

Progetto Formativo Aziendale
ORGANIZZAZIONE E UTILIZZO DI DATASET DI DATI ANAMNESTICI
EDESITI DIAGNOSTICI: RACCOLTA ED ARCHIVIAZIONE DEI DATI
Roma, 7-8 giugno 2016

Alimentare e rifinire un dataset: inserimento, controllo, verifica ricodifica dei dati

Marcello Sala
Istituto Zooprofilattico Sperimentale delle Regioni Lazio e Toscana
Osservatorio Epidemiologico Veterinario Regione Lazio
Roma



"Alimentare"

Una volta definiti

- obiettivi
- variabili (campi) di interesse
- struttura del dataset
- formato del dataset

Si passa alla raccolta dati ed al loro inserimento



“Alimentare”

Significa “dare cibo”!

Quindi alimentare un dataset significa riempire la tabella precedentemente “pensata” e strutturata inserendo i dati raccolti

Come per una dieta corretta...l'alimentazione implica un comportamento corretto

..quindi non si mangia in piedi..non si spilucca..ma non si mangia neppure troppo
... ci si siede a tavola!

“Alimentare”

Implementare un dataset non è un'abbuffata di dati disordinata ma un comportamento che ha le sue regole

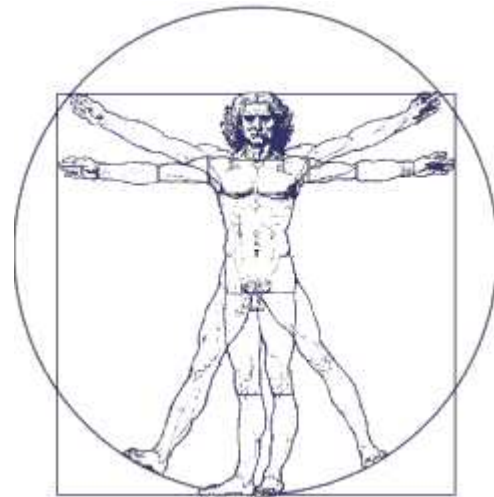


“Alimentare”

Lo scopo è ottenere un dataset con le informazioni minime necessarie, robusto, completo, pertinente ed appropriato per un pronto utilizzo



Panzaset



Dataset



“Alimentare”

Senza regole ed organizzazione il dataset che otteniamo è QUESTO...



Sarà necessario lavorare molto in seguito per renderlo utilizzabile!



"Alimentare"

Significa quindi:

- 1) Inserire progressivamente (o più raramente in modo massivo) nuovi record nel dataset
- 2) inserire in modo completo, appropriato e pertinente i dati relativi ai campi del dataset

Implica quindi l'**inserimento** dei dati e l'**aggiornamento** del dataset



Come “Alimentare” il dataset 1

- Fonte scheda di rilevamento (es. Questionario)
- Fonte Archivio informatico (es. Estrazione SIL)

Ricordate “dal foglio di carta alla creazione di un dataset”



Come “Alimentare” il dataset 2

- Alimentazione diretta della tabella (da scheda o da file)
- Alimentazione mediante maschera (da scheda)



Come “Alimentare” il dataset 2

- L’argomento è stato trattato da Massimo Mari


Il concetto guida è rappresentato dalla
necessità di garantire la “stabilità del dato”

In alimentazione del dataset GARANTIRE "Stabilità" del dato

Il dato deve essere riportato sempre in modo omogeneo all'interno dei campi

Se il dato non è stabile, l'informazione ne risulta inficiata

Controlli - verifiche



Progressivo campione	Data Prelievo	Azienda	matricola	Indirizzo produttivo	sexso	ANNO NASCITA	esito	lt/die	n° parti
1	02/09/2016	001XX001	00000001	latte	F	2012	pos	28	1
2	24/03/2015	002XX003	00000002	carne	F	2010	neg	24	2
3	02/09/2016	001XX001	pippo	latte	Femmina	2006	negativo	20	3
4	2015/05/22	003XX002	00000004	carne	Femmina	2016	neg	26	3
5	02/09/2016	001XX001	00000005	latte	M	2012	pos	32	1
6		003XX002	00000006	carne	F	2010	0,32	18	2
7	24/03/2015	002XX003	00000007	carne	F	2014	pos	38	1
8	?	003XX002	00000008	carne	F	2013	neg	51	1
9	24/03/2015	002XX003	00000009	latte	F	2012	neg	?	2
10		003XX002	ignota	carne	F	1992	neg	19	1

Dopo essere stati attenti in inserimento
Bisogna comunque VERIFICARE il dataset prima dell'utilizzo

Controlli - verifiche

Perché?

