



Identificazione, sperimentazione e validazione di alcuni indicatori di qualità delle attività sanitarie e socio-sanitarie del territorio

Roma 14, 15 settembre 2006

ASSR

AGENZIA per i
SERVIZI SANITARI
REGIONALI

UNIVERSITÀ DEGLI STUDI DI ROMA
U **TorVergata**

Il record linkage probabilistico – teoria

Idee basilari di Newcombe
La teoria di Fellegi-Sunter

Le idee del record linkage moderno sono nate dal genetista Howard Newcombe, che ha introdotto il rapporto di frequenza e le regole di decisione per delineare match e non match.

Linkage probabilistico: teoria - *Idee basilari di Newcombe*

Da Newcombe, "Handbook of Record Linkage"

L'idea di base è semplice

Se un nome o un'iniziale, o un mese di nascita, o un qualunque altro identificatore concordano o discordano, oppure sono più o meno simili tra loro, è naturale chiedersi quanto tipico sia l'esito di un confronto tra le coppie APPAIATE confrontato con coppie NON APPAIATE messe insieme a caso

Linkage probabilistico: teoria - *Idee basilari di Newcombe*

Il principio basilare del linkage probabilistico può essere rappresentato da una semplice formula di rapporto di frequenze (FREQUENCY RATIO).

Il rapporto tra la frequenza di un dato esito nel confronto tra coppie che costituiscono un match e la frequenza dello stesso tra le coppie che non costituiscono un match.

Linkage probabilistico: teoria - *Idee basilari di Newcombe*

FREQUENCY RATIO=

$$= \frac{\text{frequency of outcome } (x, y) \text{ among LINKED pairs}}{\text{frequency of outcome } (x, y) \text{ among UNLINKABLE pairs}}$$

dove

x indica l'identificatore e il suo valore nel primo file

y indica l'identificatore e il suo valore nel record nel secondo file

LINKED pairs si può riferire a tutte le coppie legate

UNLINKABLE pairs si può riferite a tutte le coppie che non si possono legare

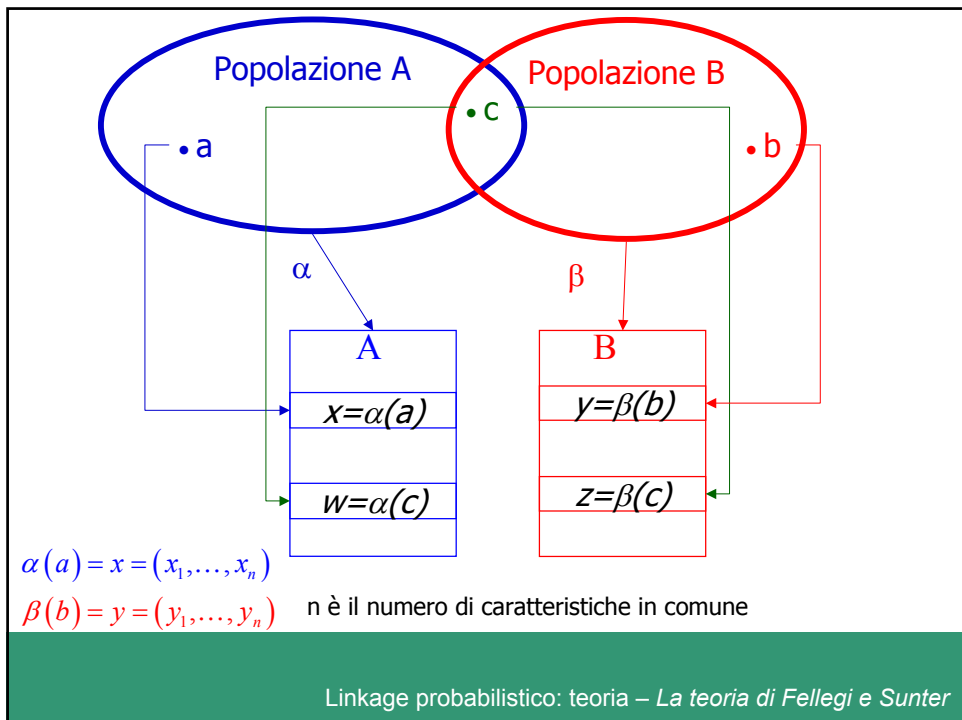
Ogni rapporto di frequenza (frequency ratio) in un certo senso rappresenta l'evidenza in favore di un vero match

