QUADERNI DEL DIPARTIMENTO DI SCIENZE DELL'UOMO

DSU 02/2007

Giovanni Delli Zotti (*dellizottig@sp.units.it*)

ARE YOU EXPERIENCED …?
Trasformazione di variabili e creazione di indici con SPSS

Marzo 2007

Università degli Studi di Trieste
www.dsu.units.it

Quaderni del Dipartimento di Scienze dell'Uomo
Università degli Studi di Trieste
DSU: 02/2007
(www.sp.units.it/working papers)

ARE YOU EXPERIENCED …?
Trasformazione di variabili e creazione di indici con SPSS

di Giovanni Delli Zotti

INDICE

1. Introduzione

Probabilmente molti ritengono che per un utente del programma di elaborazione di dati statistici SPSS® la cosa più complicata sia addentrarsi nella complessa articolazione delle procedure statistiche. Queste ultime sono infatti molto numerose, offrono varianti che utilizzano algoritmi di calcolo alternativi e opzioni che consentono di produrre una mole sterminata di tabulati. Le finestre di dialogo di Spss sono però utilizzabili in modo intuitivo e perciò non è difficile ottenere dal programma quello che si vuole, posto che si sappia qual è il modello di analisi statistica utile per rispondere ai nostri problemi conoscitivi. Se lo si conosce davvero, si conoscono anche le diverse opzioni e varianti e si è probabilmente in grado di interpretare i tabulati che il programma produce con una certa facilità.

Più complesso è invece "ripulire" i dati, trasformare, se necessario, le variabili e aggiungere indici o tipologie costruiti a partire dalle variabili originali. Senza pretesa di essere esaustivi, cercheremo di mostrare in questo lavoro quante e quali cose si possano e a volte si debbano fare prima di iniziare ad analizzare compiutamente la matrice dei dati. Può forse sorprendere che ciò accada anche con dati che provengono da un questionario completamente strutturato, come quello qui utilizzato per le esemplificazioni.

Dobbiamo al lettore una precisazione riguardo al titolo un po' stravagante di questo lavoro (fortunatamente il sottotitolo precisa di cosa davvero si parla in queste pagine): *Are you experienced…?* vuole essere un omaggio a Jimi Hendirx, un grande innovatore che non si è fatto circoscrivere all'interno delle sicurezze dell'esistente, ma ha voluto battere strade nuove che lo rendono ancora attualissimo a trent'anni dalla sua scomparsa. L'omaggio è un po' pretenzioso per un quaderno con finalità didattiche, e infatti l'idea iniziale era quella di attribuire questo titolo ad una pagina Internet collegata al sito della Memoria storica della sezione Metodologia dell'Ais (Associazione Italiana di Sociologia)[1]. La pagina Internet non è stata ancora realizzata, comunque nelle intenzioni di chi scrive potrebbe costituire un forum di discussione e deposito di esperienze per utenti più o meno esperti di Spss che non si accontentano, si chiedono se sono veramente *experienced* e hanno consapevolezza che c'è sempre qualcosa da imparare dagli errori, ma ovviamente anche dalle intuizioni proprie e degli altri.

Infine, il lettore non si sorprenda per l'uso intensivo dell'autocitazione: questo Quaderno costituisce l'ultimo, per ora, episodio di una serie di lavori nati con intendimento didattico e per i quali si può trovare un filo conduttore nel proposito di illustrare e discutere le alternative disponibili in alcuni passaggi cruciali della costruzione degli strumenti di rilevazione ed analisi dei dati.

2. Uso della sintassi

In Spss sono presenti vari comandi e funzioni che servono principalmente a:

– *definire i dati*: fornire informazioni sul tipo di variabili, fornire etichette per i nomi sintetici delle variabili e per i valori, segnalare la presenza di codici che indicano la presenza di dati mancanti;
– *trasformare i dati*: per creare nuove variabili, modificare quelle esistenti, selezionare un sottocampione di casi per le analisi successive, pesare i casi, ecc.;
– *attivare procedure*: per realizzare analisi statistiche o rappresentazioni grafiche, salvare i dati, esportare l'output, ecc.

Tutti o quasi questi comandi e funzioni possono essere attivati mediante il sistema a finestre, selezionando le opzioni adatte e/o scrivendo le informazioni richieste all'interno delle ca-

[1] La Memoria storica è accessibile dal sito: www.uniud.it/dest/docenti/dellizotti/dellizot.htm, dove sono scaricabili i testi on line citati in questa sede.

selle di testo fornite dal programma. Ci dobbiamo però chiedere se sia questo il modo più efficiente di operare: chi ha esperienza nell'uso del programma sa che è imprescindibile conoscere la sintassi dei comandi di Spss, anche se ciò non significa rinunciare all'ausilio del sistema di finestre, pulsanti, caselle di selezione, ecc.

Non pretendiamo che si comperi "a scatola chiusa" questa nostra affermazione, e perciò forniremo alcune motivazioni, nella certezza che l'uso del programma farà scoprire al lettore ulteriori motivi di soddisfazione per avere un po' faticato ad apprendere le semplici regole della sintassi di Spss. Tra i motivi che consigliano la pratica sistematica delle registrazione delle trasformazioni su un foglio di sintassi, si possono perlomeno indicare i seguenti:

- *Velocizzazione delle operazioni*: per quanto possa sembrare strano al lettore meno smaliziato, l'uso intensivo e spesso inappropriato del mouse comporta (anche) un certo degrado dell'efficienza. Ad esempio, abbandonare la tastiera mentre si scrive un testo, afferrare il mouse e selezionare la piccola icona che nella barra degli strumenti indica il corsivo, comporta un dispendio di tempo maggiore rispetto all'uso della combinazione di tasti **`CTRL+I`**. Allo stesso modo, come è facile constatare con un po' di pratica, in Spss è più semplice scrivere una lista di etichette di variabili o di valori adattando il questionario che possediamo già in formato digitale, piuttosto che inserirle una ad una utilizzando finestre e finestrelle e spesso, dovendo notificare uno per uno i cambiamenti, si rischia di perdere il lavoro svolto.
- *Documentazione interna*: le trasformazioni di variabili implicano decisioni che facilmente si possono dimenticare e spesso non sono desumibili dal lavoro finito. Ad esempio, se per costruire un indice inseriamo una formula nell'apposita finestra e non la trascriviamo su un foglio di sintassi, ben presto rischiamo di non ricordarci più quali siano le variabili utilizzate ed i pesi eventualmente loro attribuiti. Analogamente, possiamo ridurre i valori di una variabile cardinale alle categorie "basso/medio/alto", ma in seguito non saremo probabilmente in grado di ricordare quali siano i valori utilizzati per frazionare la distribuzione (specialmente se si tratta di punteggi di qualche "astratto" indice additivo)[2].
- *Rendicontazione esterna*: documentare le trasformazioni dei dati per poter ricostruire le convenzioni e decisioni adottate o le procedure implementate è anche un'esigenza "esterna" di trasparenza e rendicontazione che ci può essere richiesta per verificare e replicare il nostro lavoro.
- *Ripetizione delle analisi*: molto spesso lo stesso ricercatore vuole applicare le "ricette" sperimentate a nuovi file di dati, oppure ripetere le analisi e/o le trasformazioni in quanto pervengono nuovi casi da aggiungere in matrice. L'esperienza insegna infatti che è molto utile memorizzare anche i comandi usati per realizzare le analisi dei dati, specialmente se abbiamo usato una numerosa serie di procedure statistiche diverse, alcune precedute da una selezione di casi e/o specificando alcune opzioni. Terminato il lavoro, ci si potrebbe rendere conto che è opportuno escludere alcuni questionari compilati da persone troppo giovani o troppo anziane, oppure aggiungere questionari arrivati in ritardo o, infine, ripetere le stesse analisi separatamente per gli intervistati di ognuna delle quattro province del Friuli-Venezia Giulia. Se abbiamo registrato su un file di sintassi la sequenza di procedure della prima analisi, replicarla su un sottoinsieme di casi o sulla matrice integrata dei nuovi casi diventa questione di pochi secondi.
- *Opportunità in più*: non tutte le opzioni di Spss sono state implementate nel sistema a finestre di dialogo. Ad esempio, se vogliamo calcolare la correlazione tra una variabile dipendente ed una serie di altre nove variabili indipendenti, scopriremo che esiste una sola finestra di dialogo all'interno della quale collocare le variabili. Il programma produrrà dunque l'intera matrice di correlazione di dimensione 10x10, anche se a noi servivano solo nove coefficienti. Invece con la sintassi è possibile specificare quello che esattamente vogliamo inserendo la parola chiave

[2] Per un'ampia discussione dei criteri che si possono adottare nel caso in cui si debba o voglia suddividere in classi o categorie una variabile cardinale, si veda *Come "fare a fette" una distribuzione di frequenza* (Delli Zotti 2005). Recentemente Spss ha implementato un modulo per automatizzare le procedure di segmentazione delle variabili. Non a caso le strategie adottabili sono sostanzialmente le stesse quattro che sono state esposte nel Quaderno: si può infatti automaticamente segmentare la variabile mediante intervalli uguali di valori oppure uguale consistenza numerica dei gruppi creati, ed inoltre, indicare manualmente i punti di taglio individuando graficamente le fratture presenti nella distribuzione, oppure adottare valori-soglia significativi.

WITH tra la variabile dipendente e le indipendenti (**CORRELATION VARIABLES = indip WITH dip1 TO dip9**). Inoltre, i gruppi di variabili creati per tabulare le risposte multiple non vengono salvate nella matrice dei dati e perciò vanno reimpostati ogni volta che si accede al programma. Si può ovviare se si possiede il modulo opzionale **TABLES** che consente di ritrovare i gruppi creati al suo interno alla riapertura del programma, oppure semplicemente ricordandosi di incollare le istruzioni in un foglio di sintassi dal quale si possono immediatamente mandare in esecuzione quando necessario.

3. Operatori e parole chiave

In Spss i nomi delle variabili devono iniziare con una lettera e non possono avere lunghezza superiore a 64 byte. Non è possibile utilizzare spazi, caratteri speciali (ad esempio !, ?, ’e *) e le parole chiave standard di Spss che identificano operatori logici, relazionali e alcune parole riservate.

Con questa prescrizione siamo arrivati ad un nodo problematico nell’uso della sintassi di Spss: la necessità di tenere presenti alcune regole e, più di tutto, di ricordare parole chiave che peraltro non sono numerose e hanno nomi sufficientemente evocativi della loro funzione.

3.1 Operatori relazionali e logici

All’interno dei comandi di Spss possiamo usare i seguenti operatori relazionali che, come si vede dalla lista, possono essere sostituiti dai simboli che troviamo sulla calcolatrice.

EQ, =	uguale a
NE, <>	non uguale a
LT, <	inferiore a
LE, <=	inferiore o uguale a
GT, >	maggiore di
GE, >=	maggiore a uguale a

Gli operatori relazionali si possono anche combinare mediante i seguenti operatori logici.

AND, &	Entrambe le relazioni devono essere vere
OR, |	Almeno una relazione deve essere vera
NOT	Inverte il risultato dell’espressione

Per esemplificare l’uso degli operatori logici e relazionali prendiamo in considerazione la seguente distribuzione di frequenza della variabile `d3matur` (diploma).

		Frequenza	Percentuale
Validi	1 Scientifica	23	17,4
	2 Scienze sociali	17	12,9
	3 Altro liceo	34	25,8
	4 Tec. commerciale	25	18,9
	5 Altro tecnico	21	15,9
	6 Professionale	12	9,1
	Totale	132	100,0

Se vogliamo selezionare solo gli intervistati in possesso del diploma del liceo di scienze sociali possiamo usare l’istruzione che nell’esempio seguente è collocata a sinistra; se, al contrario, li vogliamo escludere, dobbiamo usare quella a destra. Per evitare che i casi non sele-

zionati vengano cancellati dal file e rimangano perciò disponibili per successive elaborazioni, il comando **SELECT IF** va preceduto da **TEMPORARY.**

```
SELECT IF d3matur EQ 2.
```

		Frequenza	Percentuale
Validi	2 Scienze sociali	17	100,0

```
SELECT IF d3matur NE 2.
```

		Frequenza	Percentuale
Validi	1 Scientifica	23	20,0
	3 Altro liceo	34	29,6
	4 Tec. commerciale	25	21,7
	5 Altro tecnico	21	18,3
	6 Professionale	12	10,4
	Totale	115	100,0

Nei due esempi successivi abbiamo selezionato i soli licei sulla destra e invece li abbiamo esclusi sulla sinistra. E' possibile farlo utilizzando un solo operatore relazionale quando i valori da selezionare sono contigui e si collocano tra un estremo ed un valore intermedio nella distribuzione. Va ovviamente utilizzata una sola delle due istruzioni perfettamente equivalenti.

```
SELECT IF d3matur LT 4.
SELECT IF d3matur LE 3.
```

		Frequenza	Percentuale
Validi	1 Scientifica	23	31,1
	2 Scienze sociali	17	23,0
	3 Altro liceo	34	45,9
	Totale	74	100,0

```
SELECT IF d3matur GT 3.
SELECT IF d3matur GE 4.
```

		Frequenza	Percentuale
Validi	4 Tec. commerciale	25	43,1
	5 Altro tecnico	21	36,2
	6 Professionale	12	20,7
	Totale	58	100,0

Come detto, gli operatori relazionali possono essere combinati, come nell'esempio seguente con il quale sono stati selezionate le sole studentesse femmine (codice 2 della variabile d2) in possesso del diploma di scienze sociali.

```
SELECT IF d3matur EQ 2 AND d2 EQ 2.
```

		Frequenza	Percentuale
Validi	2 Scienze sociali	13	100,0

L'uso degli operatori logici è sufficientemente intuitivo, ma è facile commettere errori. Se scriviamo la seguente istruzione con il proposito di selezionare i diplomati in scienze sociale e quelli in possesso del diploma tecnico commerciale, ci aspettiamo di ottenere 42 studenti; ma il programma stamperà un avviso se chiediamo la distribuzione di frequenza.

```
SELECT IF d3matur EQ 2 AND d3matur EQ 4.
```

Avvisi

Per questa procedura non sono stati specificati casi. È possibile che nel file dati di lavoro non siano presenti casi o che tutti i casi siano stati esclusi dal filtro attivo.