- Spesso ci si accorge dell'errore dopo aver fatto l'elaborazione e quindi si deve correggere e ricominciare
- Se è passato tempo dall'inserimento o non si è tenuta la scheda o il file di origine, l'errore è irrecuperabile (Valore vero?)
- Spesso l'errore può passare inosservato e determinare distorsioni dei risultati

Controlli - verifiche

Perché?

- I controlli non servono solo a individuare errori ma anche a risolvere problemi di omogeneità dei dati sfuggiti in inserimento

Controlli – verifiche di base

1. Omogeneità
2. Date e rapporti tra date (calcolo delle età ecc)
3. Completezza campi
4. Pertinenza
5. Espressione esiti
6. Plausibilità
7. Congruenza tra campi
8.

Controlli – verifiche di base

1. Omogeneità dei campi

Verificare che i dati inseriti in un campo

- Stessa unità di misura
- Sesso formato
- Stessa estensione

Controlli – verifiche di base

1. Omogeneità - Stessa unità di misura

Agendo sul filtro del foglio excel e
scorrendo....o operando una verifica
a vista di possono notare anomalie
grossolane

**Risultato espresso in
parti per bilione**

Esito reale = 5,91 mg/Kg

id campione	Matrice	esito (mg/Kg)
1	latte di massa	< 0.0005
2	latte di massa	< 0.0005
3	latte di massa	0.0008
4	latte di massa	0.0006
5	latte di massa	< 0.0005
6	latte di massa	0.0007
7	latte di massa	0.015
8	latte di massa	0.019
9	latte di massa	0.001
10	latte di massa	< 0.0005
11	latte di massa	< 0.0005
12	latte di massa	< 0.0005
13	latte di massa	0.0007
14	latte di massa	0.0016
15	latte di massa	< 0.0005
16	latte di massa	0.0033
17	latte di massa	0.005
18	latte di massa	60
19	latte di massa	0.001
20	latte di massa	0.0046
21	latte di massa	0.035
22	latte di massa	0.01
23	latte di massa	0.0007
24	latte di massa	0.0005
25	latte di massa	0.508
26	latte di massa	< 0.0005
27	latte di massa	0.009
28	latte di massa	0.015
29	latte di massa	< 0.0005
30	latte di massa	0.007
31	latte di massa	0.002
32	latte di massa	< 0.0005
33	latte di massa	< 0.0005
34	latte di massa	0.015

Controlli – verifiche di base

1. Omogeneità - Stesso formato

Agendo sul filtro del foglio excel o verifica a vista

*Succede spesso nel
copia & incolla
di righe da file diversi*

Progressivo campione	matricola	esito
1	00000001	0,006
2	00000002	0,003
3	pippo	<0,005
4	00000004	neg
5	00000005	pos
6	00000006	5.91
7	00000007	pos
8	00000008	neg
9	00000009	neg
10	ignota	neg

**Numerico
decimale (,)**

**Numerico
decimale (.)**

Testo

Controlli – verifiche di base

1. Omogeneità - Stesso formato

Agendo sul filtro del foglio excel o verifica a vista

*Succede spesso nel
copia & incolla
di righe da file diversi*

Progressivo campione	Data Prelievo
1	02/09/2016
2	24/03/2015
3	02/09/2016
4	2015/05/22
5	02/09/2016
6	
7	24/03/2015
8	
9	24/03/2015
10	

Data Europeo

**Data
Americano**

Controlli – verifiche di base

1. Omogeneità – Stessa estensione

Agendo sul filtro del foglio excel o verifica a vista

Progressivo campione	matricola	sex	esito
1	00000001	F	pos
2	00000002	F	neg
3	pippo	Femmina	negativo
4	00000004	Femmina	neg
5	00000005	M	pos
6	00000006	F	positivo
7	00000007	F	pos
8	00000008	F	neg
9	00000009	F	neg
10	ignota	F	neg

Estensione diversa per "sex"