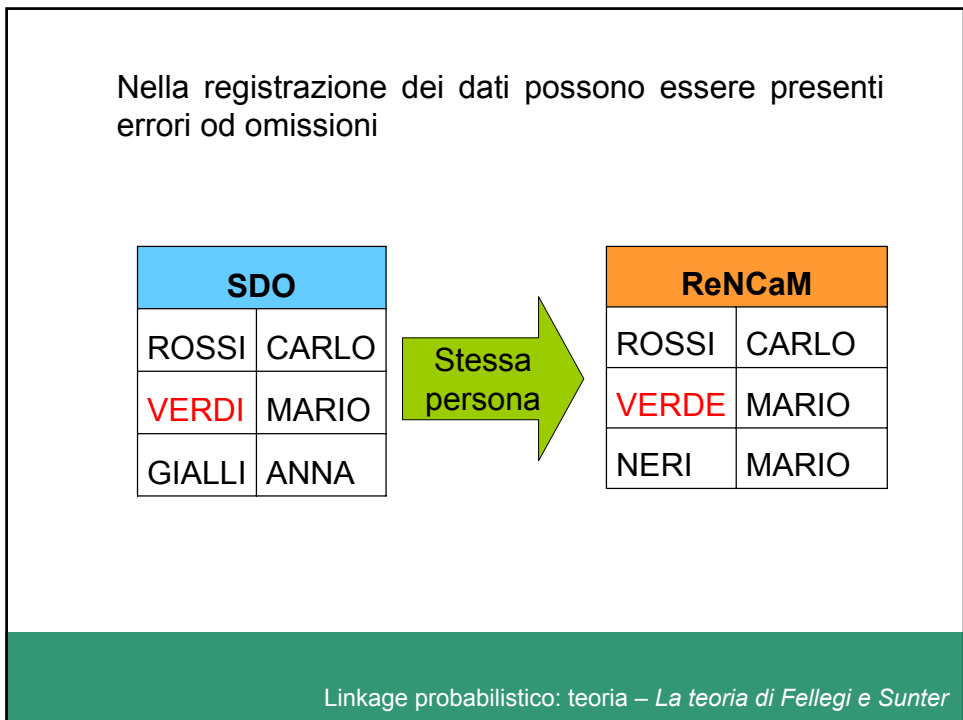
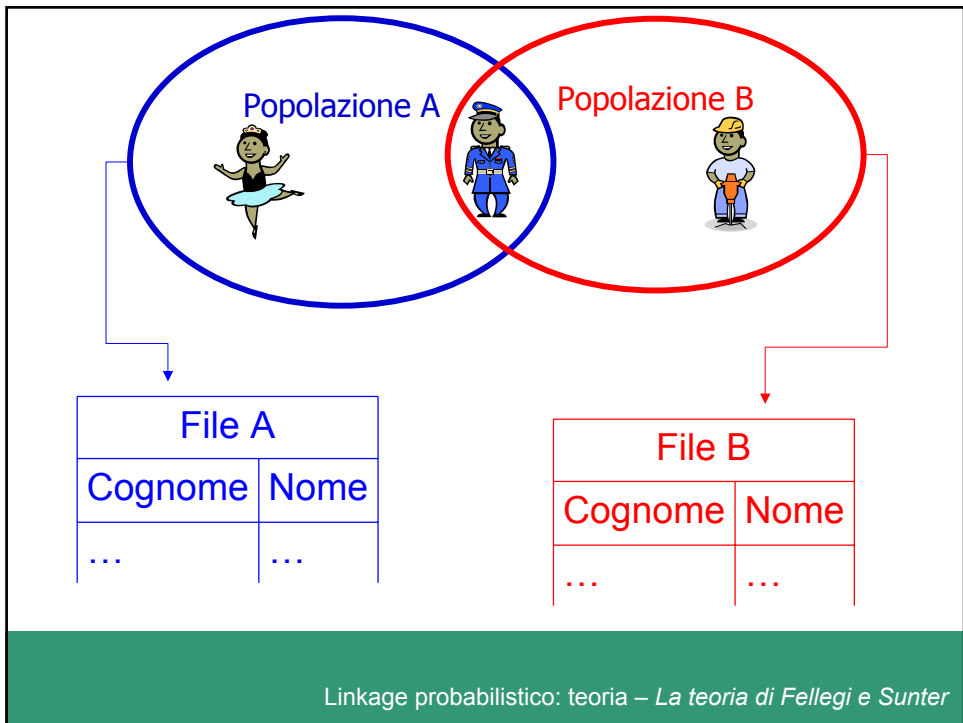
Linkage probabilistico: teoria - *Idee basilari di Newcombe*