Questo comando non viene eseguito.

Anche se può sembrare contro intuitivo, l'istruzione corretta è la seguente, che utilizza l'operatore logico **OR** invece di **AND**.

```
SELECT IF d3matur EQ 2 OR d3matur EQ 4.
```

		Frequenza	Percentuale
Validi	2 Scienze sociali	17	40,5
	4 Tec. commerciale	25	59,5
	Totale	42	100,0

Nel caso dell'istruzione **SELECT IF**, se il risultato dell'espressione logica è vero, il caso verrà selezionato e se il risultato di un'espressione logica è falso o mancante, il caso non verrà selezionato. Usando il connettore logico **AND** entrambe le condizioni devono essere soddisfatte e il risultato nel nostro esempio è stato nullo perché con quell'istruzione si pretendeva che fossero estratti studenti in possesso del diploma in scienze sociali e anche tecnico commerciali, condizione impossibile da verificarsi perché le categorie di una stessa variabile per definizione si escludono mutuamente. La richiesta di estrarre dal campione le femmine diplomate in scienze sociali è invece plausibile, perché i due requisiti di appartenenza si riferiscono a categorie di due distinte variabili. Ha anche funzionato l'istruzione con l'operatore **OR,** in quanto con esso intendiamo che debba essere soddisfatta la condizione di appartenenza alla categoria "scienze sociali" oppure a quella "tecnico commerciale".

Il problema deriva probabilmente dal fatto che ci dimentichiamo che con **SELECT IF** non ci riferiamo al risultato, ma alle condizioni che devono essere soddisfatte per ottenere un determinato risultato. Può dunque capitare di usare distrattamente **AND** perché vogliamo entrambi, non rendendoci conto che con **SELECT IF** si richiede invece che entrambe le condizioni siano soddisfatte e talvolta queste condizioni possono essere incompatibili. D'altro canto, può non essere spontaneo usare **OR** perché vogliamo entrambe le categorie e non l'una o l'altra, non rendendoci conto che **OR**, in un'istruzione di selezione, significa che è sufficiente che anche una sola delle due condizioni sia soddisfatta.

3.2 Parole riservate

Nei comandi di definizione dei dati, di trasformazione degli stessi e nella richiesta di attivazione di procedure, possono essere presenti parole riservate come le seguenti, oltre agli operatori aritmetici (**+**, **-**, **/**, ***** e ****** per l'elevazione a potenza) e a numerosissime funzioni statistiche e di altra natura, tra le quali nel seguito vedremo la sola funzione **MEANS**.

Le parole riservate principali sono le seguenti, delle quali scopriremo l'uso commentando gli esempi che seguiranno. Vedremo anche che alcune di esse possono essere usate in combinazione (ad esempio, **LO THRU 18** all'interno di un'istruzione **RECODE** significa dal valore più basso della distribuzione fino a 18).

ELSE	Tutti gli altri valori
INTO	Nella nuova variabile
THRU	Dal valore precedente a quello seguente
LOWEST o **LO**	Dal valore più basso
HIGHEST o **HI**	Dal valore più elevato
TO	Dalla variabile precedente alla variabile seguente
WITH	Con
BY	Per
SYSMIS	Mancante di sistema
ALL	Tutte le variabili

4. Comandi di trasformazione dei dati

I principali comandi per la trasformazione dei dati possono contenere espressioni ed esse possono includere nomi di variabili, costanti, operatori aritmetici, funzioni numeriche e di altro tipo, variabili logiche e operatori relazionali.

4.1 Il comando **COMPUTE**

La sintassi del comando **COMPUTE** (calcola) è la seguente:

```
COMPUTE variabile destinazione = espressione.
```

Si noti l'inversione rispetto alla notazione nelle normali espressioni aritmetiche nelle quali il risultato si scrive dopo l'espressione. Negli algoritmi informatici invece si deve indicare per primo il nome della variabile nella quale verrà riportato per ognuno dei casi il risultato del calcolo e solo dopo l'espressione da calcolare.

Ecco alcuni esempi di applicazione del comando:

COMPUTE newvar = 0. Crea la variabile `newar` e pone il suo valore uguale a 0.

COMPUTE totred=redlav+redpens. Crea la variabile `totred` sommando i redditi da pensione e da lavoro.

COMPUTE index=(d1+d2+d3+d4)/4. Crea la variabile `index` sommando le riposte fornite alle domande da `d1` a `d4` e divide il risultato per 4.

4.2 Il comando **IF**

Un altro comando che genera nuove variabili è **IF** (se):

```
IF [(] espressione logica [)] variabile destinazione = espressione.
```

Con **IF** il risultato dall'espressione collocata in fondo al comando viene attribuito ai casi che soddisfano l'espressione logica collocata all'inizio. Ad esempio:

IF (prof EQ 1) status=4. Se la professione è 1 assegna il valore 4 nella nuova variabile `status` con la quale viene attribuito un punteggio di status ad ognuno dei casi.

4.3 Il comando **RECODE**

Con **RECODE** (ricodifica) vengono modificati i valori di una variabile esistente. Se si conclude il comando con **INTO** e dal nome di una nuova variabile, il risultato della trasformazione viene attribuito alla nuova variabile e resta immutata quella originaria.

```
RECODE lista di variabili (lista di valori = valore)...(lista di valori =
  valore) [INTO lista di variabili].
```

Il comando è apparentemente complesso, ma intuitivo se guardiamo al seguente esempio:

```
RECODE età (18 THRU 35=1)(36 THRU 65=2)(66 THRU HI=3) INTO etarec.
```

Il valori della variabile `età` vengono raggruppati in classi (da 18 a 35, da 36 a 65, oltre 65 anni) e viene creata la nuova variabile `etarec` nella quale invece dell'età in anni verranno registrati i nuovi valori 1, 2 e 3, cui corrispondono le tre classi di età. Se avessimo omesso **INTO** `etarec,` avremmo effettuato una trasformazione permanente della variabile originaria e non sarebbe più disponibile l'età in anni degli intervistati. **INTO** é un'aggiunta opzionale, segnalata dalle parentesi quadre; infatti, a volte non è utile creare una nuova variabile, partico-

larmente se si tratta semplicemente di ricondurre ad un'unica categoria residuale alcune categorie originarie che si sono rivelate essere esiguamente rappresentate nel campione.

È possibile usare **RECODE** anche per trasformare una serie di variabili in una nuova serie di variabili modificate.

```
RECODE dom1 TO dom8 (1 2=0)(3 4=1) INTO rdom1 TO rdom8.
```

Con il comando precedente abbiamo trasformato i valori da 1 a 4 (molto contrario/abbastanza contrario/abbastanza d'accordo/molto d'accordo) di una serie di otto domande nei nuovi valori 0 e 1 (contrario/d'accordo) creando al contempo otto nuove variabili.

Come si sarà notato, oltre a **INTO**, negli esempi abbiamo usato alcune altre parole riservate. **TO** indica che vanno operate la trasformazioni per tutte le variabili comprese tra quella che lo precede e quella che lo segue; **THRU** che vanno trasformati tutti i valori compresi tra quello che lo precede e quello che lo segue. Si è anche fatto seguire **THRU** da **HI**, ad indicare che quello specifico intervallo si chiude con il valore più alto della distribuzione.

4.4 Il comando **COUNT**

Il comando **COUNT** (conta) attribuisce ad una nuova variabile il risultato del conteggio delle occorrenze di un valore o di una lista di valori all'interno di una lista di variabili.

```
COUNT nome di variabile = lista di variabili (lista di valori).
```

Ad esempio, con l'istruzione seguente si è chiesto di contare per ogni intervistato il numero di volte che ha risposto 3 (abbastanza) oppure 4 (molta) in una serie di domande con le quali si è chiesto di esprimere il grado di fiducia verso 10 istituzioni. In questo modo si è ottenuto un semplice indicatore di fiducia generalizzata verso le istituzioni.

```
COUNT fiducia = fid1 TO fid10 (3 4).
```

4.5 Le strutture **DO IF–END IF** *e* **DO REPEAT–END REPEAT**

Con la struttura **DO IF—END IF** si chiede al programma di effettuare le trasformazioni solo per i casi che soddisfano la condizione indicata nell'espressione logica. Opzionalmente si può utilizzare **ELSE IF**, per indicare esplicitamente come operare nel caso in cui l'espressione logica non sia vera.

```
DO IF [(] espressione logica [)].
Comandi di trasformazione.
[ELSE IF [(] espressione logica [)]].
Comandi di trasformazione.
END IF.
```

Nel seguente esempio di applicazione di questa struttura si chiede che vengano modificati i valori della variabile `var1` solo per i casi che riportano il valore 1 nella variabile `contr1`.

```
DO IF (contr1=1).
RECODE var1 (1 2=1)(3 4=2).
END IF.
```

La struttura può essere sostituita con il più semplice **IF** se per la trasformazione si deve usare **COMPUTE**. La struttura **DO IF-END IF** seguente, infatti, permette di giungere allo stesso risultato che nel paragrafo 5.3 si è ottenuto con **IF (d4s=1) d4=d4/60*100**.

```
DO IF (d4s=1).
```

```
COMPUTE d4=d4/60*100.
END IF.
```

Con la struttura **DO REPEAT–END REPEAT** si ripetono le stesse trasformazioni su un insieme di variabili e in questo modo si possono ridurre di molto i comandi da usare per ottenere un determinato risultato.

```
DO REPEAT variabile segnaposto = {lista di variabili}.
Comandi di trasformazione.
END REPEAT.
```

Al posto di una lista di variabili viene indicata una variabile "segnaposto" e le trasformazioni vengono ripetute per tutte le variabili citate nella lista che vanno a turno a sostituire la variabile segnaposto. Con il seguente esempio vengono generate le 5 nuove variabili denominate da `test1` a `test5` e ad esse è attribuito il valore iniziale 0.

```
DO REPEAT T = test1 TO test5.
COMPUTE T = 0.
END REPEAT.
```

5. Esemplificazioni e digressioni

La base di dati utilizzata per le esemplificazioni è costituita da 132 questionari somministrati a studenti della Facoltà di Scienze politiche nell'autunno del 2005. I risultati saranno pubblicati prossimamente, congiuntamente a quelli della rilevazione effettuata nel 2007 e per ora sono disponibili on line i risultati di una precedente rilevazione (Delli Zotti 2001).

Il questionario era strutturato e ciononostante si vedrà come praticamente per quasi tutte le domande sia necessario o utile effettuare modifiche e come le variabili possano essere combinate per produrre indici o tipologie.

5.1 Calcolo dell'età dall'anno di nascita e creazione di classi (**COMPUTE**, **RECODE**)

Nella matrice dei dati è stato registrato l'anno di nascita usando le sole ultime due cifre (es. 86, invece di 1986) e la distribuzione di frequenza è quella mostrata nella tabella seguente, in parte modificata, come le altre nel seguito, per migliorare la visualizzazione o impaginazione. In particolare, nelle distribuzioni di frequenza sono state escluse le colonne dedicate alle percentuali sui casi validi e cumulate.

d1 Anno di nascita

		Frequenza	Percentuale
Validi	50	1	,8
	53	1	,8
	54	1	,8
	55	1	,8
	56	1	,8
	58	1	,8
	61	2	1,5
	63	2	1,5
	64	1	,8
	67	2	1,5
	68	1	,8
	69	1	,8
	71	2	1,5
	72	1	,8
	74	1	,8
	75	1	,8
	76	1	,8
	78	1	,8
	79	5	3,8
	80	1	,8
	81	3	2,3
	82	5	3,8
	83	6	4,5
	84	12	9,1
	85	26	19,7
	86	52	39,4
	Totale	132	100,0

L'età è calcolata mediante l'istruzione **COMPUTE**, sottraendo il valore della variabile `d1` da 105 (l'anno della rilevazione è il 2005). Si è anche fornita di un'etichetta descrittiva la nuova variabile `d1eta`.

```
COMPUTE d1eta=105-d1.
VARIABLE LABELS d1eta Età.
```

d1eta Età

		Frequenza	Percentuale
Validi	19	52	39,4
	20	26	19,7
	21	12	9,1
	22	6	4,5
	23	5	3,8
	24	3	2,3
	25	1	,8
	26	5	3,8
	27	1	,8
	29	1	,8
	30	1	,8
	31	1	,8
	33	1	,8
	34	2	1,5
	36	1	,8
	37	1	,8
	38	2	1,5
	41	1	,8
	42	2	1,5
	44	2	1,5
	47	1	,8
	49	1	,8
	50	1	,8
	51	1	,8
	52	1	,8
	55	1	,8
	Totale	132	100,0

