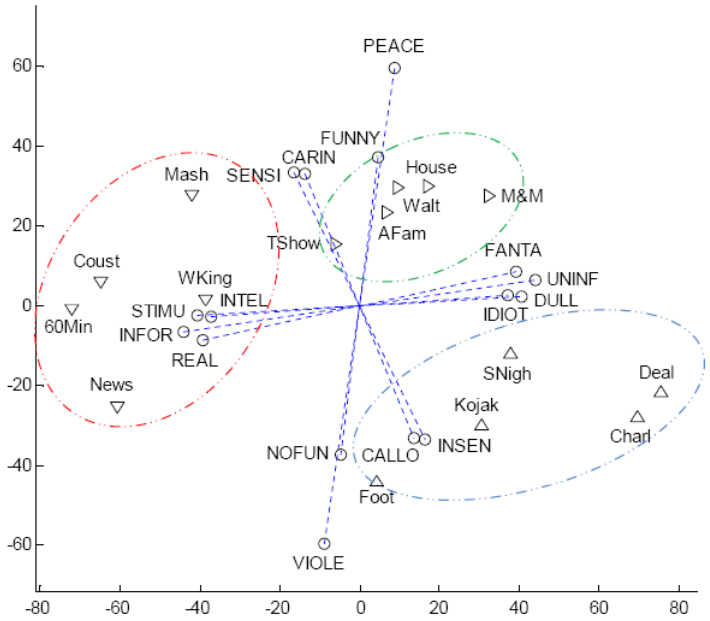


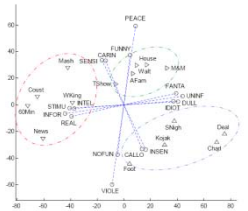
Metodi Statistici per il Management

Richiami di Inferenza Statistica

Simone Borra - Roberto Rocci

Indice

- Probabilità
- Variabili Casuali
- Metodologie dell'Inferenza Statistica



Probabilità

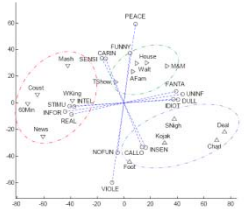
Siano dati i seguenti elementi:

- Ω un insieme, detto spazio campionario o fondamentale;
- $\mathfrak{I}$ una σ-*algebra* su Ω. Indicheremo con E i suoi elementi, sottoinsiemi di Ω detti *eventi*;
- **P** una funzione **P**: $\mathfrak{I} \rightarrow [0,1]$.

Definizione

Si dice che **P** è una (*misura di*) *probabilità* se:

- $P(E) \geq 0$;
- $P(\Omega) = 1$;
- per ogni successione numerabile disgiunta di eventi E_i, cioè con $E_i \cap E_j = \varnothing$, vale la numerabile *additività*, cioè $P\{\cup_i E_i\} = \Sigma_i \mathbf{P}\{E_i\}$.



Probabilità

Probabilità dell'unione

- $P(E \cup F) = P(E) + P(F) – P(E \cap F)$

Probabilità della negazione

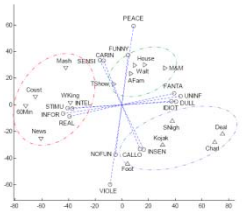
- $P(E^c) = 1- P(E)$

Probabilità condizionata

- $P(E|F) = P(E \cap F)/P(F)$

Eventi indipendenti

- $P(E \cap F) = P(E)P(F)$



Probabilità

Esempio

Studio delle probabilità di incasso delle fatture. Codici di pagamento:

PP = Puntuale, RP = Ritardato, MP = Mancato.