Estensione diversa per "esito"

Controlli – verifiche di base

2. Date e rapporti tra date

Verificare

- Date abbiano stesso formato
- Consequenzialità delle date

Controlli – verifiche di base

2. Date e rapporti tra date – Data nascita precedente a data prelievo
 - Fondamentale verifica in funzione del calcolo dell'età dell'animale al prelievo

Simile importanza: data trattamento, data di produzione dei lotti, delle partite, delle scadenze dei prodotti

Controlli – verifiche di base

2. Date e rapporti tra date – Data nascita precedente a data prelievo
- Verifica effettuabile a "vista" per dataset piccoli
 - Si può impostare una semplice finzione in excel per dataset grossi

Data nascita (data trattamento, produzione...)	Data prelievo	Delta giorni	
01/05/2014	08/09/2015	487	=GIORNO360(A61;B61;VERO)
04/10/2015	08/09/2015	-26	
....	

Valore negativo indica un problema di consequenzialità date

Controlli – verifiche di base

3. Completezza campi

Verificare l'esistenza di "vuoti" nei dati inseriti

Agendo sul filtro del foglio excel o a vista si possono notare
i Buchi

Un dato mancante potrebbe determinare l'esclusione
dell'intero record dall'analisi

Controlli – verifiche di base

3. Completezza dei campi

Progressivo campione	Data Prelievo	Azienda	matricola	Indirizzo produttivo	sexso	ANNO NASCITA	esito	lt/die	n° parti
1	02/09/2016	001XX001	00000001	latte	F	2012	0,006	28	1
2	24/03/2015	002XX003	00000002	carne	F	2010	0,003	24	2
3	02/09/2016	001XX001	00000003	latte	F	2006	<0,005	20	
4	X	003XX002	00000004	carne	F	2016	neg	26	3
5	02/09/2016	001XX001	00000005	latte	M	2012	pos	32	1
6	X	003XX002	X	X	F	X	5.91	18	2
7	24/03/2015	002XX003	00000007	carne	F	2014	pos	38	1
8	X	003XX002	00000008	carne	F	2013	X	51	1
9	24/03/2015	002XX003	X	latte	F	2012	neg	X	2
10	X	003XX002	00000010	carne	F	1882	neg	19	1

Controlli – verifiche di base

3. Completezza campi

.. bisogna definire se il dato non è disponibile (missing) o non è stato inserito seppur disponibile...quindi recuperarlo

Progressivo campione	Data Prelievo	Azienda	matricola	Indirizzo produttivo	sexso	ANNO NASCITA	esito	lt/die	n° parti
1	02/09/2016	001XX001	00000001	latte	F	2012	0,006	28	1
2	24/03/2015	002XX003	00000002	carne	F	2010	0,003	24	2
3	02/09/2016	001XX001	00000003	latte	F	2006	<0,005	20	
4	24/03/2015	003XX002	00000004	carne	F	2016	neg	26	3
5	02/09/2016	001XX001	00000005	latte	M	2012	pos	32	1
6	02/09/2016	003XX002	MISSING	latte	F	MISSING	5.91	18	2
7	24/03/2015	002XX003	00000007	carne	F	2014	pos	38	1
8	02/09/2016	003XX002	00000008	carne	F	2013	0,003	51	1
9	24/03/2015	002XX003	00000009	latte	F	2012	neg	MISSING	2
10	24/03/2015	003XX002	00000010	carne	F	1882	neg	19	1

Il record non potrà essere usato per analizzare i dati in funzione della variabile

Controlli – verifiche di base

4. Pertinenza

Il dato contenuto in un campo DEVE rappresentare la variabile a cui si riferisce

Questa caratteristica viene “gestita” nella fase di creazione del dataset ed in inserimento ma....