Mentre la variabile cardinale `d1eta` sarà utilizzata con strumenti statistici come la media o la correlazione, è opportuno costruire classi d'età per usare questa informazione nelle tabelle di contingenza. Per conservare anche la variabile cardinale originale, le trasformazioni vengono assegnate mediante la parola chiave **INTO** alla nuova variabile `d1etarec`.

```
RECODE d1eta (19=1)(20 21=2)(22 THRU HI=3) INTO d1etarec.
VARIABLE LABELS d1etarec Classe di età.
VALUE LABELS d1etarec 1 "19 anni" 2 "20-21" 3 "22 e più".
```

d1etarec Classe di età

		Frequenza	Percentuale
Validi	1 19 anni	52	39,4
	2 20-21	38	28,8
	3 22 e più	42	31,8
	Totale	132	100,0

Se per qualche motivo si preferisce che l'anno di nascita venga scritto con tutte le quattro cifre che lo compongono, con un'istruzione **COMPUTE** si può aggiungere 1900 al valore registrato in matrice, senza che venga creata una nuova variabile. Ovviamente, in questo caso per ottenere l'età l'anno di nascita dovrà essere detratto da 2005.

```
COMPUTE d1=1900+d1.
COMPUTE d1eta=2005-d1.
VARIABLE LABELS d1eta Età.
```

5.2 Aggregazione di una variabile categoriale (**RECODE**)

Se alcuni diplomi sono scarsamente presenti tra gli studenti intervistati si può decidere di aggregarli ad altri. Anche in questo caso si è mantuta in matrice la variabile originale e la nuova variabile è stata corredata di un'etichetta e di etichette per i nuovi valori.

```
RECODE d3 (1=1)(9=2)(2 3 4=3)(6=4)(5 7=5)(8=6) INTO d3matur.
```

```
VARIABLE LABELS d3matur "Tipo di diploma".
VALUE LABELS d3matur
1 “Scientifico” 2 “Scienze sociali” 3 “Altro liceo”
4 “Commerciale” 5 “Altro tecnico” 6 “Professionale”.
```

d3 Diploma di maturità

		Frequenza	Perc.
Validi	1 Liceo scientifico	23	17,4
	2 Liceo classico	13	9,8
	3 Liceo linguistico	17	12,9
	4 Altro liceo	4	3,0
	5 Tecnico industriale	9	6,8
	6 Commerciale	25	18,9
	7 Altro tecnico	12	9,1
	8 Professionale	12	9,1
	9 Scienze sociali	17	12,9
	Totale	132	100,0

d3matur Tipo di diploma

		Frequenza	Per.
Validi	1 Scientifico	23	17,4
	2 Scienze sociali	17	12,9
	3 Altro liceo	34	25,8
	4 Commerciale	25	18,9
	5 Altro tecnico	21	15,9
	6 Professionale	12	9,1
	Totale	132	100,0

5.3 Digressione su valori mancanti e mancanti di sistema

Prima di affrontare il problema della standardizzazione è opportuno notare che nella tabella posta all’inizio del prossimo paragrafo incontriamo per la prima volta l’indicazione “mancante di sistema”. Si tratta di un valore che il programma assegna automaticamente alle celle nelle quali non trova alcun codice e conviene ricodificarlo (con la parola riservata **SYSMIS** nell’istruzione **RECODE**), assegnargli un’etichetta (mediante **VALUE LABELS**) e ridefinirlo come mancante (mediante **MISSING VALUES**).

Nel decidere quale nuovo valore attribuire a **SYSMIS** (mancante di sistema), si consideri che il valore “0” non è adeguato in quanto spesso esso indica una possibile risposta valida (ad esempio, nessun figlio). Si può usare un valore come “99”, apparentemente estremo e perciò improbabile per una variabile di due cifre, ma nel nostro esempio esso indica una risposta valida (voto 99). Perciò chi scrive preferisce in genere utilizzare un valore negativo (-1) che, tra l’altro, per la presenza del segno “-” è più facilmente individuabile nella matrice dei dati.

In generale non è necessario e nemmeno utile assegnare etichette ai valori di una variabile cardinale, ma fa eccezione il valore mancante che deve essere contrassegnato con un’adeguata etichetta. Si deve anche segnalare al programma che il nuovo valore deve essere considerato un dato mancante, cosa che avveniva automaticamente per il “mancante di sistema”.

```
RECODE d4 (SYSMIS=-1).
VALUE LABELS d4 -1 Non risposto.
MISSING VALUES d4 (-1).
```

Può sembrare stravagante darsi tanta pena per arrivare apparentemente allo stesso punto da cui si è partiti, ma bisogna considerare attentamente una differenza la cui importanza può essere sottovalutata. Il nuovo valore mancante è infatti gestibile dall’utente, quindi, invece di essere escluso automaticamente, si può scegliere se usarlo o meno nelle elaborazioni. Ad esempio, potrà essere usato in una tabella di contingenza per appurare se sono più numerosi i maschi oppure le femmine che hanno deciso di non rivelare il loro voto di diploma. Se vogliamo invece calcolare il voto medio di maschi e femmine, possiamo (anzi, dobbiamo) ovviamente escludere dal calcolo questo valore. Può essere perciò utile convertire “globalmente” i valori mancanti di sistema presenti in tutte le variabili (utilizzando la parola riservata **ALL**), indicando poi che il nuovo codice (-1) è da considerare mancante.

```
RECODE ALL (SYSMIS=-1).
MISSING VALUES ALL (-1).
```

Così come sono stati definiti mancanti, questi valori possono essere successivamente "sdoganati" al fine del loro inserimento nelle tabelle di contingenza. Si può infatti predisporre un'istruzione che annulla quella precedente, semplicemente omettendo di indicare valori all'interno delle parentesi. Alternando le due istruzioni, i dati mancanti possono essere facilmente utilizzati oppure omessi dai calcoli statistici.

```
MISSING VALUES ALL ().
```

Va ricordato che **ALL**, in comandi come **RECODE** e **MISSING VALUES**, non funziona se nella matrice vi sono variabili stringa (cioè variabili che contengono caratteri alfabetici, come il comune di residenza scritto per esteso). Infatti, la ricodifica di valori di variabili stringa avviene in modo diverso (i valori vanno ad esempio scritti tra virgolette) e le celle vuote non vengono definite automaticamente come mancanti di sistema. Perciò, se la matrice contiene 36 variabili e `dom16` è una variabile stringa, si dovrà escluderla come nel comando seguente:

```
RECODE dom1 TO dom15 dom17 TO dom36 (SYSMIS=-1).
```

5.4 Standardizzazione di unità di misura diverse (**IF** *e* **RECODE**)

Il voto di maturità, come si vede dalla tabella seguente, per gli studenti più "anziani" è stato espresso in sessantesimi e si dovrà perciò provvedere alla conversione di questi voti nella nuova metrica in centesimi. Si tenga conto che, a complicare la situazione, il voto 60 è ambiguo, in quanto per alcuni studenti corrisponde al voto più basso (se espresso in centesimi) e per altri a quello più alto (se diplomato quando i voti erano espressi in sessantesimi).

d4 Voto di diploma

		Freq.	Perc.
Validi	36	2	1,5
	37	1	,8
	38	1	,8
	40	2	1,5
	42	2	1,5
	45	1	,8
	46	1	,8
	48	1	,8
	49	1	,8
	50	1	,8
	54	1	,8
	60	9	6,8
	61	1	,8
	62	4	3,0
	63	1	,8
	64	6	4,5
	65	4	3,0
	67	2	1,5
	68	4	3,0
	69	4	3,0
	70	8	6,1
	71	4	3,0
	73	4	3,0
	75	3	2,3
	76	2	1,5
	77	2	1,5
	78	5	3,8
	80	5	3,8
	81	1	,8
	82	4	3,0
	83	2	1,5
	85	2	1,5
	86	1	,8
	87	2	1,5
	88	1	,8
	90	5	3,8
	91	1	,8
	92	4	3,0
	93	2	1,5
	95	2	1,5
	97	3	2,3
	98	2	1,5
	99	8	6,1
	Totale	122	92,4
Mancanti	Mancante di sistema	10	7,6
Totale		132	100,0

La correzione del voto avviene mediante un'istruzione che modifica il valore della variabile `d4` solo se (**IF**) nella variabile `d4s` c'è il valore 1, il quale indica un voto espresso in sessantesimi. Anche in questo caso si è provveduto a trasformare il valore mancante di sistema in un valore mancante attribuito dall'utente.

```
IF (d4s=1) d4=d4/60*100.
RECODE d4 (SYSMIS=-1).
VALUE LABELS d4c -1 Non risposto.
MISSING VALUES d4c (-1).
```

d4c Voto corretto

		Freq.	Perc.
Validi	60,0	9	6,8
	61,0	1	,8
	61,7	1	,8
	62,0	4	3,0
	63,0	1	,8
	63,3	1	,8
	64,0	6	4,5
	65,0	4	3,0
	66,7	2	1,5
	67,0	2	1,5
	68,0	4	3,0
	69,0	4	3,0
	70,0	10	7,6
	71,0	4	3,0
	73,0	4	3,0
	75,0	4	3,0
	76,0	2	1,5
	76,7	1	,8
	77,0	2	1,5
	78,0	5	3,8
	80,0	6	4,5
	81,0	1	,8
	81,7	1	,8
	82,0	4	3,0
	83,0	2	1,5
	83,3	1	,8
	85,0	2	1,5
	86,0	1	,8
	87,0	2	1,5
	88,0	1	,8
	90,0	6	4,5
	91,0	1	,8
	92,0	4	3,0
	93,0	2	1,5
	95,0	2	1,5
	97,0	3	2,3
	98,0	2	1,5
	99,0	8	6,1
	100,0	2	1,5
	Totale	122	92,4
Mancanti	-1,0 Non risposto	10	7,6
Totale		132	100,0

I voti corretti sono stato poi raggruppati in classi di voto e i nuovi valori attribuiti alla nuova variabile `d4rec`. Si tenga presente che utilizzando il comando **`RECODE … INTO`** tutti i valori della variabile originaria `d4` vanno ricodificati, altrimenti quelli omessi vengono automaticamente trasformati in valori mancanti di sistema.

```
RECODE d4 (60 THRU 69=1)(70 THRU 89=2)(90 THRU 100=3) (-1=-1) INTO d4rec.
VARIABLE LABELS d4rec Voto di diploma.
VALUE LABELS d4rec 1 60-69 “2 70-89” 3 “90-100” 9 "Non risp.".
MISSING VALUES d4rec (9).
```

d4rec Voto di diploma

		Frequenza	Percentuale
Validi	1 60-69	39	29,5
	2 70-89	53	40,2
	3 90-100	30	22,7
	Totale	122	92,4
Mancanti	-1 Non risp.	10	7,6
Totale		132	100,0

5.5 Creazione di un indice (**`COMPUTE`**, **`IF`** *e* **`RECODE`**)

Le due seguenti distribuzioni di frequenza mostrano i titoli di studio posseduti dai padri e dalle madri degli studenti intervistati. Come abbiamo esposto più articolatamente altrove (Delli Zotti 2005), per calcolare un indice del livello culturale dei genitori si possono proporre perlomeno quattro diverse soluzioni e in questa sede ne illustreremo due.

d5p Titolo di studio del padre

		Freq.	Perc.
Validi	1 Elementare o nessuno	9	6,8
	2 Media inferiore	29	22,0
	3 Diploma professionale	22	16,7
	4 Media superiore	45	34,1
	5 Laurea o diploma univ.	25	18,9
	Totale	130	98,5
Mancanti	-1 Non risposto	2	1,5
Totale		132	100,0

d5m Titolo di studio della madre

		Freq.	Perc.
Validi	1 Elementare o nessuno	7	5,3
	2 Media inferiore	34	25,8
	3 Diploma professionale	18	13,6
	4 Media superiore	50	37,9
	5 Laurea o diploma univ.	21	15,9
	Totale	130	98,5
Mancanti	-1 Non risposto	2	1,5
Totale		132	100,0

Si può costruire un semplice indice additivo considerando punteggi di scolarità i valori usati per codificare i titoli di studio (ovviamente si tratta di una forzatura, trattandosi di variabili ordinali). Per non penalizzare gli studenti che hanno indicato il titolo di studio di un solo genitore, con le due istruzioni successive si attribuisce alla nuova variabile `d5gen` il doppio del valore del titolo di studio del genitore del quale si possiede l'informazione.

```
COMPUTE d5gen=d5p+d5m.
IF (d5m=0) d5gen=d5p*2.
IF (d5p=0) d5gen=d5m*2.
VARIABLE LABELS d5gen Punteggi istruzione genitori.
```

d5gen Punteggi istruzione genitori

		Frequenza	Percentuale
Validi	2	5	3,8
	3	5	3,8
	4	19	14,4
	5	12	9,1
	6	20	15,2
	7	12	9,1
	8	27	20,5
	9	19	14,4
	10	12	9,1
	Totale	131	99,2
Mancanti	-1	1	,8
Totale		132	100,0

d5rec Livello culturale dei genitori

		Frequenza	Percentuale
Validi	-1 Non risposto	1	,8
	1 Basso	41	31,1
	2 Medio	59	44,7
	3 Alto	31	23,5
	Totale	132	100,0

