N(\mu_0, \sigma^2)$.

Si conduce una sperimentazione o, comunque, si raccolgono dei dati $X_1, \ldots, X_n$ per valutare se l'ipotesi nulla

$$H_0 : \mu = \mu_0$$

sia accettabile oppure no.

Talvolta invece si vuole vedere se l'innovazione è utile ma non è ben nota la situazione antecedente.

In queste situazioni si prelevano due campioni uno (Y) *di controllo* senza l'innovazione in oggetto che porta informazioni sul sistema ante-innovazione, .l'altro (X) come sopra porta informazioni sull'innovazione in studio.

Sia hanno così due campioni indipendenti

$$X_1, \ldots, X_n \quad iid \quad N(\mu_x, \sigma_x^2)$$

ed

$$Y_1, \ldots, Y_m \quad iid \quad N(\mu_y, \sigma_y^2).$$

L'*inutilità* dell'innovazione è rappresentata dalla condizione $H_0 : \mu_x = \mu_y$. Si prende in considerazione l'innovazione solo se c'è forte evidenza che H_0 è falsa.

Test sulla media

Interessa l'ipotesi

$$H_0 : \mu_x = \mu_y$$

Varianza nota

Si usa la statistica

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

che sotto H_0 ha distribuzione $N(0, 1)$.

Si ottiene quindi la regione di accettazione (del test bilaterale)

$$-z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} \leq \bar{X} - \bar{Y} \leq z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

Caso omoschedastico

Se $\sigma_x^2 = \sigma_y^2 = \sigma^2$ allora ovviamente le formule si semplificano e si ottiene la regione di accettazione

$$-z_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n} + \frac{1}{m}} \leq \bar{X} - \bar{Y} \leq z_{\frac{\alpha}{2}} \sigma \sqrt{\frac{1}{n} + \frac{1}{m}}$$

Varianza ignota (test t a due campioni)

Consideriamo solo il **caso omoschedastico** $\sigma_x^2 = \sigma_y^2 = \sigma^2$.

La comune varianza σ^2 è stimata dalla media delle varianze campionarie

$$S_*^2 = \frac{S_x^2(n-1) + S_y^2(m-1)}{n+m-2}.$$

e per il test a una o due code si usa la statistica t

$$t = \frac{\bar{X} - \bar{Y}}{S_* \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

che ha $n + m - 2$ gradi di libertà

Test t per dati accoppiati

Se le X e le Y sono osservate sulle stesse unità statistiche lipotesi di inidpendenza viene meno allora si considera

$$Z_i = X_i - Y_i$$

e l'ipotesi di uguaglianza è ora esprimibile come

$$H_0 : E(Z) = 0$$

Problemi

- cosa cambia rispetto al test a due campioni ?
- quale è meglio ?

Test sulla varianza

(opzionale)

Interessa ora

$$H_0 : \sigma_x^2 = \sigma_y^2.$$

Sotto H_0, la statistica

$$F = \frac{S_x^2}{S_y^2}$$

ha distribuzione F di Snedocor con $n-1$ ed $m-1$ gradi di libertà, $F_{n-1,m-1}$.

La regione di accettazione per il test bilaterale è quindi

$$F_{n-1,m-1,1-\frac{\alpha}{2}} \leq F \leq F_{n-1,m-1,\frac{\alpha}{2}}.$$

Problemi:

- Costruire il test unilaterale destro per $H_1 : \sigma_x^2 > \sigma_y^2$
- Scrivere la potenza per l'alternativa

$$H_1 : \sigma_y^2 = k\sigma_x^2,\ k > 1$$

  nel test unilaterale.
- Per il sistema di ipotesi di cui sopra, si considerino i test di dimensione α, con regioni di rifiuto date da

$$\frac{S_X^2}{S_Y^2} > r_\alpha \qquad e \qquad \frac{S_Y^2}{S_X^2} < r'_\alpha$$

  rispettivamente. **Mostrare che**
  - sono equivalenti e, in particolare,
  - hanno la stessa potenza.