Nel 1969 Fellegi e Sunter hanno introdotto i fondamenti matematici formali del record linkage e una struttura più generale di quella espressa da Newcombe.

La loro teoria dimostra la bontà delle regole decisionali utilizzate da Newcombe e introducono alcuni modi per stimare le probabilità cruciali direttamente dai file da legare.

Linkage probabilistico: teoria - *Idee basilari di Newcombe*





Per decidere se due record si riferiscono allo stesso soggetto sarà necessario confrontarne le caratteristiche comuni.

Si considera il prodotto cartesiano $A \times B$ cioè tutte le possibili coppie di record ...

e le funzioni di confronto delle
n caratteristiche dei soggetti



$$\Gamma_i : A \times B \rightarrow \{0, 1\}$$

$$(x, y) \mapsto \gamma_i$$

Ad esempio

$$\gamma_i = \begin{cases} 1 & \text{se c'è concordanza di valori} \\ 0 & \text{se c'è discordanza di valori} \end{cases}$$

Complessivamente il confronto
sarà



$$\Gamma : A \times B \rightarrow \{0, 1\}^n$$

$$(x, y) \mapsto \gamma = (\gamma_1, \dots, \gamma_n)$$

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

Consideriamo tutte le possibili coppie di record ...

SDO		ReNCaM	
ROSSI	CARLO	ROSSI	CARLO
ROSSI	CARLO	VERDE	MARIO
ROSSI	CARLO	NERI	MARIO
VERDI	MARIO	ROSSI	CARLO
VERDI	MARIO	VERDE	MARIO
VERDI	MARIO	NERI	MARIO
GIALLI	ANNA	ROSSI	CARLO
GIALLI	ANNA	VERDE	MARIO
GIALLI	ANNA	NERI	MARIO



Considerazioni sul numero di coppie

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

Esempio di funzione di confronto dicotomica ...

$$\gamma_i = \begin{cases} 1 & \text{se c'è concordanza di valori} \\ 0 & \text{se c'è disconcordanza di valori} \end{cases}$$

SDO		ReNCaM		COG	NOM
ROSSI	CARLO	ROSSI	CARLO	1	1
ROSSI	CARLO	VERDE	MARIO	0	0
ROSSI	CARLO	NERI	MARIO	0	0
VERDI	MARIO	ROSSI	CARLO	0	0
VERDI	MARIO	VERDE	MARIO	0	1
VERDI	MARIO	NERI	MARIO	0	1
GIALLI	ANNA	ROSSI	CARLO	0	0
GIALLI	ANNA	VERDE	MARIO	0	0
GIALLI	ANNA	NERI	MARIO	0	0

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

Possiamo quindi considerare le probabilità ...

$$m(\gamma_i) = \Pr[\gamma_i | M]$$

$$u(\gamma_i) = \Pr[\gamma_i | U]$$

Considerazioni sui
metodi di stima

... e le possiamo stimare

Inoltre possiamo costruire dei pesi associati ad ogni variabile ...

$$w_i = \ln \left(\frac{m(\gamma_i)}{u(\gamma_i)} \right)$$

e uno complessivo ...

$$w = w_1 + w_2 + \dots + w_n$$

Hp indep.