Come si vede nella tabella sopra a sinistra, non si è potuto calcolare il punteggio totale per lo studente aveva omesso di indicare il livello di istruzione di entrambi i genitori e perciò, nel costruire la nuova variabile con le classi di livello culturale, ci si dovrà ricordare di ricodificare anche il valore -1, in quanto, come illustrato nel paragrafo precedente, tutti i valori non esplicitamente ricodificati nella nuova variabile diventano mancanti di sistema.

```
RECODE d5gen (-1=-1)(2 3 4 5=1)(6 7 8=2)(9 10=3) INTO d5rec.
VARIABLE LABELS d5rec Livello culturale dei genitori.
VALUE LABELS d5rec -1 Non risposto 1 “Basso” 2 “Medio” 3 “Alto”.
```

La soluzione egualitaria appena vista sarebbe stata tecnicamente più corretta se le due variabili fossero cardinali, e perciò proporremo una seconda soluzione che non implica alcuna forzatura se si concorda sul “principio di dominanza”, attribuendo all'indice di livello culturale il valore del titolo di studio più elevato posseduto dai due genitori.

```
IF (d5p LE d5m) d5tit=d5m.
IF (d5m LE d5p) d5tit=d5p.
VARIABLE LABELS d5tit Titolo di studio più elevato dei genitori.
VALUE LABELS d5tit
-1 Non risposto 1 “Elementare o nessuno” 2 “Media inferiore”
3 “Diploma professionale” 4 “Media superiore” 5 “Laurea o diploma univ.”.
```

d5tit Titolo di studio più elevato dei genitori

		Frequenza	Percentuale
Validi	-1 Non risposto	1	,8
	1 Elementare o nessuno	4	3,0
	2 Media inferiore	22	16,7
	3 Diploma professionale	19	14,4
	4 Media superiore	52	39,4
	5 Laurea o diploma univ.	34	25,8
	Totale	132	100,0

In questo modo si ottiene un risultato piuttosto convincente, in quanto è condivisibile che un indice che vuole rilevare il potenziale di socializzazione culturale dei genitori tenga conto in particolare del genitore più istruito e non venga mortificato da un valore medio. Inoltre, al posto di un indice composto da punteggi astratti ridotti in classi, questa soluzione ci riporta ai più comprensibili titoli di studio proposti nella definizione operativa delle variabili originali.

5.6 Creazione di una tipologia (**COMPUTE**, **IF** *e* **RECODE**)

Nel caso della professione dei genitori costruiremo un indice tipologico perché, dovendo combinare due variabili nominali, è preclusa la strada per la costruzione di indici che implichino un livello almeno ordinale delle variabili originarie. Come spiegato nel dettaglio nel lavoro citato, si tratta di incrociare le informazioni per individuare combinazioni di professioni simili o assimilabili e definire un numero non molto elevato di tipi di famiglie.

d6p Professione del padre

		Freq.	Perc.
Validi	1 Imprend./l.prof./dirig.	33	25,0
	2 Commerc./artig./agric.	9	6,8
	3 Impiegato	31	23,5
	4 Insegnante	8	6,1
	5 Operaio	19	14,4
	7 Altro	28	21,2
	Totale	128	97,0
Mancanti	Mancante di sistema	4	3,0
Totale		132	100,0

d6m Professione della madre

		Freq.	Perc.
Validi	1 Imprend./l.prof./dirig.	10	7,6
	2 Commerc./artig./agric.	12	9,1
	3 Impiegato	37	28,0
	4 Insegnante	17	12,9
	5 Operaio	9	6,8
	6 Casalinga	31	23,5
	7 Altro	14	10,6
	Totale	130	98,5
Mancanti	Mancante di sistema	2	1,5
Totale		132	100,0

Innanzitutto, per rendere più semplici le operazioni successive, aggreghiamo alcune categorie, accorpando le prime due (accomunate dalla condizione di lavoratori autonomi) e le tre successive (lavoratori dipendenti). Accomuniamo anche il dato mancante (genitore non presente) alla risposta "altro" (presumibilmente pensionati o altre condizioni di non occupazione) e forniamo le due variabili di adeguate etichette, per un loro utilizzo in questa nuova versione anche per scopi diversi da quello della creazione di una tipologia professionale dei genitori.

```
RECODE d6p (1 2 =1)(3 4 5=2)(SYSMIS 6 7=3) INTO d6pr.
RECODE d6m (1 2 =1)(3 4 5=2)(SYSMIS 6 7=3) INTO d6mr.
VARIABLE LABELS d6pr Professione del padre /d6mr Professione della madre.
VALUE LABELS d6pr d6mr 1 "Autonomo" 2 "Dipendente" 3 "Non occupato".
```

Il risultato è evidenziato nella tabella seguente, nella quale si possono individuare i tipi di condizione professionale familiare. Ovviamente possono essere proposte tipologie alternative e non ci periteremo di difendere la plausibilità di quella che verrà qui costruita a fini esemplificativi. Come evidenziato con l'uso del colore, proponiamo di costruire un primo tipo composto da famiglie con entrambi i genitori occupati e almeno un lavoratore autonomo. Nel secondo tipo entrambi i genitori sono lavoratori dipendenti, nel terzo un solo genitore è occupato e, infine, nel quarto tipo entrambi i genitori non risultano essere occupati.

Tavola di contingenza d6pr Professione del padre * d6mr Professione della madre

		d6mr Professione della madre			Totale
		1 Autonomo	2 Dipendente	3 Non occupato	
d6pr Professione del padre	1 Autonomo	12	16	14	42
	2 Dipendente	7	37	14	58
	3 Non occupato	3	10	19	32
Totale		22	63	47	132

Per costruire la tipologia in Spss si possono utilizzare una serie di istruzioni **IF**.

```
IF (d6pr EQ 1 AND d6mr EQ 1) d6gen=1.
IF (d6pr EQ 1 AND d6mr EQ 2) d6gen=1.
IF (d6pr EQ 2 AND d6mr EQ 1) d6gen=1.
IF (d6pr EQ 2 AND d6mr EQ 2) d6gen=2.
IF (d6pr EQ 1 AND d6mr NE 3) d6gen=3.
IF (d6pr NE 3 AND d6mr EQ 1) d6gen=3.
IF (d6pr EQ 3 AND d6mr EQ 3) d6gen=4.
```

Come abbiamo visto in precedenza (par. 3.1), con le istruzioni **IF** bisogna prestare particolare attenzione nell'uso degli operatori logici e relazionali e perciò può essere più semplice utilizzare un accorgimento che ci consente di riportare le combinazioni su un'unica variabile, relativamente più semplice da ricodificare. Per fare ciò si sommano le due variabili attribuendo un valore posizionale alla prima moltiplicando per 10 il suo valore. In pratica, nel codice a due cifre risultante le decine indicano la professione del padre e le unità quella della madre.

```
COMPUTE d6gen=d6pr*10+d6mr.
VARIABLE LABELS d6gen Combinazioni professioni genitori.
```

La variabile finale può poi essere costruita facilmente usando un'unica istruzione **RECODE**.

```
RECODE d6gen (11 12 21=1)(22=2)(13 23 31 32=3) (33=4) INTO d6rec.
VARIABLE LABELS d6rec Professione dei genitori.
VALUE LABELS d6rec 1 "Almeno un autonomo" 2 "Entrambi dipendenti"
3 "Un solo occupato" 4 "Entrambi non occupati".
```

d6gen Combinazioni professioni genitori

		Frequenza	Percentuale
Validi	11	12	9,1
	12	16	12,1
	13	14	10,6
	21	7	5,3
	22	37	28,0
	23	14	10,6
	31	3	2,3
	32	10	7,6
	33	19	14,4
	Totale	132	100,0

d6rec Professione dei genitori

		Frequenza	Percentuale
Validi	1 Almeno un autonomo	35	26,5
	2 Entrambi dipendenti	37	28,0
	3 Un solo occupato	41	31,1
	4 Entrambi non occupati	19	14,4
	Totale	132	100,0

5.7 Dicotomizzazione di una variabile (**RECODE**)

La ricodifica della variabile seguente ci consente di illustrare l'utilizzo della parola chiave **ELSE** dell'istruzione **RECODE**; particolarmente potente e proprio perciò potenzialmente pericolosa, come vedremo nel prossimo paragrafo.

d8 A quale anno iscritto

		Frequenza	Percentuale
Validi	1 Primo base	95	72,0
	2 Secondo base	23	17,4
	3 Terzo base	4	3,0
	4 Fuori corso base	5	3,8
	5 Primo spec.	1	,8
	6 Secondo spec.	2	1,5
	7 Fuori corso sp.	2	1,5
	Totale	132	100,0

d8r Anno di corso

		Frequenza	Percentuale
Validi	1 Primo base	95	72,0
	2 Altro anno	37	28,0
	Totale	132	100,0

Volendo semplicemente dicotomizzare la variabile, è particolarmente utile infatti, dopo avere attribuito il valore "1" agli iscritti al primo anno della laurea di base, ricodificare tutti gli altri valori semplicemente citandoli mediante la parola chiave **ELSE**, che significa per l'appunto "tutto il resto".

```
RECODE d8 (1=1)(ELSE=2) INTO d8r.
VARIABLE LABELS d8r "Anno di corso".
VALUE LABELS d8r 1 "Primo base" 2 "Altro anno".
```

5.8 Aggregazione di una variabile territoriale (**RECODE**)

Nel questionario si è chiesto di indicare il comune di residenza e si tratta ovviamente di un'informazione molto analitica che però, proprio per questo, può essere utilizzata per diverse riaggregazioni alternative dei comuni.

d10 Comune di residenza

		Freq.	Perc.
Validi	7 Attimis	1	,8
	23 Cervignano	2	1,5
	49 Lestizza	1	,8
	58 Martignacco	1	,8
	69 Pagnacco	1	,8
	83 Precenicco	1	,8
	102 S.Giovanni al N.	1	,8
	122 Tolmezzo	1	,8
	130 Udine	3	2,3
	132 Venzone	1	,8
	141 Aviano	1	,8
	144 Brugnera	1	,8
	150 Chions	1	,8
	154 Cordenons	1	,8
	158 Fiume Veneto	1	,8
	169 Pordenone	1	,8
	172 Roveredo	1	,8
	175 S.Martino al T.	1	,8
	177 S.Vito al T.	1	,8
	178 Sequals	1	,8
	179 Sesto al Reg.	1	,8
	180 Spilimbergo	2	1,5
	185 Valvasone	1	,8
	188 Zoppola	2	1,5
	190 Cormons	1	,8
	193 Farra	1	,8
	195 Gorizia	8	6,1
	196 Gradisca	1	,8
	199 Medea	2	1,5
	200 Monfalcone	4	3,0
	204 Ronchi dei Leg.	2	1,5
	209 San Pier	2	1,5
	214 Duino-Aurisina	1	,8
	216 Muggia	4	3,0
	217 San Dorligo	1	,8
	218 Sgonico	1	,8
	219 Trieste	51	38,6
	300 Belluno	4	3,0
	303 Treviso	5	3,8
	304 Venezia	9	6,8
	320 Liguria	1	,8
	400 Slovenia	2	1,5
	401 Croazia	1	,8
	Totale	130	98,5
Mancanti	Mancante di sistema	2	1,5
Totale		132	100,0

La conoscenza della localizzazione dei comuni ed il fatto che essi sono elencati in ordine alfabetico all'interno delle province consente una facile recodifica, utilizzando la parola chiave **THRU** (che significa da … a). Si noti che vengono utilizzati nell'istruzione gli intervalli completi dei codici dei comuni (e non solo di quelli presenti in questa occasione nella distribuzione di frequenza) e perciò in questo modo l'istruzione è "riciclabile" ogniqualvolta si utilizzi lo stesso schema di codifica dei comuni.

```
RECODE d10 (1 THRU 137=1) (138 THRU 188=2) (189 THRU 213=3) (214 THRU
  219=4) (ELSE=5) INTO d10r.
VARIABLE LABELS d10r Provincia di residenza.
VALUE LABELS d10r 1 "UD" 2 "PN" 3 "GO" 4 "TS" 5 "Altro".
```

d10r Provincia di residenza

		Frequenza	Percentuale
Validi	1 UD	13	9,8
	2 PN	16	12,1
	3 GO	21	15,9
	4 TS	58	43,9
	5 Altro	24	18,2
	Totale	132	100,0

Il lettore probabilmente accoglierà con scetticismo l'affermazione che "è successo per caso", ma abbiamo appena usato **ELSE** in modo scorretto, a conferma di quanto abbiamo sostenuto nel paragrafo precedente quanto alla "pericolosità" di questa parola chiave. Attribuiti i comuni del Friuli-Venezia Giulia alle 4 province di appartenenza, abbiamo distrattamente classificato gli altri studenti come residenti altrove, senza renderci conto che, in realtà, due di essi non avevano indicato il luogo residenza e nella nuova variabile `d10r` dovrebbero essere per loro costruita la categoria "non risposto". Giocando sull'ambiguità del significato di "altro" (altro luogo di residenza/altra risposta), lasciamo però le cose come stanno.