Test Asintotici

Stima di θ asintoticamente normale

$$\hat{\theta}_n \quad approx \quad N\left(\theta, \frac{\tau^2}{n}\right)$$

$$H_0 : \theta = \theta_0$$

Regione di Accettazione:

$$\theta_0 - z_{\alpha/2}\frac{\hat{\tau}_0}{\sqrt{n}} < \hat{\theta} < \theta_0 + z_{\alpha/2}\frac{\hat{\tau}_0}{\sqrt{n}}$$

dove $\hat{\tau}_0$ è uno stimatore consistente di τ sotto H_0.

Modelli Empirici

Premessa - In questa parte impareremo a costruire un modello empirico per spiegare il legame fra una grandezza sulla base dei dati osservati. Per esempio studieremo la qualità di un processo in funzione di uno o più fattori mediante una relazione del tipo

$$y = f(\mathbf{x}).$$

Seguendo l'approccio statistico non pretenderemo che la relazione valga per ogni coppia di dati osservati $(\mathbf{x}_i, y_i)$ ma che **valga in media**

$$E(y|\mathbf{x}) = f(\mathbf{x}).$$

Considereremo il caso in cui sia la variabile dipendente o **risposta** y che le variabili **esplicative** o **regressori** $\mathbf{x}$ sono di tipo continuo. Inoltre come

notazione useremo sempre le minuscole anche per le v.c.: Riserviamo le maiuscole X ed Y per alcune speciali matrici. Useremo il **neretto minuscolo** per alcuni speciali vettori.

Correlazione

Dati: n coppie di valori o punti del piano

$$(x_1, y_1), \ldots, (x_n, y_n)$$

Coefficiente di correlazione lineare campionario

$$r = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2 \sum_{t=1}^{n}(y_t - \bar{y})^2}} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$$

dove il numeratore è la covarianza campionaria:

$$\hat{\sigma}_{xy} = \frac{1}{n}\sum(x_t - \bar{x})(y_t - \bar{y})$$

Proprietà

$$-\hat{\sigma}_x\hat{\sigma}_y \le \hat{\sigma}_{xy} \le \hat{\sigma}_x\hat{\sigma}_y$$

$$-1 \le r \le 1$$

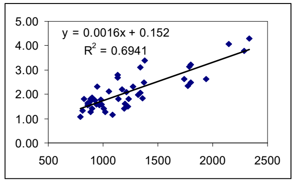
NB: $\hat{\sigma}_{xy}$ è detta covarianza campionaria e stima la vera covarianza

$$\sigma_{xy} = Cov(x, y)$$

Inoltre r, il coefficente di correlazione campionario, è stima di ρ, il coefficiente di correlazione *"nella popolazione"*

$$\rho = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

Regressione Lineare



Descriviamo il legame fra la x e la y tramite un legame lineare

$$y = \beta_0 + \beta_1 x + \epsilon$$

che vale a meno di un errore stocastico non osservabile ε.

In particolare $E(\epsilon) = 0, \quad Var(\varepsilon) = \sigma^2 = \sigma_\varepsilon^2$ e *spesso* assumeremo che ε è $N(0,\sigma^2)$.

Nel seguito le x saranno v.c. o valori noti ma in ogni caso saranno **indipendenti** da ε.

Con i dati campionari:

$$(x_1,y_1),\ldots,(x_n,y_n)$$

si stimano a **Minimi Quadrati** i parametri $\boldsymbol{\beta} = (\beta_0,\beta_1)'$:

$$Q(\beta) = \sum_{t=1}^{n}(y_t - \beta_0 - \beta_1 x_t)^2$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} Q(\beta)$$

In particolare si ha la seg. soluzione esplicita

$$\hat{\beta}_1 = b = \frac{\sum(x_t-\bar{x})(y_t-\bar{y})}{\sum(x_t-\bar{x})^2} \qquad e \qquad \hat{\beta}_0 = a = \bar{y} - b\bar{x}.$$