La terna $(\Omega, \mathfrak{I}, P)$ si può così formare

- Ω = {PP,RP,MP}

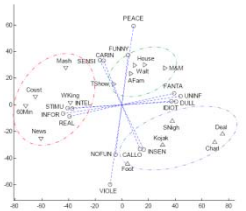
- Come σ-algebra $\mathfrak{I}$ la scelta più naturale e semplice, quando l'insieme Ω è finito, è di considerare per $\mathfrak{I}$ l'insieme $\wp(\Omega)$ delle parti di Ω:

$$\mathfrak{I} = \{ \{PP\},\{RP\},\{MP\},\{PP,RP\},\{PP,MP\},\{RP,MP\},\Omega, \varnothing \}.$$

- La funzione di probabilità **P** può essere scelta assegnando una probabilità ai tre eventi elementari che rispetti l'assioma **P**(Ω) = 1, e le probabilità agli altri eventi in modo coerente con l'assioma di additività di **P**. Procediamo con una delle possibili scelte

P{PP} = 0.6, **P**{RP} = 0.3, **P**{MP} = 0.1. Conseguentemente avremo:

P {PP,RP} = 0.6 + 0.3 = 0.9; **P** {PP,MP} = 0.6 + 0.1 = 0.7; **P** {RP,MP} = 0.3 + 0.1 = 0.4. Con le probabilità usuali di "contorno" **P** $\{\Omega\}$ = 1, **P** $\{\varnothing\}$ = 0.



Variabili casuali

Definizione

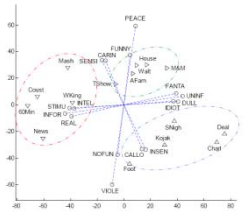
Siano dati i seguenti elementi:

- $(\Omega,\mathfrak{I},\mathbf{P})$ uno spazio di probabilità;
- $(\Re,\mathcal{R})$ l'insieme dei numeri reali e una σ-algebra $\mathcal{R}$ definita su di esso;
- un'applicazione $X: \Omega \rightarrow \Re$.

Se per ogni $C \in \mathcal{R}$, $X^{-1}(C) \in \mathfrak{I}$, allora X è una *variabile casuale*.

La misura di probabilità $\mathbf{P}_X$ indotta dalla variabile X è definita in modo naturale dalla misura di partenza **P** mediante

$$\mathbf{P}_X(C) = \mathbf{P}\{\omega \in \Omega : X(\omega) \in C\} = \mathbf{P}\{X^{-1}(C)\}, \text{ per ogni } C \in \mathcal{R}$$



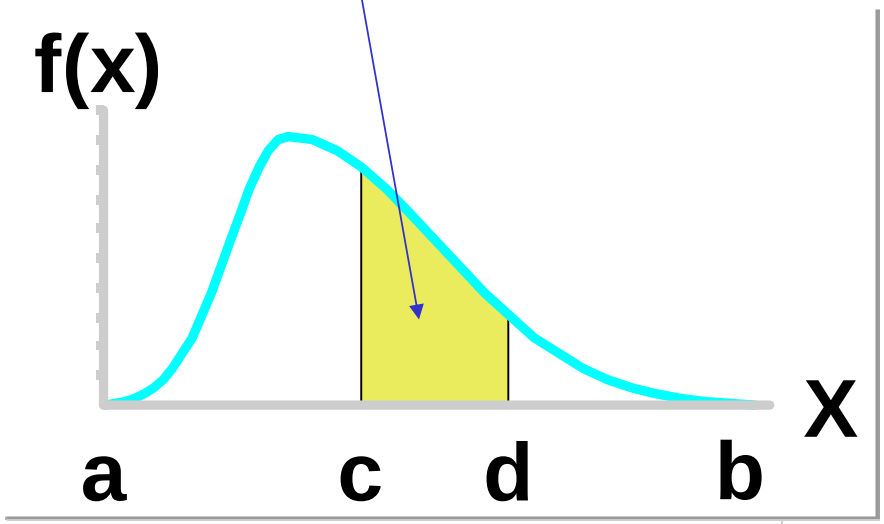
Variabili casuali

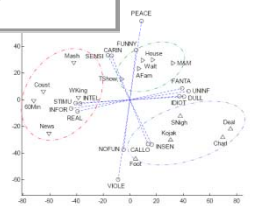
Definizioni

▪Una variabile casuale *X* si dice **discreta** se assume valori in un insieme *E* di cardinalità **finita** o **numerabile**. La sua distribuzione di probabilità è descritta dalla funzione di probabilità (assegna la probabilità ad ogni possibile realizzazione).