5.9 Inversione dei codici di una variabile (**RECODE** *o* **COMPUTE**)

I codici attribuiti alle modalità di risposta della domanda relativa alla frequenza alle lezioni sono stati assegnati senza tenere conto che sarebbe stato congruo, vista la natura ordinale della variabile, che il valore più elevato segnalasse una frequenza più assidua alle lezioni.

d13 Frequenza alle lezioni

		Freq.	Perc.
Validi	1 Tutte o quasi	102	77,3
	2 Solo alcune	22	16,7
	3 Mai, o quasi mai	6	4,5
	Totale	130	98,5
Mancanti	Mancante di sistema	2	1,5
Totale		132	100,0

d13 Frequenza alle lezioni

		Freq.	Perc.
Validi	1 Mai, o quasi mai	6	4,5
	2 Solo alcune	22	16,7
	3 Tutte o quasi	102	77,3
	Totale	130	98,5
Mancanti	-1 Non risp.	2	1,5
Totale		132	100,0

L'inversione dei codici si può ottenere facilmente mediante la seguente ricodifica dei valori e, non essendo in alcun modo modificato il contenuto informativo della variabile, è inutile creare una nuova variabile mediante **INTO**. È essenziale invece provvedere, prima di dimenticarsene, ad assegnate i nuovi significati ai valori, altrimenti si rischia di commentare una presunta debacle nella frequenza alle lezioni, dovuta semplicemente al fatto che rimarrebbe assegnata l'etichetta "mai, o quasi mai" al codice 3, che ora però significa invece "tutte o quasi".

```
RECODE d13 (1=3)(2=2)(3=1)(SYSMIS=-1).
VALUE LABELS d13
3 "Tutte o quasi" 2 "Solo alcune" 1 "Mai, o quasi mai" -1 "Non risp."
```

L'istruzione **RECODE** è intuitiva ed efficace ma, a conferma del fatto che nella programmazione si possono spesso trovare soluzioni alternative e magari più eleganti, si consideri che lo stesso risultato si può ottenere con un'appropriata istruzione **COMPUTE**.

```
COMPUTE d13=4-d13.
```

Infatti, sottraendo da 4 i valori della variabile originaria si ottiene 1 per il valore originario 3 (4 – 1=3), 2 per il valore 2 (4 – 2=2) e, infine, 3 per il valore 1 (4 – 3=1).

Come è facile constatare effettuando qualche prova, l'accorgimento funziona anche qualora siano omessi alcuni codici nella sequenza: si dovrà semplicemente sottrarre i valori della variabile originale da un numero superiore di un'unità rispetto al valore massimo della distribuzione. Se, ad esempio, una scala auto-ancorante utilizzasse punteggi da 1 a 5 e si fosse attribuito il codice 9 a "non risposto", basterà sottrarre da 10 i valori della variabile originale. In questo modo i codici si invertono correttamente, ma la sequenza dei valori validi sarà spostata di 4 punti verso l'altro. Ciò non costituisce un problema con una variabile ordinale e basterà assegnare le etichette giuste ai valori.

Se si preferisce che i codici siano gli stessi della variabile originale, ad esempio per calcolare un valore medio che si vuole ricada all'interno dei valori della variabile originale, basterà inserire nella formula una correzione che elimini i punti in esubero. La correzione consiste nella distanza tra l'ultimo valore presente nella sequenza di codici validi e quello fuori sequenza. Nel nostro caso si tratta di 4, in quanto il valore più elevato della scala era 5 e 9 il valore assegnato a "non risposto". Il comando per l'esempio proposto è dunque il seguente:

```
COMPUTE test1=(10-d13)-4.
```

Si tenga conto che con **RECODE** diventano mancanti di sistema tutti i valori non esplicitamente citati nell'istruzione di ricodifica, anche quelli non definiti mancanti nella variabile originale; con **COMPUTE** diventano mancanti di sistema tutti i valori che non è possibile calcolare in quanto nella variabile originale erano mancanti di sistema o definiti mancanti dall'utente.

La procedura può essere percepita come un po' complicata, ma è certamente vantaggiosa quando i valori da invertire sono molto numerosi.

5.10 Conteggio delle ricorrenze di valori in un serie di variabili (**COUNT**)

Nel questionario si chiedeva agli studenti di esprimere un giudizio riguardo all'utilità di una serie di mezzi di informazione al fine di giungere ad una decisione riguardo al corso di laurea a cui iscriversi. La tabella seguente mostra la distribuzione di frequenza delle risposte riguardo alle pagine Internet.

d16_1 Pagine Internet

		Frequenza	Percentuale
Validi	1 Per nulla	11	8,3
	2 Poco	27	20,5
	3 Abbastanza	42	31,8
	4 Molto	49	37,1
	Totale	129	97,7
Mancanti	Mancante di sistema	3	2,3
Totale		132	100,0

Le risposte fornite per tutti i mezzi elencati nel questionario possono essere riassunte in un'unica tabella delle distribuzione di frequenza se si possiede una versione di Spss dotata del modulo **TABLES**. Nella tabella sono omessi i valori mancanti di sistema; è comunque agevole calcolare quanti siano sottraendo per ogni variabile il numero dei casi validi dal totale di 132.

Inoltre, si sono effettuate alcune modifiche rispetto alle scelte iniziali del programma, come l'aggiunta delle percentuali e del totale. Se avessimo eliminato i valori assoluti, con le sole percentuali probabilmente non ci saremmo nemmeno resi conto che le distribuzioni si riferiscono ad un numero variabile di casi. Si è anche chiesto che le distribuzioni di frequenza vengano disposte sulle righe e non sulle colonne e si sono modificate in "N" e "%" le intestazione delle colonne. Infine, si è scelto che le percentuali vengano scritte senza il ridondante simbolo "%" accanto ad ogni singolo valore.

Le percentuali calcolate con esclusione delle mancate risposte non sono ovviamente le stesse che avremmo ottenuto se tutti i casi figurassero nella distribuzione. Però, come abbiamo visto in precedenza, con i valori definiti mancanti di sistema non c'è possibilità di scelta. Invece, trasformandoli in valori gestibili dall'utente (ad esempio -1) potremo produrre la tabella con tutti i casi e le percentuali calcolate comprendendo anche le mancate risposte.

	1 Per nulla		2 Poco		3 Abbastanza		4 Molto		Totale	
	N	%	N	%	N	%	N	%	N	%
Pagina Internet	11	8,5	27	20,9	42	32,6	49	38,0	129	100,0
Guida dello studente	15	11,8	40	31,5	53	41,7	19	15,0	127	100,0
Salone dello studente	76	60,8	26	20,8	16	12,8	7	5,6	125	100,0
Visite all'università	54	43,5	41	33,1	21	16,9	8	6,5	124	100,0
Incontri di orientamento presso la scuola	91	72,8	21	16,8	9	7,2	4	3,2	125	100,0
Incontri con associazioni studentesche	97	78,2	23	18,5	4	3,2			124	100,0
Informazione e pubblicità sulla stampa	68	54,4	42	33,6	14	11,2	1	,8	125	100,0
Informazione e pubblicità radio e televisiva	92	73,6	26	20,8	7	5,6			125	100,0
Singoli incontri con docenti/tutor	83	66,4	18	14,4	17	13,6	7	5,6	125	100,0
Materiali illustrativi ricevuti a domicilio	89	72,4	19	15,4	13	10,6	2	1,6	123	100,0
Informazioni da amici e conoscenti	16	12,8	20	16,0	50	40,0	39	31,2	125	100,0

Un modo ancor più sintetico di rappresentare parte dell'informazione contenuta nella tabella consiste nell'utilizzare la procedura **MULT RESPONSE** per tabulare solo le risposte che definiscono i diversi mezzi come "molto" utili (valore 4)[3]. Con il sistema a finestre si opera in due fasi: si fornisce dapprima un nome e un'etichetta al gruppo di "dicotomie", si indicando le variabili che ne fanno parte ed il valore da tabulare. Si può poi chiedere la distribuzione di frequenza del gruppo di dicotomie e anche realizzare tabelle di contingenza. Per ottenere la distribuzione di frequenza, l'istruzione, creata mediante le finestre di dialogo e poi incollata su un foglio di sintassi, è la seguente:

```
MULT RESPONSE GROUPS=$dom16 "Mezzi d'informazione molto utili"
  (d16_1 TO d16_11 (4))/FREQUENCIES=$dom16.
```

Riepilogo dei casi

	Validi		Mancanti		Totale	
	N	Percentuale	N	Percentuale	N	Percentuale
$Dom16(a)	92	69,7%	40	30,3%	132	100,0%

$Dom16 Frequenze

		Risposte		Percentuale di casi
		N	Percentuale	
Mezzi d'informazione molto utili(a)	Pagine Internet	49	36,0%	53,3%
	Guida dello studente	19	14,0%	20,7%
	Salone dello studente	7	5,1%	7,6%
	Visite all'università	8	5,9%	8,7%
	Incontri di orientamento presso la scuola	4	2,9%	4,3%
	Informazione e pubblicità sulla stampa	1	,7%	1,1%
	Singoli incontri con docenti/tutor	7	5,1%	7,6%
	Materiali illustrativi ricevuti a domicilio	2	1,5%	2,2%
	Informazioni da amici e conoscenti	39	28,7%	42,4%
Totale		136	100,0%	147,8%

a Gruppo a dicotomie incluso nella tabella al valore4.

[3] In Delli Zotti (1999) abbiamo trattato il tema della costruzione di tabelle di contingenza multiple, illustrando come si possono ottenere con **MULT RESPONSE**, anche se non si possiede il modulo **TABLES**.

Si noti che nel nostro esempio la dicotomia è artificiale (la risposta 4 a fronte di tutte le altre) e ciò dimostra però la versatilità dello strumento che può essere utilizzato non solo con dicotomie naturali come sì/no o vero/falso.

Dalla tabella si vede anche che non tutti i mezzi di informazione sono presenti: infatti, mancano i due mezzi (incontro con associazioni studentesche e pubblicità su radio e televisione) che nessuno degli studenti ha ritenuto "molto" utile.

Il lettore attento si sarà accorto che i 49 studenti che ritengono essere stato molto utile l'uso di Internet sono pari al 37,1% nella tabella della distribuzione di frequenza della singola variabile, ma anche pari a esattamente il 38% se ricalcoliamo la percentuale escludendo i 3 studenti che non rispondono alla domanda (come si vede nella tabella della pagina precedente che visualizza le distribuzioni di frequenza di tutte le variabili, escluse le mancate risposte). Può perciò sorprendere notare che nell'ultima tabella gli stessi 49 studenti sono pari al 36% e che, addirittura, nella colonna accanto è indicata una "percentuale di casi" pari a 53,3%.

Per risolvere questo apparente mistero, è utile innanzitutto osservare che, nella tabella "riepilogo dei casi", vengono indicati come mancanti 40 casi. Possiamo scoprire perché costruendo una nuova variabile con il comando **COUNT,** finora non ancora usato negli esempi.

```
COUNT d16n = d16_1 TO d16_11 (4).
VARIABLE LABELS d16n Numero mezzi molto utili.
```

d16n Numero mezzi molto utili

		Frequenza	Percentuale
Validi	0	40	30,3
	1	59	44,7
	2	26	19,7
	3	5	3,8
	5	2	1,5
	Totale	132	100,0

Contando quante volte il codice 4 (molto) viene usato da ognuno degli studenti abbiamo ottenuto una nuova variabile la cui distribuzione mostra che per 40 studenti nessuno degli 11 mezzi di informazione si è rivelato molto utile, altri 59 studenti ritengono sia stato molto utile uno solo e così via, fino ai 2 studenti che hanno definito molto utili cinque mezzi.

Abbiamo così individuato i 40 casi che, non avendo mai utilizzato la risposta "molto", sono stati proprio per questo omessi dalla tabella. Le percentuali perciò erano corrette, ma calcolate su una base diversa e cioè sui soli 92 studenti che ritengono molto utile almeno uno tra gli 11 mezzi di informazione.

Trattandosi di una serie di risposte multiple, non deve sorprendere che il totale delle percentuali sia 147,8%, in che significa che gli studenti che ritengono molto utile almeno un mezzo di informazione attribuiscono nel complesso questa valutazione a 1,5 di essi.

L'altra percentuale esposta nella tabella è, a nostro avviso, di minore utilità e un po' astratta in quanto la base di calcolo è costituita dalle complessive 136 risposte "molto" che hanno fornito i 92 studenti. Questa percentuale consente di valutare correttamente quanto è apprezzato un singolo mezzo in confronto agli altri e sul totale, ma non si tratta di un'informazione particolarmente interessante se non sappiamo anche quanti sono gli studenti che si esprimono in maniera positiva riguardo all'utilità dei mezzi di informazione.

Dopo avere appreso, ad esempio, che un po' più di un terzo delle valutazioni molto positive riguarda la pagine Internet, siamo certamente curiosi di sapere qual è la quota di studenti che così si esprime. Dalle percentuali sui casi vediamo che si tratta di un po' più della metà degli studenti, ma il peso (sulle risposte) avrebbe potuto essere lo stesso anche nel caso di una debacle dei mezzi di informazione, con solo pochi studenti che affermano la loro utilità (se così fosse, citerebbe Internet un

terzo dei pochissimi studenti che ritengono utile qualcuno dei mezzi di informazione elencati). Noi riteniamo che, dovendo scegliere, siano più evocative le percentuali che si riferiscono a "concreti" rispondenti, piuttosto che quelle che si riferiscono ad un "astratto" insieme di risposte.