Noto $\hat{\beta}$ si possono calcolare i **valori interpolati**:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$$

che hanno media uguale a quella dei dati

$$\frac{1}{n}\sum \hat{y}_t = \bar{y}.$$

Inoltre si possono calcolare **i residui** della regressione

$$e_t = y_t - \hat{y}_t$$

che hanno somma nulla e sono importanti per comprendere le proprietà del modello trovato e stimare σ_ε^2 :

$$s^2 = \frac{1}{n-2}\sum e_t^2.$$

Problemi

1. Dimostrare che $\hat{\boldsymbol{\beta}} = (a,b)$.
   Suggerimento: risolvere il sistema
   $$\frac{\partial}{\partial\beta_0}Q(\boldsymbol{\beta}) = 0$$
   $$\frac{\partial}{\partial\beta_1}Q(\boldsymbol{\beta}) = 0$$
2. $\hat{\sigma}_{xy} = \frac{1}{n}\sum(x_t-\bar{x})(y_t-\bar{y})$ è detta covarianza campionaria e stima $\sigma_{xy} = Cov(x,y)$
3. Verificare che
   $$b = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}.$$
4. Osservare che $\quad \beta_1 = \frac{\sigma_{xy}}{\sigma_x^2}.$

Adattamento dei dati al modello

Devianza di totale di y

$$D_{tot} = \sum (y_t - \bar{y})^2$$

Devianza residua

$$D_{res} = \sum (y_t - \hat{y}_t)^2$$

Devianza spiegata

$$D_{sp} = \sum (\hat{y}_t - \bar{y})^2$$

Scomposizione Devianza:

Se $\hat{y} = a + bx$ con a e b a minimi quadrati, allora

$$D_{tot} = D_{res} + D_{sp}$$

Coefficiente di determinazione

Dalla scomposizione Devianza

$$D_{tot} = D_{res} + D_{sp}$$

si ha che l'indice di determinazione

$$R^2 = 1 - \frac{D_{res}}{D_{tot}}$$

nella regressione lineare con intercetta R^2 è interpretabile come % di varianza spiegata, infatti:

$$R^2 = 1 - \frac{D_{res}}{D_{tot}} = \frac{D_{sp}}{D_{tot}} \quad \text{con} \quad 0 \le R^2 \le 1$$

Correlazione e varianza residua

Nel modello retta di regressione con intercetta, il coefficiente di correlazione al quadrato

$$r^2 = \left(\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} \right)^2$$

e l'indice di determinazione

$$R^2 = \frac{D_{sp}}{D_{tot}}$$

sono uguali.

Per vedere ciò, si osserva che:

$$\hat{\sigma}^2(\hat{y}) = \hat{\sigma}^2(a + bx) = b^2 \hat{\sigma}(x) = \left(\frac{\hat{\sigma}(x,y)}{\hat{\sigma}^2(x)} \right)^2 \hat{\sigma}^2(x) = r^2 \hat{\sigma}(y)$$

da cui

$$r^2 = \frac{\hat{\sigma}^2(\hat{y})}{\hat{\sigma}^2(y)} = R^2.$$

In questo ambito la devianza residua può essere calcolata come:

$$D_{res} = \sum_i e_i^2 = (1 - r^2) D_{tot}$$

e la varianza residua

$$s^2 = \frac{D_{res}}{n - 2} = (1 - r^2) \frac{D_{tot}}{n - 2}$$

Correlazione e incorrelazione

1. Coefficiente di correlazione lineare campionario
$$r = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum (x_t - \bar{x})^2 \sum (y_t - \bar{y})^2}} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$$
2. r è stima di ρ, il coefficiente di correlazione *"nella popolazione"*
$$\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$
3. $\beta = 0$ sse $\rho = 0$
4. $b = 0$ sse $r = 0$
5. $\rho^2 =$ % di varianza spiegata nella popolazione
6. $r^2 =$ % di varianza spiegata nel campione
7. sotto $H_0 : \rho = 0$, per n grande, $r \approx N(0, \frac{1}{n})$.