Spesso nei copia e incolla si sfasa l’allineamento dei record nei confronti dei campi e si inficia la pertinenza

Questa verifica consente di individuare gli “sfasamenti”

Controlli – verifiche di base

4. Pertinenza

Dataset

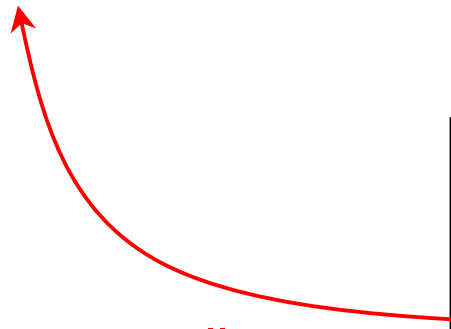
Progressivo campione	Data Prelievo	Azienda	matricola	Indirizzo produttivo	sexso	ANNO NASCITA	esito	lt/die	n°parti
1	02/09/2016	001XX001	00000001	latte	F	2012	0,006	28	1
2	24/03/2015	002XX003	00000002	carne	F	2010	0,003	24	2
3	02/09/2016	001XX001	00000003	latte	F	2006	<0,005	20	
4	24/03/2015	003XX002	00000004	carne	F	2016	neg	26	3
5	02/09/2016	001XX001	00000005	latte	M	2012	pos	32	1
02/09/2016	003XX002	MISSING	latte	F	MISSING	5.91	18	2	
24/03/2015	002XX003	00000007	carne	F	2014	pos	38	1	
02/09/2016	003XX002	00000008	carne	F	2013	0,003	51	1	
24/03/2015	002XX003	00000009	latte	F	2012	neg	MISSING	2	
24/03/2015	003XX002	00000010	carne	F	1882	neg	19	1	



Tabella esterna

Progressivo campione	Data Prelievo	Azienda	matricola	Indirizzo produttivo	sexso	ANNO NASCITA	esito	lt/die	n°parti
6	02/09/2016	003XX002	MISSING	latte	F	MISSING	5.91	18	2
7	24/03/2015	002XX003	00000007	carne	F	2014	pos	38	1
8	02/09/2016	003XX002	00000008	carne	F	2013	0,003	51	1
9	24/03/2015	002XX003	00000009	latte	F	2012	neg	MISSING	2
10	24/03/2015	003XX002	00000010	carne	F	1882	neg	19	1

Copia&incolla



Controlli – verifiche di base

5. Espressione degli esiti

E' un controllo specifico fondamentale che racchiude tutte le caratteristiche delle precedenti verifiche:

1. Omogeneità
3. Completezza
4. Pertinenza

La sua importanza è cruciale per l'utilizzo finale del dataset

Perché questo campo è spesso oggetto di **codifica** e **ri-codifica**

Controlli – verifiche di base

6. Plausibilità

E' un controllo che esula dalla verifica puramente formale
(1.-5.)

Consiste nell'assicurarsi che il dato inserito per uno o più
record in corrispondenza di un determinato campo
(variabile) **abbia senso!**

**Sono valori formalmente corretti ma insensati o non
plausibili**

Controlli – verifiche di base

6. Plausibilità

Progressivo campione	Data Prelievo	Azienda	matricola	Indirizzo produttivo	sexso	ANNO NASCITA	esito	lt/die	n° parti
1	02/09/2016	001XX001	00000001	latte	F	2012	pos	28	1
2	24/03/2015	002XX003	00000002	carne	F	2010	neg	24	2
3	02/09/2016	001XX001	00000003	latte	F	2006	neg	20	3
4	24/03/2015	003XX002	00000004	carne	F	2016	neg	26	3
5	02/09/2016	001XX001	00000005	latte	M	2012	pos	32	1
6	02/09/2016	003XX002	00000006	carne	F	2010	pos	18	2
7	24/03/2015	002XX003	00000007	carne	F	2014	pos	38	1
8	02/09/2016	003XX002	00000008	carne	F	2013	neg	51	1
9	24/03/2015	002XX003	00000009	latte	F	2012	neg	180	2
10	24/03/2015	003XX002	00000010	carne	F	1882	neg	19	1

Errori inserimento?

Dove sta l'errore?

Controlli – verifiche di base

7. Congruenza tra campi

E' un controllo che esula dalla verifica puramente formale
(1.-5.)

Consiste nell'assicurarsi che il dato inserito in un campo sia
congruente con il dato inserito in un altro campo
collegato secondo logica o gerarchia

Sono valori formalmente corretti ma in contraddizione

Controlli – verifiche di base

7. Congruenza tra campi (esempi)

(I controlli tra date, maschi con 2 parti.....)