Sempre a nostro avviso, anche l'insieme di "studenti che definiscono molto utile almeno un mezzo di informazione" è un po' artificioso e dunque può essere utile trovare un modo per tabulare sinteticamente le risposte "molto" senza rinunciare alla base di calcolo delle percentuali costituita dall'insieme di tutti gli studenti. Ciò può essere fatto ricodificando come "4" il valore "0" che identifica nella distribuzione della variabile `d16n` i 40 studenti che non ritengono sia stato utile alcun mezzo.

```
RECODE d16n (0=4)(ELSE=0) INTO d16rec.
```

d16rec Nessun mezzo molto utile

		Frequenza	Percentuale
Validi	0 No	92	69,7
	4 Sì	40	30,3
	Totale	132	100,0

Con la trasformazione è stata creata la nuova variabile `d16rec` per lasciare intatto il potenziale informativo della variabile `d16n` con la quale abbiamo contato il numero di mezzi ritenuti molto utili. Anche in questo caso i lettori più attenti si saranno resi conto che, per risolvere un problema con le percentuali, abbiamo ottenuto nel contempo anche un indicatore dell'utilità attribuita dagli studenti all'insieme dei mezzi di informazione.

Si pensi, ad esempio, al caso in cui la stessa operazione venga svolta contando il numero di istituzioni che i rispondenti considerano degne di "molta" fiducia o, per tornare ad una dicotomia naturale, al conteggio della associazioni cui gli intervistati sono iscritti, o, ancora, degli elettrodomestici e/o dei prodotti di alta tecnologia posseduti da una famiglia. Si noti anche che il conteggio produce una variabile cardinale, "preziosa" in quanto usabile con tutte le tecniche statistiche, ma di norma assai rara nelle matrici dei dati di sondaggio.

Ritornando alla distribuzione di frequenza delle risposte multiple, possiamo ora tabulare anche questa ultima variabile congiuntamente alle altre con la procedura **`MULT RESPONSE`**.

```
MULT RESPONSE GROUPS=$Dom16 "Mezzi d'informazione molto utili"
  (d16_1 TO d16_11 d16rec (4)) /FREQUENCIES=$Dom16.
```

$Dom16 Frequenze

		Risposte N	Risposte Percentuale	Percentuale di casi
Mezzi d'informazione molto utili(a)	Pagine Internet	49	27,8%	37,1%
	Guida dello studente	19	10,8%	14,4%
	Salone dello studente	7	4,0%	5,3%
	Visite all'università	8	4,5%	6,1%
	Incontri di orientamento presso la scuola	4	2,3%	3,0%
	Informazione e pubblicità sulla stampa	1	,6%	,8%
	Singoli incontri con docenti/tutor	7	4,0%	5,3%
	Materiali illustrativi ricevuti a domicilio	2	1,1%	1,5%
	Informazioni da amici e conoscenti	39	22,2%	29,5%
	Nessun mezzo molto utile	40	22,7%	30,3%
Totale		176	100,0%	133,3%

a Gruppo a dicotomie incluso nella tabella al valore4.

Dalla tabella finale si vede che la percentuale riferita ai 132 casi torna ad essere pari all'originario 37,1%. La percentuale sulle risposte (27,8%) è diventata ancor meno utile in quanto la base è del tutto artificiosa, comprendendo ora anche le 40 risposte che indicano nessun "molto".

5.11 Calcolo della risposta media ad una batteria di domande (**COMPUTE**)

La batteria di domande appena esplorata si può prestare anche ad un'altra utile operazione: il calcolo del punteggio medio attribuito da ogni intervistato agli item che compongono la batteria. In mancanza di variabili cardinali nell'archivio che stiamo usando, a fini didattici operiamo consapevolmente una forzatura, calcolando le medie di variabili ordinali.

Innanzitutto, calcoliamo separatamente il punteggio medio per ognuna delle 11 domande, chiedendo inoltre di ordinare i mezzi di informazione secondo valori decrescenti delle medie.

```
DESCRIPTIVES VARIABLES=d16_1 TO d16_11 /STATISTICS=MEAN /SORT=MEAN (D).
```

Statistiche descrittive

	N	Media
Pagine Internet	129	3,00
Informazioni da amici e conoscenti	125	2,90
Guida dello studente	127	2,60
Visite all'università	124	1,86
Salone dello studente	125	1,63
Singoli incontri con docenti/tutor	125	1,58
Informazione e pubblicità sulla stampa	125	1,58
Materiali illustrativi ricevuti a domicilio	123	1,41
Incontri di orientamento presso la scuola	125	1,41
Informazione e pubblicità radio e televisiva	125	1,32
Incontri con associazioni studentesche	124	1,25
Validi (listwise)	120	

Dalla tabella si evince che il mezzo di cui è stata maggiormente apprezzata l'utilità ha raggiunto un punteggio di 3,00 (in quella che con qualche forzatura potremmo considerare una scala da 1 a 4) e quello apprezzato di meno il punteggio di 1,25. Analogamente alla tabella in cui si sono tabulate le percentuali di risposta, anche in questo si nota che risposte valide sono state fornite da un numero di studenti che varia da 123 a 129, a seconda del mezzo e che i casi validi **listwise** (cioè quelli che hanno fornito risposte valide per tutti gli item) sono 120.

Per calcolare il punteggio medio su tutti gli item si possono ovviamente sommare i punteggi e dividerli per 11.

```
COMPUTE d16punt=(d16_1+d16_2+d16_3+d16_4+d16_5+d16_6+d16_7+d16_8+d16_9+
  d16_10+d16_11)/11.
VARIABLE LABELS d16punt Utilità media dei mezzi di informazione.
FREQUENCIES VARIABLES=d16punt /STATISTICS=MEAN /ORDER=ANALYSIS.
```

Come si vede nelle tabelle seguenti il calcolo è stato ovviamente effettuato solo per i 120 studenti che avevano espresso il giudizio di utilità per tutti gli 11 mezzi di informazioni, come era scritto anche in fondo alla tabella che riportava le medie per ogni singolo mezzo. I valori "mancanti di sistema", oltre che nel caso di celle vuote, appaiono infatti anche quando non può essere calcolato il valore di una nuova variabile per la mancanza di dati validi in uno o più dei termini di un'espressione calcolata mediante l'istruzione **COMPUTE**.

Statistiche

N	Validi	120
	Mancanti	12
Media		1,858

d16punt Utilità media dei mezzi di informazione

		Frequenza	Percentuale cumulata
Validi	1,2	1	,8
	1,3	6	5,8
	1,5	9	13,3
	1,5	10	21,7
	1,6	19	37,5
	1,7	8	44,2
	1,8	15	56,7
	1,9	13	67,5
	2,0	8	74,2
	2,1	7	80,0
	2,2	4	83,3
	2,3	5	87,5
	2,4	6	92,5
	2,5	4	95,8
	2,5	2	97,5
	2,8	1	98,3
	2,9	1	99,2
	3,1	1	100,0
	Totale	120	
Mancanti	Mancante di sistema	12	
Totale		132	

Infatti, nell'espressione **(var1+var2+var3)/3** il risultato è mancante se un caso include un valore mancante anche per una sola delle tre variabili e ciò può tradursi in una drastica perdita di casi. Per ovviare a questo inconveniente si può usare in alternativa la funzione **MEAN (var1 var2 var3)** con la quale il risultato è mancante solo se un caso include valori mancanti per tutte le variabili. È anche possibile specificare il numero minimo di argomenti che devono includere valori non mancanti, digitando un punto e il numero minimo dopo il nome della funzione. Se vogliamo che la media venga calcolata nel caso in cui sino state espresse le valutazioni per almeno due mezzi di informazione, l'espressione è la seguente:

```
COMPUTE punt16=MEAN.2 (d16_1 TO d16_11).
```

Statistiche descrittive

	N	Media
Utilità media dei mezzi di informazione	127	1,89
Validi (listwise)	127	

I casi su cui è calcolata la media salgono a 127, con il recupero di 7 rispetto alla situazione precedente. Si riesce così ad ovviare ad una perdita di casi che può diventare molto elevata al crescere del numero di variabili e delle mancate risposte. Non va però dimenticato che le media generale è stata calcolata per alcuni studenti utilizzando tutte le informazioni e per altri sulla base della valutazione dell'utilità anche di due soli mezzi di informazione.

5.12 Digressione sulle mancate risposte (**COUNT**, **DO IF**, **RECODE**)

Osservando più attentamente i 12 casi nei quali si sono registrate mancate risposte per alcuni dei mezzi di informazione, ci accorgiamo che, a voler ben interpretare l' "intenzione di voto" degli studenti, forse non si tratta sempre di vere e proprie mancate risposte. Infatti, alcuni studenti hanno definito molto utili uno o due mezzi di informazione, e magari qualcun altro abbastanza utile, lasciando in bianco le altre caselle del questionario. A nostro avviso, è molto ragionevole presumere, a maggior ragione trattandosi di un questionario auto compilato, che gli studenti abbiano voluto segnalare solo i mezzi più o meno utili, sottintendendo che quelli non contrassegnati non li hanno ritenuti utili.

L'esperienza insegna che, nonostante le raccomandazioni fatte da chi coordina la rilevazione, ciò accade a volte anche con intervistatori opportunamente addestrati. Per velocizzare l'esecuzione dell'intervista si limitano a segnare le caselle "Sì" che corrispondono ad esempio alle associazioni di cui gli intervistati fanno parte, sembrando loro superfluo contrassegnare le caselle "No" per le altre. Ovviamente, questa nostra interpretazione è particolarmente credibile in presenza di almeno una risposta "Sì" all'insieme delle domande o, nell'esempio che stiamo facendo in questa sede, di una risposta valida per almeno uno dei mezzi di informazio-

ne. Se, invece, per tutta la batteria di domande in questione non ci fosse alcun segno di risposta, è meno certo che ci si trovi di fronte ad una totalità di risposte negative e potrebbe essere invece realistico ritenere che l'intervistato non abbia davvero inteso rispondere alla domanda.

Il ragionamento regge ovviamente a patto che non ci sia il ragionevole dubbio che, pur avendo risposto per alcune delle domande nella batteria, l'intervistato davvero non abbia voluto o potuto esprimersi su alcune delle altre. Alcune domande, ad esempio, potrebbero presupporre conoscenze che l'intervistato non possiede o comportamenti non praticati, oppure potrebbero riferirsi a questioni sulle quali l'intervistato vuole mantenere il riserbo. Tutto ciò comunque, mette in luce il fatto che, nel caso delle batterie di domande, ci sono almeno due tipi di mancate risposte che è opportuno tenere distinte, anche al fine di un loro utilizzo selettivo. C'è la mancata risposta "globale", cioè il rifiuto o la difficoltà di rispondere all'insieme delle domande, e poi la mancata risposta "locale" o "selettiva", cioè il rifiuto o la difficoltà di rispondere ad un singolo item in quanto, ad esempio, quel singolo aspetto non è conosciuto o non pertinente, o considerato un fatto privato, e così via.

Quando si può ragionevolmente ritenere che i dati mancanti nella matrice dei dati si possano interpretare come risposta negativa (nel nostro caso "per nulla utile") si possono ricodificare i codici "non risposto" eventualmente attribuiti in fase di codifica. In presenza di sole mancate risposte locali, come nel caso della matrice che stiamo utilizzando nelle presenti esemplificazioni, la ricodifica può essere effettuata semplicemente mediante la seguente istruzione:

```
RECODE d16_1 TO d16_11 (-1=1).
```

Volendo generalizzare la soluzione, conviene usare una batteria nella quale siano compresenti le mancate risposte globali e quelle locali, come la domanda 29 che chiedeva di esprimere un giudizio sui servizi offerti dalla Facoltà e dall'Università. Come si vede dalla tabella, era già stato previsto in fase di costruzione del questionario che alcuni studenti, tra i quali numerose matricole, potessero non essere in grado di esprimersi per i servizi usati più di rado. Siccome gli studenti sono in tutto 132, accanto alle esplicite risposte "non so/non conosco", vi è un numero variabile di "mancanti di sistema", da un minimo di due (completezza informazioni e orari delle lezioni), ad un massimo di 6 (frequenza appelli).

	1 Insufficiente		2 Sufficiente		3 Buono		4 Non so/ non conosco		Totale	
	N	%	N	%	N	%	N	%	N	%
Orario Centro Servizi	24	18,8	49	38,3	33	25,8	22	17,2	128	100,0
Efficienza Centro Servizi	16	12,4	45	34,9	48	37,2	20	15,5	129	100,0
Completezza informazioni	14	10,8	52	40,0	61	46,9	3	2,3	130	100,0
Tempestività informazioni	23	18,0	47	36,7	51	39,8	7	5,5	128	100,0
Informazione su appelli	6	4,7	45	34,9	51	39,5	27	20,9	129	100,0
Frequenza appelli	11	8,7	24	19,0	52	41,3	39	31,0	126	100,0
Sportello dello studente	25	19,4	40	31,0	25	19,4	39	30,2	129	100,0
Disponibilità computer	28	21,7	34	26,4	37	28,7	30	23,3	129	100,0
Collegamenti a Internet	29	22,5	29	22,5	42	32,6	29	22,5	129	100,0
Disponibilità fotocopiatrici	36	27,9	34	26,4	12	9,3	47	36,4	129	100,0
Spazi per studio	12	9,4	45	35,4	48	37,8	22	17,3	127	100,0
Orari delle lezioni	6	4,6	59	45,4	59	45,4	6	4,6	130	100,0
Aule per le lezioni	9	7,0	40	31,0	77	59,7	3	2,3	129	100,0

Si tratta ora di individuare le risposte mancanti globali, in modo da poterle distinguere dalla mancata risposta a singole domande. Conosciamo di già l'istruzione che ci consente di contare il numero di occorrenze di uno o più valori in un insieme di variabili e nel nostro caso essa può essere la seguente.

```
COUNT d16mis = d16_1 TO d16_11 (-1).
VARIABLE LABELS d16mis Mancate risposte alla dom16.
```

d29mis Mancate risposte alla dom29

		Frequenza	Percentuale
Validi	0	49	37,1
	1	13	9,8
	2	8	6,1
	3	17	12,9
	4	10	7,6
	5	14	10,6
	6	12	9,1
	7	6	4,5
	8	1	,8
	13	2	1,5
	Totale	132	100,0

Come si vede, oltre a 49 studenti che hanno fornito risposte complete, ve ne sono 13 che hanno omesso una sola risposta e, all'opposto, due studenti che non hanno risposto a tutte le domande. Queste ultime possono essere considerata mancate risposte globali e possiamo assegnare loro uno specifico codice, al fine di poterle distinguere dalle mancate risposte locali, condizionando la decodifica, mediante la struttura **DO IF-END IF**, alla condizione che la variabile `d29mis` sia uguale a 13.