▪Una variabile casuale è detta (*assolutamente*) **continua** se assume valori in un intervallo dei reali. Le probabilità sono assegnate ad intervalli e non a singoli punti mediante la funzione di densità di probabilità (ogni punto ha probabilità zero!)

$$P(c \le X \le d) = \int_c^d f(x)\,dx$$





Funzione di ripartizione di v.c.

Definizione. Data una variabile casuale X, la funzione

$$F: \Re \rightarrow [0,1],\ F(x) = \mathbf{P}_X \{ X \leq x \} = \mathbf{P}\{ X^{-1} (-\infty, x] \}$$

è detta funzione di ripartizione di X.

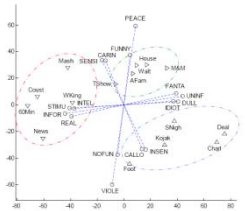
La funzione di ripartizione gode della seguenti proprietà:

- monotona non decrescente;
- continua a destra;
- $F(-\infty) = 0$, $F(+\infty) = 1$.

Per una variabile casuale X, la probabilità dell'intervallo [a,b] è data da:

$$\mathbf{P}_X \{[a,b]\} = F(b) - F(a)$$

Tale proprietà è alla base della rilevanza operativa della funzione di ripartizione.

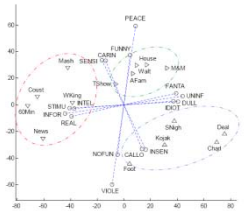


Valore atteso di variabili casuali

Definizione

Data una variabile casuale X, il suo valore atteso $E(X)$ è dato da:

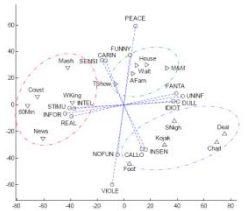
- $E(X) = \Sigma_x \, x \cdot P(X=x)$ se la v.c. è *discreta;*
- $E(X) = \int_{-\infty}^{+\infty} x f(x) dx$ se la v.c. è *continua.*



Valore atteso di variabili casuali

Esempio

Scenario x (guadagno 1000 €)	Prob (x)	$X \cdot \text{Prob}(x)$
-2	0.4	- 0.8
+6	0.4	+ 2.4
+12	0.2	+ 2.4
	$E(x)= \Sigma\, x\, P(x)$	**+ 4.0**

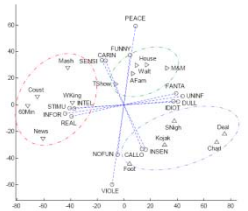


Varianza di variabili casuali

Definizione

È detta *varianza*, e si indica con σ^2, il valore atteso del quadrato dello scostamento della variabile casuale dalla sua media μ, $\sigma^2 = E(X-\mu)^2$:

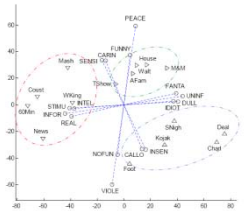
- $\sigma^2 = \Sigma_x (x-\mu)^2 P(X=x)$ se la v.c. è *discreta;*
- $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx$ se la v.c. è *continua.*



Varianza di variabili casuali

Esempio

Scenario x (guadagno 1000 €)	Prob (x)	$(x - \mu)^2 \text{Prob}(x)$
-2	0.4	14.4
+6	0.4	1.6
+12	0.2	12.8
	$\sigma^2 =$ $\sigma =$	**28.8** **5.367**



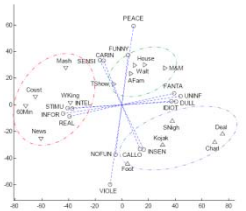
Variabili casuali notevoli

Discrete

- **Binomiale**: numero di successi in *n* prove
- **Poisson**: numero di eventi nel tempo
- **Binomiale negativa**: numero di tentativi per il primo successo

Continue

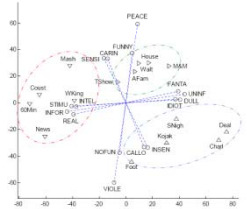
- **Gaussiana**: distribuzione utile per molti fenomeni
- **Lognormale**: come sopra, in finanza per i livelli di prezzo
- **Uniforme**: utilizzata per descrivere la massima incertezza
- **Gamma**: per fenomeni a valor positivi
- **Beta**: per fenomeni con supporto un intervallo (Es. un tasso percentuale aleatorio)
- **Esponenziale**: per descrivere il tempo tra due eventi (nascite, morti, arrivi, ..)