Dataset Scrapie - congruenza tra motivo del prelievo e luogo del prelievo:

- Luogo mattatoio (3)=motivo reg. macellati (1) o focolaio (3)
- Luogo allevamento (1) = motivo morto allevamento (2)
- Tutte le altre combinazioni sono incongruenti

7. Congruenza tra campi (esempi)

ASL	data_prelievo	luogoprel	comune_prel	provincia_prel	cod_istat	specie	identificativo	codazALL	Motivo_prelievo
L107	30/12/2014	1	TERME	SI	090	O	IT052000806345	009SI030	2
L102	30/12/2014	1	CAPANNORI	LU	090	O	IT046000000265	007LU006	2
O106	02/01/2015	1	MANZIANA	RM	120	O	851850	054RM021	2
O109	02/01/2015	1	ENTE	VT	120	O	056000439757	001VT021	2
O107	02/01/2015	1	ROMANO	RM	120	O	IT058000333154	113RM004	2
L106	03/07/2014	1	SANTA LUCE	PI	090	O	050000072914	034PI033	2
O107	02/01/2015	1	MANDELA	RM	120	O	38005400016416	053RM013	2
O107	02/01/2015	1	MANDELA	RM	120	O	38005400016409	053RM013	2
O107	02/01/2015	1	GERANO	RM	120	O	IT058000272592	044RM010	2
L107	02/01/2015	1	O	SI	090	O	IT052000021083	014SI248	2
L102	02/01/2015	1	ANTELMINEL	LU	090	C	IT046000027837	011LU399	2
O107	04/01/2015	1	NUOVA	RM	120	O	IT058000331085	014RM021	2
O107	05/01/2015	1	GERANO	RM	120	O	IT058000272419	044RM010	2
O107	05/01/2015	1	GERANO	RM	120	O	IT058000272494	044RM010	2
O107	03/01/2015	1	MANDELA	RM	120	O	IT054000121562	053RM013	2
O107	03/01/2015	1	ROMANO	RM	120	O	IT058000348082	008RM014	2
O107	03/01/2015	1	ROMANO	RM	120	O	IT058000088647	008RM012	2
O107	03/01/2015	1	ROMANO	RM	120	O	IT058000348042	008RM012	2
L109	02/01/2015	1	GROSSETO	GR	090	O	IT053000110527	011GR499	2
L109	02/01/2015	1	GROSSETO	GR	090	O	IT053000143820	011GR499	2
L109	02/01/2015	1	GROSSETO	GR	090	O	IT013GR12101M7	011GR499	2
L109	02/01/2015	1	GROSSETO	GR	090	O	IT011GR499-308	011GR499	2
O112	05/01/2015	3	PALIANO	FR	120	O	IT060000180284	006FR688	3
O112	05/01/2015	3	PALIANO	FR	120	O	IT060000180224	006FR688	3
O112	05/01/2015	3	PALIANO	FR	120	O	IT060000180351	006FR688	3
O112	05/01/2015	3	PALIANO	FR	120	O	IT060000180295	006FR688	3
O112	05/01/2015	3	PALIANO	FR	120	O	IT060000180283	006FR688	3
O112	05/01/2015	3	PALIANO	FR	120	O	IT060000180217	006FR688	3
O112	05/01/2015	3	PALIANO	FR	120	O	IT060000180570	006FR688	3
O112	05/01/2015	3	PALIANO	FR	120	O	IT060000180486	006FR688	3

Controlli – verifiche di base

7. Congruenza tra campi (esempi)

Con una pivot individuo l'entità delle incongruenze

luogoprel	Motivo_prelievo			Totale
	reg. mac (1)	Morto (2)	Focolaio (3)	
Allevamento (1)	24	1929	12	1965
Mattatoio (3)	1544	221	206	1971
Totale complessivo	1568	2150	218	3936

Incongruenti:

Verificare e rettificare

Controlli – verifiche di base

NOTA

La maggior parte di queste verifiche può essere effettuata in una prima fase esplorativa dell'elaborazione dati (ne parleremo domani) predisponendo delle semplici tabelle di frequenza

REGOLA D'ORO

Maggiore sarà la precisione nell'inserimento dati iniziale minore risulterà l'entità delle verifiche (e delle anomalie rilevate) nel dataset finale

Le codifiche dei dati in inserimento aiutano a limitare le anomalie

Codifica – RI-codifica

Definizione generale

rappresenta una “SEMPLIFICAZIONE” del dato originario inserito in un campo, all’interno del dataset, attraverso l’adozione di “etichette” o di valori convenzionali in sostituzione del dato grezzo