Avendo assegnato il valore -1 (non risposto) alle risposte mancanti globali, possiamo ora uniformare le mancate risposte locali ricodificando come 4 (non so/non conosco) i residui valori mancanti di sistema (**SYSMIS**).

```
DO IF (d29mis EQ 13).
RECODE d29_1 TO d29_13 (SYSMIS 4=-1).
END IF.
RECODE d29_1 TO d29_13 (SYSMIS=4).
```

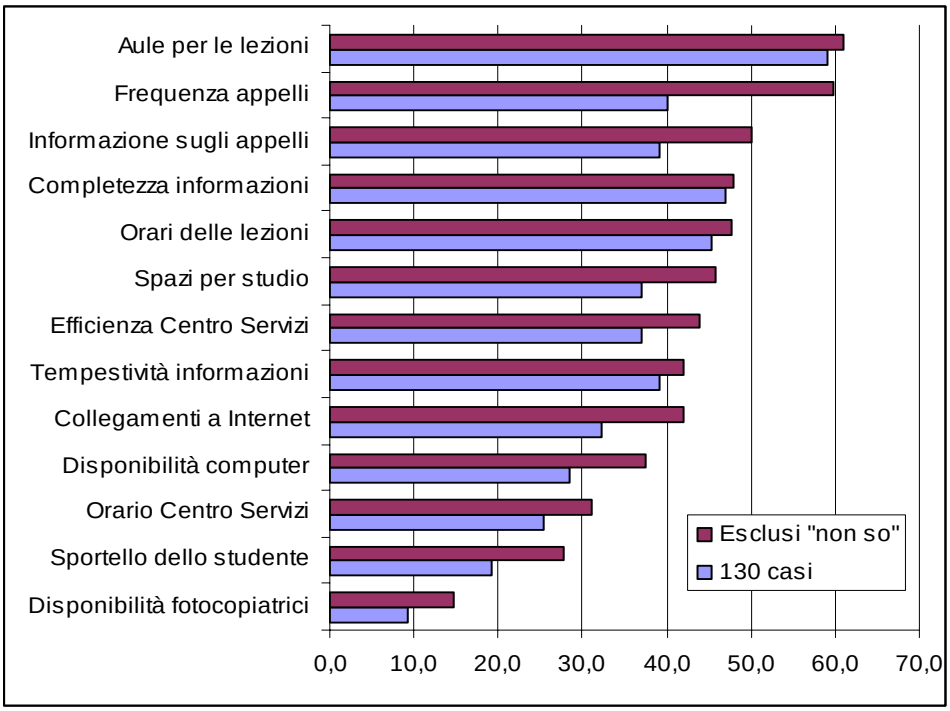
	1 Insufficiente		2 Sufficiente		3 Buono		4 Non so/ non conosco		Totale	
	N	%	N	%	N	%	N	%	N	%
Orario Centro Servizi	24	18,5	49	37,7	33	25,4	24	18,5	130	100,0
Efficienza Centro Servizi	16	12,3	45	34,6	48	36,9	21	16,2	130	100,0
Completezza informazioni	14	10,8	52	40,0	61	46,9	3	2,3	130	100,0
Tempestività informazioni	23	17,7	47	36,2	51	39,2	9	6,9	130	100,0
Informazione sugli appelli	6	4,6	45	34,6	51	39,2	28	21,5	130	100,0
Frequenza appelli	11	8,5	24	18,5	52	40,0	43	33,1	130	100,0
Sportello dello studente	25	19,2	40	30,8	25	19,2	40	30,8	130	100,0
Disponibilità computer	28	21,5	34	26,2	37	28,5	31	23,8	130	100,0
Collegamenti a Internet	29	22,3	29	22,3	42	32,3	30	23,1	130	100,0
Disponibilità fotocopiatrici	36	27,7	34	26,2	12	9,2	48	36,9	130	100,0
Spazi per studio	12	9,2	45	34,6	48	36,9	25	19,2	130	100,0
Orari delle lezioni	6	4,6	59	45,4	59	45,4	6	4,6	130	100,0
Aule per le lezioni	9	6,9	40	30,8	77	59,2	4	3,1	130	100,0

A questo punto le mancate risposte si possono gestire a piacere, reintegrando le mancate risposte globali, in modo che le percentuali siano calcolate su un totale di 132 studenti (dichiarando non mancante il valore –1), oppure dichiarando come mancante anche il valore 4 (non

so/non conosco) in modo che le percentuali siano calcolate sul numero variabile di studenti che hanno espresso la valutazione sui singoli servizi. Il programma infatti consente di definire come mancanti fino a 3 valori o intervalli di valori.

Nel grafico che segue si possono notare gli effetti dell'inclusione/esclusione delle risposte mancanti. In generale aumenta la percentuale di risposte positive (come del resto di quelle che danno un giudizio di insufficienza, non visualizzate nel grafico), ma la diversa consistenza delle risposte "non so" a seconda del servizio considerato fa sì che a volte il cambiamento sia minimo e altre volte invece assai consistente, al punto che cambia la posizione di alcuni servizi nella gerarchia di "gradimento". Ad esempio, le valutazioni positive collocano la frequenza e le informazioni sugli appelli d'esame al secondo e terzo posto, ma si troverebbero tre posizioni più in basso se le percentuali tenessero conto anche del parere degli studenti che non si ritengono sufficientemente informati su questi aspetti.

Percentuale di risposte "buono" per una serie di servizi erogati



5.13 Dicotomizzazione da risposte multiple (**RECODE**, **COMPUTE**, **IF** *e* **DO REPEAT**)

Riprendendo la scansione del questionario utilizzato per le presenti esemplificazioni, troviamo una domanda con la quale si potevano effettuare al massimo tre scelte da una lista unica che proponeva i motivi che potrebbero avere influenzato la scelta di studiare all'Università. Si tratta di scelte multiple che vanno tabulate mediante il modulo **MULT RESPONSE** ma, a differenza di quanto visto a proposito dei mezzi di informazione, si tratta di un gruppo di risposte multiple, invece che dicotomie multiple. Quando si crea un gruppo di dicotomie multiple ogni variabile con almeno una risposta di quelle che si è deciso di tabulare (ad esempio 1=Sì), diventa una categoria. Quando invece si crea un gruppo di risposte multiple ogni valore diventa una categoria ed il programma calcola le sue frequenze aggiungendo quelle di tutte le variabili del gruppo che contengono quel valore.

La sintassi è molto simile, in quanto si tratta di indicare tra parentesi, invece dell'unico valore delle dicotomia da tabulare, l'intervallo di valori utilizzati per codificare tutte le risposte, compresi gli eventuali "non risposto", se non definiti come valori mancanti. Nell'istruzione abbiamo "largheggiato", per non controllare fino a che valore arrivassero le risposte valide e inoltre si nota un curioso -3 che va spiegato. Come si vede dall'istruzione **RECODE**, si sono ri-

codificati rispettivamente come -2 e -3 i valori mancanti della seconda e terza variabile, inizialmente anch'essi uguali a -1, al fine di poter distinguere le risposte non date alla prima, alla seconda e alla terza richiesta di indicare un motivo per la scelta di proseguire gli studi. In questo modo, invece del complessivo 63% circa di mancate risposte, sappiamo nel dettaglio che quasi tutti hanno fornito almeno un motivo, un po' più dell'80% ne ha fornito un secondo e più della metà ha indicato anche un terzo motivo. Il 63% circa di mancate risposte complessive si può anche interpretare sotto forma di numero medio di motivi indicati, calcolabile come (300–63=227)/100=2,27, si può cioè affermare che mediamente gli studenti hanno fornito 2,3 motivi circa e, a nostro avviso, anche questa informazione residuale può essere molto interessante. Infatti, con questo semplice parametro possiamo confrontare gruppi di studenti diversi considerandolo un indicatore che con valori bassi può indicare scelte di tipo "vocazionale" e con valori alti invece minore determinazione nella scelta.

```
RECODE d19_2 (-1=-2) /d19_3 (-1=-3).
MULT RESPONSE GROUPS=$dom19 'Motivi scelta studiare all'università'
  (d19_1 d19_2 d19_3 (-3,20)) /FREQUENCIES=$dom19.
```

		Risposte		Percentuale di casi
		N	Percentuale	
Motivi scelta studiare all'Università (a)	Non risp. III	58	14,6%	43,9%
	Non risp. II	24	6,1%	18,2%
	Non risp. I	1	,3%	,8%
	Sempre buoni risultati a scuola	33	8,3%	25,0%
	Difficile trovare lavoro	19	4,8%	14,4%
	Genitori si aspettavano che continuassi	28	7,1%	21,2%
	Interessava la ricerca scientifica	14	3,5%	10,6%
	Preciso settore di interesse	60	15,2%	45,5%
	Con laurea più facile trovare lavoro	65	16,4%	49,2%
	Titolo universitario dà prestigio	40	10,1%	30,3%
	Laurea necessaria per attività che intendo svolgere	37	9,3%	28,0%
	Altro motivo	17	4,3%	12,9%
Totale		396	100,0%	300,0%

a Gruppo

Dalla tabella si vede una volta di più, ad avviso di chi scrive, che sono molto più utili le percentuali di risposta basate sui casi, dalle quali si apprende che quasi esattamente la metà degli studenti indica tra i tre motivi anche la speranza, del resto fondata, che con la laurea sia più facile trovare lavoro. Si tratta di un'informazione di gran lunga più significativa rispetto ad apprendere che, su una lista di otto motivi, cui si aggiunge un eterogeneo "altri motivi" e circa un 20% di mancate risposte, quello specifico motivo pesa per circa il 16%.

Diversi motivi sono state scelti da un numero elevato di intervistati, al punto che forse si sarebbe potuto chiedere direttamente in forma dicotomica "Sì/No", quali motivi abbiano influenzato la scelta. Si è deciso invece di limitare a tra le possibili scelte e al contempo si può concordare che le variabili multiple sono un po' complesse ed il loro utilizzo limitato alla realizzazione di distribuzioni di frequenza e tabelle di contingenza. Perciò, una volta ottenuto lo scopo di costringere gli studenti ad essere selettivi, ci si può chiedere se non sia possibile ricondurre le risposte alla forma dicotomica in modo da poter utilizzare ogni singola informazione indipendentemente dalle altre.

Il procedimento è noto nella letteratura metodologica come uso delle variabili *dummy* (di comodo) e consiste nel ridurre ad alcune dicotomie le categorie principali di una variabile categoriale come, ad esempio, la religione di appartenenza. Partendo da: 1=cristiano, 2=ebreo, 3=musulmano, 4=altro, si crea una variabile per i cristiani ricodificando come 0=No le reli-

gioni diverse dalla prima, e poi ricodificando in 1=Sì/0=No l'appartenenza alla religione ebraica e così via.

Nel nostro caso la risposta positiva per uno qualsiasi dei motivi può essere presente in ognuna delle tre variabili che compongono il gruppo e perciò le istruzioni diventano solo un po' più complesse. Innanzitutto conviene costruire con una struttura **DO REPEAT-END REPEAT** una serie di variabili per i 9 motivi, poste inizialmente uguali a 0=No. Poi, con una serie di **IF** si modifica in 1=Sì il valore di ognuna delle variabili nel caso in cui ognuno dei 9 diversi valori siano presenti nelle tre variabili sulle quali sono state registrate le scelte. La variabile segnaposto DOM sostituisce la lista di variabili d19_1d **TO** d19_9d all'interno della struttura **DO REPEAT-END REPEAT,** evitando di ripetere per 9 volte le stesse istruzioni**.**

```
DO REPEAT DOM = d19_1d TO d19_9d.
COMPUTE DOM=0.
IF (d19_1 EQ 1 OR d19_2 EQ 1 OR d19_3 EQ 1) DOM=1.
END REPEAT.
VARIABLE LABELS d19_1d Sempre buoni risultati a scuola
...
/d19_9d Laurea necessaria per attività che intendo svolgere.
VALUE LABELS d19_1d TO d19_9d
0 No
1 Sì.
```

La tabella seguente mostra il risultato definitivo e, come si vede, le percentuali di risposte affermative sono esattamente le stesse viste sopra, ma ora sono il complemento a 100% di una serie di variabili usabili indipendentemente le une dalle altre. L'unica novità consiste nell'apparizione di una variabile che si riferisce al rinvio del servizio di leva, risposta effettivamente prevista nel questionario, ma che nessuno aveva scelto e perciò non era apparsa nella tabulazione multipla delle risposte.