Problemi:

1. Correlazione o Causalità ?
2. Dimostrare che $\rho^2 = \frac{Var(\beta_0 + \beta_1 x)}{\sigma_y^2}$
3. quante rette a minimi quadrati: 1 o 2 ?
4. Regressione inversa e taratura: x misurando e y misura.
5. **Matlab:** regress, regstat, lsline

Premesse Regressione Multipla

Vettori stocastici e Matrice di Varianze-covarianze

Sia $\mathbf{x} = (x_1, \ldots, x_k)'$ un vettore casuale $k - \dim$.

La media di $\mathbf{x}$ è un vettore:
$$\mu = E(\mathbf{x}) = (\mu_1, \ldots, \mu_k)'$$
La matrice di varianze-covarianze contiene le varianze
$$\sigma_i^2 = Var(x_i)$$
e le covarianze
$$\sigma_{i,j} = cov(x_i, x_j).$$

E' data da
$$\Sigma = (\sigma_{i,j})_{i,j=1,\ldots,k} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \cdots & \sigma_{1,k} \\ \sigma_{2,1} & \sigma_2^2 & & & \sigma_{2,k} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \sigma_{k,1} & \sigma_{k,2} & \cdots & \cdots & \sigma_k^2 \end{pmatrix}$$

NB:

1. $\sigma_i^2 = \sigma_{i,i}$
2. E' simmetrica: $\sigma_{i,j} = \sigma_{j,i}$
3. E' semidefinita positiva: $\det(\Sigma) \geq 0$

Varianza di una combinazione lineare

Se $\mathbf{x}$ è un vettore casuale $k-\dim$ con matrice di varianze-covarianze Σ, e b è un vettore non casuale allora

$$Var(b'\mathbf{x}) = b'\Sigma b.$$

Regressione Multipla

(cenni)

Consideriamo il Modello Lineare

$$\begin{aligned} y &= \boldsymbol{\beta}'\mathbf{x} + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \end{aligned}$$

dove

1. $x = (x_1, \ldots, x_k)'$ variabili esplicative o regressori ed $\mathbf{x} = (1, x')'$;
2. ε è $N(0, \sigma^2)$, σ^2 parametro di disturbo per $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$

Esempi:

1. Modelli senza intercetta, $\beta_0 = 0$
2. Modelli polinomiali

Stima a Minimi Quadrati

Date n osservazioni $(x_1, y_1), \ldots, (x_n, y_n)$ si ha la forma matriciale,

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

data da

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

dove

- $Y = (y_1, \ldots, y_n)'$ è $n \times 1$
- X è $n \times (k+1)$, ha come $i-esima$ riga $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ik})$
- $\boldsymbol{\varepsilon}$ è $n \times 1$

Usando i minimi quadrati

$$Q(\beta) = \sum (y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2 = (Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta})$$

definiamo stima **Least Squares** (LS) la soluzione dei minimi quadrati:

$$\hat{\boldsymbol{\beta}}_{LS} = \hat{\boldsymbol{\beta}} = \arg\min Q(\beta).$$

che ha forma esplicita data da

$$\hat{\boldsymbol{\beta}}_{LS} = (X'X)^{-1}X'Y.$$

L'espressione di $\hat{\boldsymbol\beta}$ si trova dalle $k+1$ condizioni del prim'ordine:

$$\frac{\partial Q}{\partial \beta_j} = -2\sum x'_{ij}(y_i - \boldsymbol\beta'\mathbf{x}_i) = 0$$

o, in forma matriciale,

$$\frac{\partial Q}{\partial \boldsymbol\beta} = -2X'(Y - X\boldsymbol\beta) = \mathbf{0}.$$