Statistica descrittiva e inferenza

STATISTICA: strumento conoscitivo atto ad analizzare in termini quantitativi un fenomeno collettivo.

- **Statistica descrittiva:** Insieme di metodologie atte a descrivere, riassumere le caratteristiche della distribuzione di una o più variabili statistiche. L'obiettivo viene perseguito rilevando le variabili di interesse su di un collettivo (popolazione).
- **Statistica inferenziale:** Ha gli stessi obiettivi della statistica descrittiva ma tenta di perseguirli rilevando solo una "parte" della popolazione. La "parte" (campione) viene scelta casualmente. Questo implica (vantaggio) l'uso dello strumento matematico del calcolo delle probabilità.



Le metodologie dell'inferenza

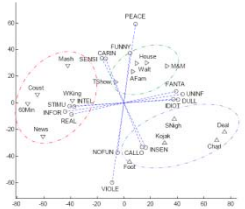
Stima e stimatori

Sia $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ un campione di n osservazioni con funzione di massa o densità di probabilità $f(\boldsymbol{x};\theta)$ dipendente da uno o più parametri, non noti, oggetto dell'inferenza.

Il problema della stima consiste nel costruire una funzione $T(x_1, x_2, \ldots, x_n)$ che permette di stimare il parametro non noto θ.

Va ricordato che lo stimatore, in quanto funzione di n variabili casuali, è esso stesso una variabile casuale, e andrebbe indicata in modo più appropriato con $T(X_1, X_2, \ldots, X_n)$, mentre rappresentiamo con t_n il risultato (detto appunto stima) su un campione particolare.

La deviazione standard dello stimatore è detta **standard error**.



Le metodologie dell'inferenza

Proprietà auspicabili per uno stimatore

- ***Correttezza* o *non distorsione***

$$E(T_n) = \theta \Leftrightarrow D(T_n) = E(T_n) - \theta = 0$$

- ***Efficienza* (errore quadratico medio minimo)**

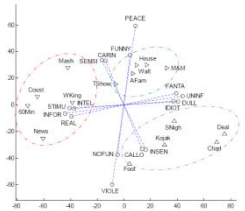
$$EQM(T_n) = E(T_n - \theta)^2 = V(T_n) + [D(T_n)]^2$$

- ***Consistenza debole***

$$\text{Lim}_n \, \mathbf{P}\{|T_n - \theta| > \delta\} = 0 \text{ per ogni } \delta > 0.$$

- ***Consistenza in media quadratica* (*implica quella debole*)**

$$\text{Lim}_n \, EQM(T_n) = 0.$$



Le metodologie dell'inferenza

Strategie per la costruzione di stimatori

- ***Analogia***

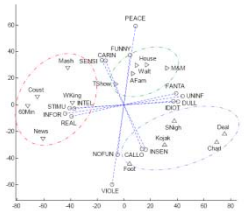
Le stime sono calcolate sul campione trattandolo in modo "analogo" alla popolazione.

- ***Metodo dei momenti***

Le stime sono calcolate uguagliando i momenti della popolazione (media, varianza, etc.) con quelli del campione.

- ***Massima verosimiglianza***

Le stime sono calcolate massimizzando la probabilità di ottenere il campione selezionato. Generalmente il metodo fornisce stimatori asintoticamente normali ed efficienti.



Le metodologie dell'inferenza

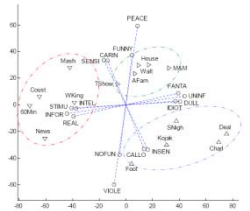
La verifica delle ipotesi

Si deve capire sulla scorta del campione $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$ se una certa affermazione sul parametro θ sia ragionevolmente vera o meno. In generale si ha:

- $\boldsymbol{X}$ un campione di n osservazioni con funzione di massa o densità di probabilità $f(\boldsymbol{x};\theta)$
- Un'affermazione sul parametro θ, detta ipotesi *nulla*, indicata con H_0;
- Una ipotesi *alternativa*, indicata con H_1 (H_0 falsa).