	0 No		1 Sì		Totale	
	N	%	N	%	N	%
Sempre buoni risultati a scuola	99	75,0	33	25,0	132	100,0
Difficile trovare lavoro	113	85,6	19	14,4	132	100,0
Genitori si aspettavano che continuassi	104	78,8	28	21,2	132	100,0
Interessava la ricerca scientifica	118	89,4	14	10,6	132	100,0
Rinviare servizio di leva	132	100,0			132	100,0
Preciso settore di interesse	72	54,5	60	45,5	132	100,0
Con laurea più facile trovare lavoro	67	50,8	65	49,2	132	100,0
Titolo universitario dà prestigio	92	69,7	40	30,3	132	100,0
Laurea necessaria per attività che intendo svolgere	100	75,8	32	24,2	132	100,0

5.14 La differenza tra **RECODE … INTO** *e* **COMPUTE/RECODE**

I nomi dei comandi **COMPUTE** e **RECODE** alludono a funzioni diverse (calcolo di valori per il primo e ricodifica di valori per il secondo) e vi sono tra loro differenze di sintassi. Si sarà però notato che entrambi possono servire a costruire nuove variabili (**COMPUTE** lo fa esplicitamente e **RECODE** utilizzando **INTO**) e, d'altro canto, anche **COMPUTE** (come **RECODE)** può servire a trasformare i valori di variabili esistenti, come nell'esempio visto in precedenza della standardizzazione dei voti (**COMPUTE d4=d4/60*100**).

Ci sono poi casi in cui i due comandi possono essere combinati al fine di ottenere un determinato risultato: nel paragrafo precedente abbiamo usato **COMPUTE** per generare una variabile e assegnarle un valore iniziale come una costante o i valori di una variabile già esistente. Questi valori possono poi essere modificati con successivi comandi (ad esempio **IF** o **RECO-**

DE). Perciò, si potrebbe ritenere che la coppia di istruzioni **COMPUTE/RECODE** e il **RECODE … INTO** che seguono siano perfettamente equivalenti. Con la prima soluzione creiamo una nuova variabile (**COMPUTE)** e poi la ricodifichiamo (**RECODE)**; con la seconda dapprima ricodifichiamo i valori di una variabile esistente (**RECODE)** e poi assegniamo il risultato della trasformazione ad una nuova variabile (**INTO)**. In realtà, il risultato è equivalente se tutto gira per il verso giusto, ma la prima soluzione, apparentemente più complessa, è in alcuni casi quella consigliabile.

```
COMPUTE newvar = oldvar.
RECODE newvar (1 THRU 15=1)(16 THRU XX=2)(ecc).

RECODE oldvar (1 THRU 15=1)(16 THRU XX=2)(ecc) INTO newvar.
```

Partendo dalla distribuzione di frequenza dei comuni di residenza gli studenti (vedi par. 5.8), potremmo effettuare una ricodifica per zone che tenga conto della vicinanza dei comuni dalle linee ferroviarie che partono da Trieste, per individuare gli studenti che presumibilmente ne hanno tenuto conto nel decidere di scegliere questa sede universitaria. La ricodifica non è semplice come quella in province, la quale era facilitata dal fatto che i codici sono stati assegnati ai comuni secondo l'ordine alfabetico all'interno delle quattro province. Può essere perciò prudente procedere per gradi, cioè attribuendo alle diverse zone dapprima i comuni dei quali si è certi e, sgombrato il campo da numerosi nomi di comuni e codici, procedere successivamente all'assegnazione di quelli che si trovano in posizione intermedia, magari tenendo conto della consistenza dei gruppi che si vanno a formare.

Nelle due tabelle che seguono però si vede che è diverso il risultato intermedio che si ottiene adottando le due soluzioni.

d10zona Zona di residenza

		Freq.	Perc.
Validi	1 Trieste e provincia	58	43,9
	2 Limitrofa ferrovia	45	34,1
	4 Altro	8	6,1
	7	1	,8
	58	1	,8
	69	1	,8
	102	1	,8
	122	1	,8
	132	1	,8
	141	1	,8
	178	1	,8
	179	1	,8
	180	2	1,5
	185	1	,8
	188	2	1,5
	193	1	,8
	199	2	1,5
	209	2	1,5
	Totale	130	98,5
Mancanti	Mancante di sistema	2	1,5
Totale		132	100,0

d10zona Zona di residenza

		Freq.	Perc.
Validi	1 Trieste e provincia	58	43,9
	2 Limitrofa ferrovia	45	34,1
	4 Altro	8	6,1
	Totale	111	84,1
Mancanti	Mancante di sistema	21	15,9
Totale		132	100,0

Nel primo caso abbiamo inizialmente assegnato alla variabile `d10zona` i valori della variabile originaria `d10` e con la ricodifica parziale rimangano immutati i codici dei comuni non ancora assegnati alle tre zone che sono state create. Successivamente si potrà decidere se vale la pena di creare una quarta area, visto che la categoria "altro" per ora si riferisce sostanzial-

mente a studenti che vengono da lontano o dall'estero, oppure assegnare qualche altro comune alla zona influenzata dalla presenza della ferrovia e aggregare i comuni rimanenti alla categoria "altro".

Se invece utilizziamo l'istruzione **RECODE** con **INTO**, la distribuzione di frequenza intermedia che viene prodotta impedisce l'ulteriore perfezionamento dell'assegnazione dei comuni alle aree in quanto tutti i valori non assegnati vengono automaticamente inseriti nella categoria "mancante di sistema" e diventano perciò indistinguibili.

È comunque tranquillizzante sapere che in entrambi i casi abbiamo operato in maniera reversibile perché in entrambi i casi rimane immutata la variabile originale `d10` e perciò siamo sempre in tempo a rimediare rilanciando le istruzioni corrette, cosa che non accade se invece ci limitiamo a ricodificare la variabile originale senza crearne una nuova.

6. Conclusioni

Il testo doveva essere completato con un'ultima esemplificazione che alla fine chi scrive ha deciso di non utilizzare e il motivo dell'esclusione si presta particolarmente bene per offrire al lettore qualche consiglio conclusivo.

L'ultimo esempio di trasformazione doveva riguardare il problema del calcolo della distanza temporale tra due eventi. Una soluzione elegante consiste nel registrare le informazioni in una variabile Spss "data", che tra i vari formati offre anche quello in ore e minuti (hh:mm). Per calcolare la distanza temporale in questo caso basta detrarre, mediante il comando **COMPUTE**, la data e/o l'orario più recente da quello del passato e il computer calcolerà la distanza automaticamente tenendo conto ovviamente che, ad esempio, le ore sono 24 e i minuti 60.

Ma i dati possono essere stati registrati su un programma esterno in campi separati per ore, minuti ed eventualmente il giorno; se si vuole ugualmente calcolare la distanza tra i due eventi si doveva nel passato fare un po' di sfoggio di abilità nell'utilizzo dei comandi di trasformazione dei dati e perciò il problema si prestava particolarmente bene per chiudere con le esemplificazioni.

Infatti, siccome l'informazione completa si trova su più variabili bisognava ricondurle ad una sola, aggiungendo ai minuti le ore trasformate in minuti moltiplicandole per 60. In questo modo, in pratica, si ottiene la distanza in muniti dalla mezzanotte e, se l'ora d'inizio è collocata in un giorno precedente a quello della conclusione, bisognerà aggiungere i minuti trascorsi nella giornata o nelle giornate precedenti (per ogni giorno si dovranno aggiungere 1440 minuti - 24 ore x 30 minuti). A questo punto, sottraendo i minuti della fine dell'evento da quelli del suo inizio, abbiamo la durata dell'evento in minuti. Basterà ora dividere i minuti per sessanta ed avremo la durata in ore (il resto della divisione corrisponde a centesimi di ora che vanno poi trasformati in minuti).

Come si vede, si tratta di un problema complesso che si risolve con l'utilizzo di diversi comandi tra quelli che abbiamo fin qui illustrato, ma è un problema che Spss ha già risolto e dunque che non vale più la pena di illustrarlo. Infatti, Spss non solo consente di inserire le informazioni dall'inizio come tipo "data/ora", ma anche di indicare le variabili nelle quali abbiamo registrato le ore, i minuti, ed eventualmente i giorni, etc. e, sulla base di queste indicazioni, il programma provvederà automaticamente a costruire la variabile "data/ora" e a proporre addirittura diversi formati nei quale scriverla.

Dunque, un problema risolto, ma da questa vicenda possiamo trarre ugualmente alcuni insegnamenti che si prestano a concludere il presente lavoro. Il primo insegnamento è che programmi come Spss diventano complessi e sempre più completi in quanto il loro successo e dunque la loro crescente diffusione comportano l'accostarsi a nuove frange di utenza o "nicchie di mercato" con specifici problemi conoscitivi. Questi problemi vengono alla luce attraverso pubblicazioni, forum di discussione o contatti diretti con il produttore che, aggiornando

il programma, include nuovi moduli o opzioni che risolvono questi problemi. Quindi, chi utilizza il programma farà bene a controllare con sistematicità le caratteristiche delle nuove *releases*, anche utilizzando le tabelle comparative che indicano all'utente che usava versioni precedenti del programma quali siano le modifiche importanti introdotte nella nuova versione.

Allo stesso tempo, però, quest'ultimo esempio dimostra che, nell'attesa che il programma si adegui alle nostre esigenze, non siamo costretti a rinunciare al tentativo di ottenere dal programma quanto ci serve. Gli strumenti di base, magari per strade un po' contorte e con qualche fatica, ci possono spesso aiutare a risolvere i nostri problemi con le variabili senza attendere che, magari in maniera più elegante, Spss risolva i problema per noi.

Si può scoprire, tra l'altro, che, alla fin fine, la soluzione di Spss è spesso la stessa o simile a quella che avevamo noi stessi trovato con la differenza che Spss, traducendola in opzione o funzione, la generalizza, mentre l'utente la applica al caso specifico. È capitato proprio con il problema della distanza temporale tra due eventi per la quale Spss, volendo per l'appunto generalizzare la soluzione a tutte le eventualità, traduce gli orari nell'unità più piccola e cioè in secondi.

Si può chiudere veramente tornando alla giustificazione che abbiamo presentato all'inizio riguardo all'utilità dell'apprendimento delle regole di sintassi e della registrazione delle trasformazioni su un foglio di sintassi. Come abbiamo anticipato, i comandi, e più di tutto le sequenze di comandi qui esemplificati per risolvere un problema specifico, possono diventare facilmente soluzioni generalizzabili, se li abbiamo memorizzati in quanto, a quel punto, basta un semplice adeguamento agli specifici nomi e valori delle variabili presenti in altri contesti. Il lettore è dunque incoraggiato a utilizzare liberamente le esemplificazioni qui esposte, a sperimentarle sui propri dati e a costruirsi una sua piccola "biblioteca" di soluzioni da riprendere e adattare ogniqualvolta un problema simile si ripresenti.

Alcuni dei testi disponibili sul sito: http://web.uniud.it/dest/docenti/dellizotti/Testionl.pdf

Delli Zotti, G. (2006), *Tabelle di mobilità e cambiamenti di opinione. Il caso delle giurie dei cittadini a Torino e Bologna*, Quaderni del Dipartimento di Scienze dell'Uomo, Quad-DSU, 2-2006, Trieste.
Delli Zotti, G. (2005), *Analisi e sintesi di una variabile*, Quaderni del Dipartimento di Scienze dell'Uomo, Quad-DSU, 3-2005, Trieste.
Delli Zotti, G. (2005), *Come creare un indice o una tipologia*, Quaderni del Dipartimento di Scienze dell'Uomo, Quad-DSU, 2-2005, Trieste.
Delli Zotti, G. (2005), *Come "fare a fette" una distribuzione di frequenza*, Quaderni del Dipartimento di Scienze dell'Uomo, Quad-DSU, 1-2005, Trieste.
Delli Zotti, G. (2004), *Le nuove fonti dei dati*, in G. Amendola (a cura di), *Anni in salita. Speranze e paure degli italiani*, Angeli, Milano.
Delli Zotti, G. (1999), *L'analisi esplorativa delle tabelle di contingenza*. Nuova edizione - esempi realizzati con Spss per Windows 7.5, Quaderni del Dipartimento Est, 99-15, Dest, Udine.
Delli Zotti G. (1996), *Il metodo comparato in sociologia*, in A. Gasparini, R. Strassoldo (a cura di), *Tipi ideali e società*, Angeli, Milano.
Delli Zotti G. (1996), *Quale quantità e quanta qualità nella ricerca sociale: tra integrazione e convergenza*, in C. Cipolla, A. De Lillo (cur.), *Il sociologo e le sirene. La sfida dei metodi qualitativi*, Angeli, Milano.
Delli Zotti G. (1992), *Il problema più importante per noi ... Opzioni nella formulazione, codifica ed elaborazione di domande di atteggiamento*, Quaderni dell'Isig. Programma "Metodologia", 92-1, Isig, Gorizia.
Delli Zotti G. (1986), *L'intervista assistita dal computer (C.A.I. - Computer Assisted Interviewing)*, in A. Ardigò, G. Amendola (a cura di), *Ricerca sociologica, informatica e società italiana*, Angeli, Milano.
Delli Zotti G. (1985), *"Tipologia delle matrici utilizzate nella ricerca sociale"*, Rassegna Italiana di Sociologia, XXVI, 2.