Si ha così il sistema detto delle **eq**^{ni} **normali**

$$X'X\hat{\boldsymbol\beta} = X'Y$$

che ha, appunto, soluzione

$$\hat{\boldsymbol\beta}_{LS} = (X'X)^{-1}X'Y.$$

Esistenza

La condizione $\det(X'X) > 0$ è sempre soddisfatta a meno che una o più colonne della matrice X non sia una combinazione lineare delle altre. Supponiamo, per esempio, che l'ultima colonna sia una tale combinazione:

$$x_{ik} = \sum_{j=1}^{k-1} a_j x_{ij}$$

allora l'osservazione della corrispondente variabile esplicativa x_k non porta informazioni aggiuntive rispetto alle altre per il sistema che si sta studiando.e va eliminata dal modello.

Problemi

1. dato $\hat{y} = x\beta + \varepsilon$, con $x \in R^1$ mostrare che $\hat\beta = \frac{\sum x_t y_t}{\sum x_t^2}$
2. $\hat{y} = \beta_0 + x_1\beta_1$ calcolare $X'X$. ed $(X'X)^{-1}$.
3. $f(x;\beta) = \beta_0 + \beta_1 x + \ldots + \beta_k x^k$ con $\beta = (\beta_0, \ldots, \beta_k)'$: studiare $\hat\beta$.
4. Posto $\hat\Sigma = (\hat\sigma_{ij})$ dimostrare che

$$\hat\beta_0 = \bar{y} - \sum_{j=1}^{k} \hat\beta_j \bar{x}_j$$

e

$$\begin{pmatrix} \hat\beta_1 \\ \vdots \\ \hat\beta_k \end{pmatrix} = \hat\Sigma^{-1} \begin{pmatrix} \hat\sigma(x_1, y) \\ \vdots \\ \hat\sigma(x_k, y) \end{pmatrix}$$

Proprietà della stima LS

1 - Nondistorsione

$$E\left(\hat{\boldsymbol\beta}\right) = \boldsymbol\beta$$

2 - Matrice di varianze-covarianze

$$Var\left(\hat{\boldsymbol\beta}\right) = \sigma_\varepsilon^2 (X'X)^{-1}$$

- $\det(X'X) \cong 0 \Rightarrow$ stime scadenti.
- Posto

$$v = diag\left((X'X)^{-1}\right)$$

si ha per $j = 0, .., k$

$$Var\left(\hat\beta_j\right) = \sigma_\varepsilon^2 v_{j+1}$$

3 - Normalità

- piccoli campioni ε *iid* $N(0,\sigma^2) \Rightarrow \hat{\boldsymbol{\beta}}$ è $N_{k+1}\left(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}\right)$
- grandi campioni: se

$$\frac{1}{n}X'X \to \Gamma > 0$$

allora, se ε *iid* $(0,\sigma^2)$ con $E(\varepsilon^4) < \infty$, vale il seg. TLC:

$$\hat{\boldsymbol{\beta}} \cong N\left(\beta, \frac{\sigma^2}{n}\Gamma^{-1}\right).$$

Scomposizione della Varianza

(opzionale)

In modo analogo alla retta di regressione abbiamo:

Devianza totale

$$D_{tot} = \sum(y_t - \bar{y})^2 \approx \sigma^2\chi^2_{n-1}$$

Devianza residua

$$D_{res} = \sum(y_t - \hat{y}_t)^2 \approx \sigma^2\chi^2_{n-k-1}$$

Devianza spiegata

$$D_{sp} = \sum(\hat{y}_t - \bar{y})^2 = D_{tot} - D_{res} \approx \sigma^2\chi^2_k$$

Adattamento

(opzionale)

Varianza Residua, stima di σ^2_ε :

$$s^2 = \frac{1}{n-k-1}D_{res}.$$

Coefficiente di Determinazione Multipla

$$R^2 = 1 - \frac{D_{res}}{D_{tot}}$$

sotto $H_0 : \beta = 0$, per n grande, nR^2 ha distribuzione approssimata di tipo χ^2_k.