I dati trasformati possono assumere valori nominali, ordinali, discreti, dicotomici

Codifica-ricodifica dei dati

La codifica è decisa in fase di inserimento dati

La ricodifica avviene invece a posteriori, una volta concluse le verifiche (e le eventuali correzioni) del dataset

In questo corso adotteremo semplificazioni e non entreremo troppo nel dettaglio

Codifica dei dati

Scopo

1. Assicurare omogeneità e pertinenza ai dati inseriti in un campo
2. Ridurre il livello di riempimento del dataset (occupare meno memoria)
3. Attribuire classi gerarchiche di valori ai dati
4. Predisporre la tabella secondo il formato richiesto dai software statistici

Codifica

Si attribuisce ad un valore grezzo di un campo un codice alfabetico (a 1 o più lettere) o numerico

Il formato delle codifiche è (quasi) sempre "Testo"

I codici attribuiti possono essere incrementali
(sottendono una gerarchia di valori crescente)
oppure sono indipendenti

Codifiche più frequenti

Codifica in variabile DICOTOMICA (esprime presenza assenza di due valori mutualmente esclusivi)

Esito qualitativo (positivo-negativo) codificato in "1" e "0"

...oppure

Sesso Maschio e femmina codificati come "1" e "0" (o "M" – "F")

Variabili di esposizione codificate in termini di presenza/assenza "1" e "0" (o "SI" – "NO")

(es. vaccinazione, trattamento si/no)

Codifiche più frequenti

Codifica in variabile NOMINALE (esprime più alternative di valore)

Esito qualitativo (negativo-conforme-non conforme)
vengono codificati in "0" - "1" - "2"

...oppure

Sesso Maschio e femmina e castrone codificati come "1" -
"0" - "2" (o "M" - "F" - "C")

Variabili di esposizione (es indirizzo produttivo -motivi di
prelievo) "1" - "2" - "3" ("A" - "B" - "C") ecc...

Es. trattamento1, trattamento 2, placebo)

Codifiche più frequenti

Dati mancanti o non disponibili

Devono sempre essere codificati in formato testo con un valore: "**missing**" oppure "**99**"

Ricordate un campo **vuoto** in una tabella o foglio di calcolo costituisce un valore (ad es 0 in excel)

Codifiche più frequenti

Codifica in variabile incrementale (esprime diversi livelli gerarchici del valore) – qualitativa ordinale

Più livelli di risultato o di esposizione in altrettanti codici incrementali – i dati hanno un ordine “naturale”

Es.

- 1) alto, medio, basso = “2”, “1”, “0”
- 2) Mai, qualche volta, sempre = “0”, “1”, “2”
- 3) Ecc..

Progressivo campione	matricola	Indirizzo produttivo	sexso	ANNO NASCITA	esito	lt/die	n°parti	TRATTAMENTI	vaccinazione
1	00000001	latte	F	2012	pos	28	1	MAI	vaccino 1
2	00000002	carne	F	2010	neg	24	2	QUALCHE VOLTA	nessuna
3	00000003	latte	F	2006	neg	20	3	MAI	nessuna
4	00000004	carne	F	2016	neg	26	3	QUALCHE VOLTA	nessuna
5	00000005	latte	M	2012	pos	32	1	QUALCHE VOLTA	vaccino 1
6	00000006	carne	F	2010	pos	18	2	SEMPRE	vaccino 1
7	00000007	carne	F	2014	pos	38	1	SEMPRE	nessuna
8	00000008	carne	MISSING	2013	neg	51	1	MAI	nessuna
9	00000009	latte	F	2012	neg	180	2	QUALCHE VOLTA	vaccino 1
10	00000010	misto	F	1992	neg	19	1	SEMPRE	nessuna