Esistono due importanti e distinte tipologie di errore:

- **Errore del *I tipo***: Rifiutare H_0, quando è vera. La sua probabilità si indica con α**;**
- **Errore del *II tipo***: Rifiutare H_1, quando è vera. La sua probabilità si indica con β.



Test Z

- Ipotesi nulla H_0: $\vartheta=\vartheta_0$

  Quando non altrimenti specificato, assumeremo che i parametri possano assumere qualsiasi valore reale. L'ipotesi alternativa è quindi sempre semplicemente la negazione dell'ipotesi nulla.

- Statistica test

$$z_{oss} = \left(\hat{\vartheta} - \vartheta_0\right) / (\text{Std. Err.}\hat{\vartheta}) \sim N(0,1) \text{ se } H_0 \text{ vera}$$

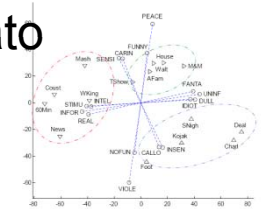
- Regola di rifiuto (probabilità di errore del primo tipo pari ad α)

$$| z_{oss} | > z_{\alpha/2}$$

- P-value

$$2\Pr\{Z > | z_{oss}|\}$$

  E' la probabilità di osservare un valore "più estremo" di quello osservato calcolato assumendo H_0 vera. Varia tra 0 e 1 e può essere interpretato come un indice di coerenza tra dati e ipotesi nulla.



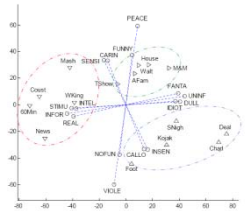
Le metodologie dell'inferenza

Intervalli di confidenza

Lo stimatore del parametro ϑ, in quanto variabile casuale, produce stime con un margine di errore. Al fine di fornire un idea circa questo margine, è utile ricercare per il parametro ϑ un **intervallo**. Tale intervallo è funzione dei dati campionari e quindi aleatorio. Detto I_ϑ tale intervallo, si chiede dunque che sia verificata la proprietà $P\{\vartheta \in I_\vartheta\} \geq c$. La costante c è detta livello di confidenza.

Di solito i livelli richiesti per c sono 95%, 99%, e 99.9%. La costruzione dell'intervallo di confidenza, generalmente parte da uno **stimatore** opportuno T_n.

Un intervallo di confidenza si può definire **ottimale** se, a parità di confidenza, risulta ragionevolmente "**piccolo"**



Intervalli di confidenza Z

Se esiste la variabile aleatoria

$$Z = \frac{\hat{\vartheta} - \vartheta}{se(\hat{\vartheta})} \sim N(0,1)$$

possiamo scegliere $z_{\alpha/2}$ in modo tale che

$$\Pr\{ Z > z_{\alpha/2} \} = \alpha/2 \Rightarrow \Pr\{ -z_{\alpha/2} < Z < z_{\alpha/2} \} = 1-\alpha$$

$$\Rightarrow \Pr\{ -z_{\alpha/2} < \frac{\hat{\vartheta} - \vartheta}{se(\hat{\vartheta})} < z_{\alpha/2} \} = 1-\alpha$$

$$\Rightarrow \Pr\{ \hat{\vartheta} - z_{\alpha/2} se(\hat{\vartheta}) < \vartheta < \hat{\vartheta} + z_{\alpha/2} se(\hat{\vartheta}) \} = 1-\alpha$$

In questo modo costruiamo un intervallo aleatorio

$$I_{\vartheta} = [\hat{\vartheta} - z_{\alpha/2} se(\hat{\vartheta}), \hat{\vartheta} + z_{\alpha/2} se(\hat{\vartheta})]$$

che con probabilità $1-\alpha$ “copre” il vero valore del parametro