Coefficiente di Correlazione Multipla

$$R = +\sqrt{R^2} = r(y,\hat{y})$$

Tuttavia quando $n-k$ non è grande si possono avere R^2 alti come solo effetto di interpolazione.

A tal fine si preferisce l'R^2 corretto:

$$\bar{R} = 1 - \frac{D_{res}/(n-k-1)}{D_{tot}/(n-1)}$$

Inoltre è opportuno formulare dei test per valutare la significatività del modello trovato dai minimi quadrati.

Analisi della Varianza e Test F

(opzionale)

Il modello è significativo ?

Interessa valutare la significatività del modello nel suo insieme:

$$H_0 : \beta_1 = \ldots = \beta_k = 0$$

A tal fine usiamo la statistica

$$F = \frac{D_{sp}/k}{D_{res}/(n-k-1)}.$$

In ipotesi di normalità, sotto H_0 la statistica F ha distribuzione F **di Snedecor** con k ed $n-k-1$ gradi.di libertà

$$F \sim F_{k,n-k-1}$$

Tabella di *ANOVA*

sorgente	DF	SS	MS	F	p –value
regressione	k	D_{sp}	$MS_{sp} = \frac{D_{sp}}{k}$	$\frac{MS_{sp}}{s^2}$	$P(F_{k,n-k-1} > F)$
errori	$n-k-1$	D_{res}	$s^2 = \frac{D_{res}}{n-k-1}$		
totale	$n-1$	D_{tot}	S_y^2		

Test t sui coefficienti

Interessa valutare la significatività dei singoli coefficienti β_j :

$$H_{0j} : \beta_j = 0$$

si usa la statistica t :

$$t = \frac{\hat{\beta}_j}{s\sqrt{v_{j+1}}}$$

dove

$$v = diag\left((X'X)^{-1}\right)$$

e t ha distribuzione t di Student con $n-k-1$ gradi di libertà.

Intervalli di Confidenza nella regressione

IC sui coefficienti

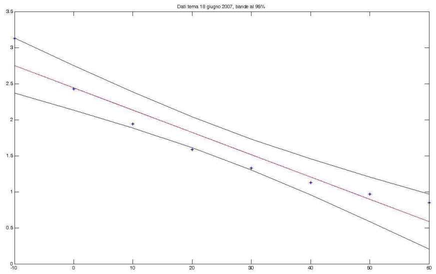
$$\hat{\beta}_j - t_{n-k-1,\frac{\alpha}{2}} s\sqrt{v_{j+1}} \le \beta_j \le \hat{\beta}_j + t_{n-k-1,\frac{\alpha}{2}} s\sqrt{v_{j+1}}$$

dove $t_{n-k-1,\frac{\alpha}{2}}$ è il valore critico della t di Student con $n-k-1$ gradi di libertà.

Grandi campioni

$$\hat{\beta}_j - z_{\frac{\alpha}{2}} s\sqrt{v_{j+1}} \le \beta_j \le \hat{\beta}_j + z_{\frac{\alpha}{2}} s\sqrt{v_{j+1}}$$

IC sulla superficie attesa



Dati Tema 18 giugno 2007

Interessa l'**IC** per

$$\mu_y(\mathbf{x}) = E(y|x)$$

in corrispondenza ad x non osservato. La sua stima LS è

$$\hat{\mu}_y(x) = \hat{\boldsymbol{\beta}}'\mathbf{x}$$

con varianza

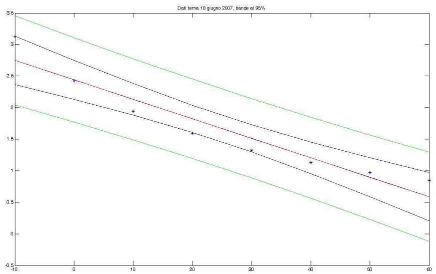
$$Var\left(\hat{\beta}'\mathbf{x}\right) = \sigma^2\mathbf{x}'(X'X)^{-1}\mathbf{x}.$$

Perciò l'**IC** per $\mu_y(\mathbf{x})$ è

$$\hat{\beta}'\mathbf{x} \mp t_{n-k-1,\frac{\alpha}{2}}s\sqrt{\mathbf{x}'(X'X)^{-1}\mathbf{x}}$$

dove $t_{n-k-1,\frac{\alpha}{2}}$ è il valore critico della t di Student con $n-k-1$ gradi di libertà.