Progressivo campione	matricola	Indirizzo produttivo	sexso	ANNO NASCITA	esito	lt/die	n°parti	TRATTAMENTI	vaccinazione
1	00000001	A	0	2012	1	28	1	0	1
2	00000002	B	0	2010	0	24	2	1	0
3	00000003	A	0	2006	0	20	3	0	0
4	00000004	B	0	2016	0	26	3	1	0
5	00000005	A	1	2012	1	32	1	1	1
6	00000006	B	0	2010	1	18	2	2	1
7	00000007	B	0	2014	1	38	1	2	0
8	00000008	B	99	2013	0	51	1	0	0
9	00000009	A	0	2012	0	180	2	1	1
10	00000010	C	0	1992	0	19	1	2	0

latte=A; F=0;
carne=B; M=1;
misto=C NON DISPONIBILE=99

POS=1;
NEG=0

MAI=0;
QLCH=1;
SEMPRE=2

vaccino 1=1;
nessuna=0

Codifiche

Tenere sempre un file o un foglio dove sono annotate tutte le codifiche o ri-codifiche eseguite

ESITO	
Valore	Codifica
POS	1
NEG	0

SESSO	
Valore	Codifica
F	0
M	1
Non disp.	99

IND. PROD	
Valore	Codifica
latte	A
carne	B
misto	C

Trattamenti	
Valore	Codifica
MAI	0
QLCH	1
SEMPRE	2

Vaccino	
Valore	Codifica
Vaccino 1	1
nessuna	0

Codifica-Ricodifica

Una forma particolare di ri-codifica, che normalmente viene condotta a posteriori sul dataset, riguarda la creazione di nuove variabili di raggruppamento partendo dalle variabili grezze o codificate nel dataset

In sostanza si parla di **riclassificazione** dei valori del campo (variabile) in un nuovo campo

E' un processo di aggregazione in classi dei valori dei campi

Ri-classificazione

Il processo prevede l'ulteriore raggruppamento dei valori di un campo in un numero minore di categorie (classi)

Ossia, i valori tra loro vicini vengono raggruppati nella stessa classe creando un nuovo campo con un numero minore di valori della variabile

I criteri di aggregazione dei valori devono essere definiti chiaramente

Ri-classificazione

..forse meglio ragionare su un esempio classico...

Le classi d'età....

Creazione nuovo campo "Classetà"

Valori di riferimento – campo "età anni"

Ri-classificazione

Prima della riclassificazione
6 possibili valori

Progressivo campione	matricola	Indirizzo produttivo	sexso	età anni	Classetà
1	00000001	A	0	3	1
2	00000002	B	0	5	2
3	00000003	A	0	9	3
4	00000004	B	0	1	1
5	00000005	A	1	3	1
6	00000006	B	0	5	2
7	00000007	B	0	1	1
8	00000008	B	99	2	1
9	00000009	A	0	3	1
10	00000010	C	0	17	3

Criteri:

Classe 1 = 1-3 anni = codifica "1"

Classe 2 = 4-6 anni = codifica "2"

Classe 3 = >6 anni = codifica "3"

numero

testo

Dopo la riclassificazione
3 possibili valori

Ri-classificazione

Altri esempi

Litri latte individuale prodotto

Creazione nuovo campo "Classe produzione"

Valori di riferimento – campo "lt/die"

Ri-classificazione

Prima della riclassificazione
11 possibili valori

Progressivo campione	matricola	Indirizzo produttivo	sezzo	lt/die	Classe_produzione
1	00000001	A	0	28	2
2	00000002	B	0	24	2
3	00000003	A	0	20	1
4	00000004	B	0	26	2
5	00000005	A	1	32	3
6	00000006	B	0	18	1
7	00000007	B	0	38	3
8	00000008	B	99	51	3
9	00000009	A	0	35	3
10	00000010	C	0	19	1

Criteri:

Classe 1 = ≤ 20 lt = codifica "1"

Classe 2 = 21-30 lt = codifica "2"

Classe 3 = >30 lt = codifica "3"

numero

Dopo la riclassificazione
3 possibili valori

testo

E così via.....

Ri-classificazione

La riclassificazione viene normalmente utilizzata per ridurre gli strati di una variabile che altrimenti sarebbe difficile analizzare con i valori originali

Va applicata quando opportuno

Semplifica le analisi

Ma ... i criteri di raggruppamento devono avere un senso o una utilità informativa

Ri-classificazione

La riclassificazione riguarda quasi sempre le
variabili di “esposizione” (fattori di rischio)
più raramente le variabili di esito

Vedremo le sue applicazioni domani....

Solo dopo aver alimentato appropriatamente un dataset, eseguito i controlli si potrà lavorare serenamente alle

Ri-codifica

Ri-classificazione