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

$$m(\gamma_i) = \Pr[\gamma_i | M]$$

E' legata **all'attendibilità** di una variabile: probabilità di commettere errori nella registrazione del dato

Tanto più è elevata l'attendibilità di una variabile tanto maggiore è il valore di m attribuito alla concordanza [$m(1)$] e quindi tanto minore di $m(0)$

$$u(\gamma_i) = \Pr[\gamma_i | U]$$

E' legata alla **capacità discriminante** di una variabile: probabilità di assumere valori differenti per soggetti differenti

Tanto più è elevata la capacità discriminante di una variabile tanto maggiore è il valore di u attribuito alla discordanza [$u(0)$] e quindi tanto minore il valore di $u(1)$

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

$$w_i = \ln \left(\frac{m(\gamma_i)}{u(\gamma_i)} \right)$$

Perciò il peso cresce con l'attendibilità e la capacità discriminante della variabile oggetto del confronto

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

Ipotizzando che per le variabili cognome e nome le stime di $m(\cdot)$ e $u(\cdot)$ siano ...

	m		u	
	1	0	1	0
COG	0,99	0,01	0,003	0,997
NOM	0,98	0,02	0,013	0,987

si ottengono i seguenti pesi ...

	1	0
w_cog	5,80	-4,60
w_nom	4,32	-3,90

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

E quindi ...

SDO		ReNCaM		COG	NOM	w_cog	w_nom	w
ROSSI	CARLO	ROSSI	CARLO	1	1	5,80	4,32	10,12
ROSSI	CARLO	VERDE	MARIO	0	0	-4,60	-3,90	-8,50
ROSSI	CARLO	NERI	MARIO	0	0	-4,60	-3,90	-8,50
VERDI	MARIO	ROSSI	CARLO	0	0	-4,60	-3,90	-8,50
VERDI	MARIO	VERDE	MARIO	0	1	-4,60	4,32	-0,28
VERDI	MARIO	NERI	MARIO	0	1	-4,60	4,32	-0,28
GIALLI	ANNA	ROSSI	CARLO	0	0	-4,60	-3,90	-8,50
GIALLI	ANNA	VERDE	MARIO	0	0	-4,60	-3,90	-8,50
GIALLI	ANNA	NERI	MARIO	0	0	-4,60	-3,90	-8,50

	1	0
w_cog	5,80	-4,60
w_nom	4,32	-3,90

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

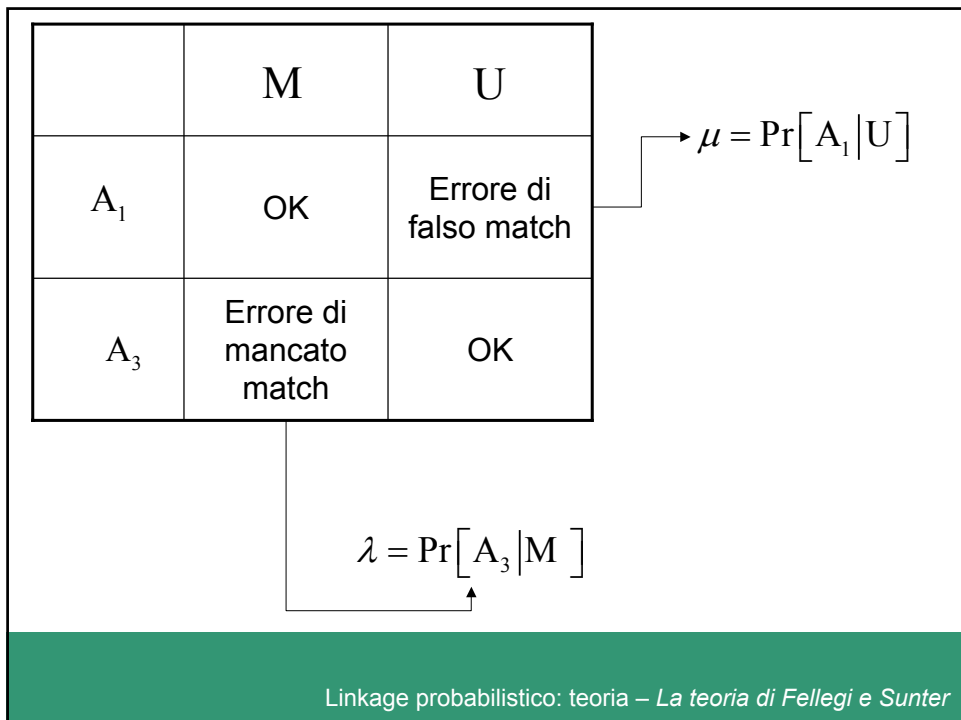
Per una coppia (x,y) di record possiamo fare una delle seguenti affermazioni:

- i record si riferiscono allo stesso soggetto nella popolazione
- non si può dire nulla
- i record non si riferiscono allo stesso soggetto

E si può definire una regola di matching suddividendo le coppie in tre insiemi:



Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*



Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

$$\lambda = \Pr[A_3 | M]$$

“Mancato match”

$$\mu = \Pr[A_1 | U]$$

“Falso match”

Secondo la teoria introdotta da Fellegi e Sunter fissati i livelli di errore λ e μ è possibile determinare una regola che minimizza la probabilità di A_2 rispetto a tutte le altre possibili regole di livello λ e μ . (regola ottimale)

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*

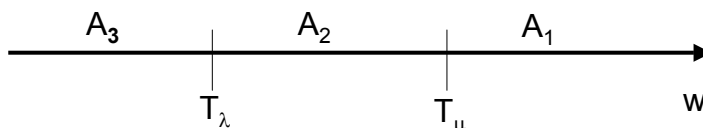
La regola

fissati i valori λ e μ si determinano due soglie T_λ e T_μ e si stabilisce che

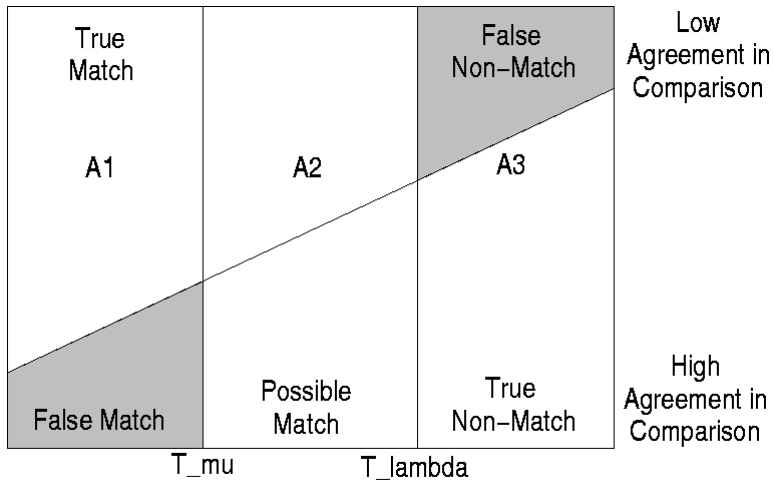
tutte le coppie con un peso superiore alla soglia maggiore saranno considerati dei match

tutte le coppie con un peso inferiore alla soglia minore saranno considerati dei non match

tutte le coppie con un peso tra le due soglie andranno rivisti manualmente



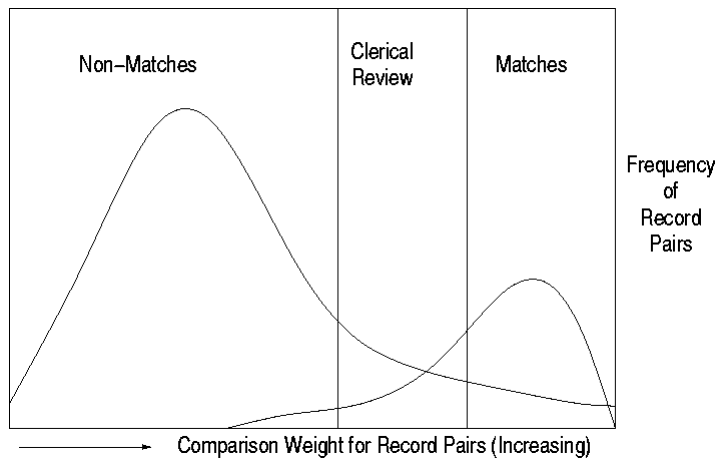
Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*



Record Pairs Ordered Monotonically by Comparison Weight

da Gu *et al.*, "Record linkage: current practice and future directions", 2003

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*



da Gu *et al.*, "Record linkage: current practice and future directions", 2003

Linkage probabilistico: teoria – *La teoria di Fellegi e Sunter*