IC sulle previsioni



Dati Tema 18 giugno 2007

Interessa l'**IC** per

$$y = \beta'\mathbf{x} + \boldsymbol{\varepsilon}$$

in corrispondenza ad una $\mathbf{x}$ non osservata.
La sua stima è

$$\hat{y} = \hat{\beta}'\mathbf{x}$$

osservando che

$$\hat{\beta} = \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon},$$

si ottiene la varianza dell'errore di previsione:

$$Var(y-\hat{y}) = \sigma^2\left[1 + \mathbf{x}'(X'X)^{-1}\mathbf{x}\right]$$

Perciò l'**IC** per $\hat{y}$ è

$$\hat{\beta}'\mathbf{x} \mp t_{n-k-1,\frac{\alpha}{2}} s \sqrt{1 + \mathbf{x}'(X'X)^{-1}\mathbf{x}}$$

dove $t_{n-k-1,\frac{\alpha}{2}}$ è il valore critico della t di Student con $n-k-1$ gradi di libertà.

Varianze per la retta di regressione

Varianza dell'intercetta:

$$Var\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}\right)$$
$$= \frac{\sigma^2}{\sum(x-\bar{x})^2} m_2(X)$$

Varianza del coeff. angolare:

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum(x-\bar{x})^2}$$

Varianza del valor medio $E(y|x)$

$$Var(\mu_y(x^0)) = \sigma^2\left(\frac{1}{n} + \frac{(x^0-\bar{x})^2}{\sum(x-\bar{x})^2}\right)$$

Varianza della previsione

$$Var(y - \hat{y}(x^0)) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x^0-\bar{x})^2}{\sum(x-\bar{x})^2}\right)$$

Dimostrazione e commenti

Osserviamo che

$$X'X = \begin{pmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{pmatrix}$$

Inoltre

$$\det(X'X) = n\Sigma x^2 - (\Sigma x)^2 = n\Sigma(x-\bar{x})^2$$

Perciò

$$(X'X)^{-1} = \frac{1}{n\Sigma(x-\bar{x})^2}\begin{pmatrix} \Sigma x^2 & -\Sigma x \\ -\Sigma x & n \end{pmatrix}$$

Segue quindi

$$
\begin{aligned}
Var\left(\hat{\beta}_0\right) &= \sigma^2 \frac{\Sigma x^2}{n\Sigma(x-\bar{x})^2} \\
&= \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\Sigma(x-\bar{x})^2}\right) \\
&= \frac{\sigma^2}{n}\left(1+\frac{\bar{x}^2}{\hat{\sigma}_x^2}\right) \\
&= \frac{\sigma^2}{n}\frac{m_2(x)}{\bar{m}_2(x)}
\end{aligned}
$$

$$
\begin{aligned}
Var\left(\hat{\beta}_1\right) &= \frac{\sigma^2}{\Sigma(x-\bar{x})^2} \\
&= \frac{\sigma^2}{n}\frac{1}{\hat{\sigma}_x^2}
\end{aligned}
$$

Bibliografia

(MRH – *inglese*) D.C. Montgomery, G.C. Runger & N.F. Hubele (2000): **Engineering Statistics**, Second Edition, John Wiley&Sons, Inc. New York.

D.C. Montgomery, G.C. Runger & N.F. Hubele (2004): **Statistica per Ingegneria**, Egea

S. Ross (2003) **Probabilità e Statistica per l'Ingegneria e le Scienze**, Apogeo