Introduzione a SAS

a.a. 2011/12

Analisi multivariata

Carla Rampichini
rampichini@ds.unifi.it

Struttura del sistema SAS

Analisi dei dati ➔ 2 passi fondamentali:

1. organizzazione dei dati;
2. analisi dei dati.

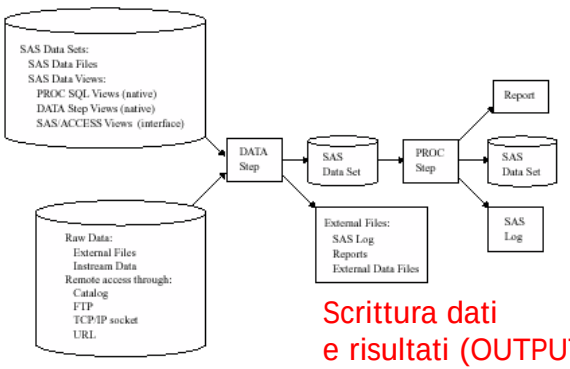
Sistema SAS:

- DATA step: creazione e manipolazione SAS data set;
- PROC step: analisi dei dati attraverso le procedure SAS o procedure utente.

SAS PROCESSING

Lettura dei dati (INPUT):
•SAS data set
•Viste
•Dati grezzi

Figure 2.1 SAS Processing



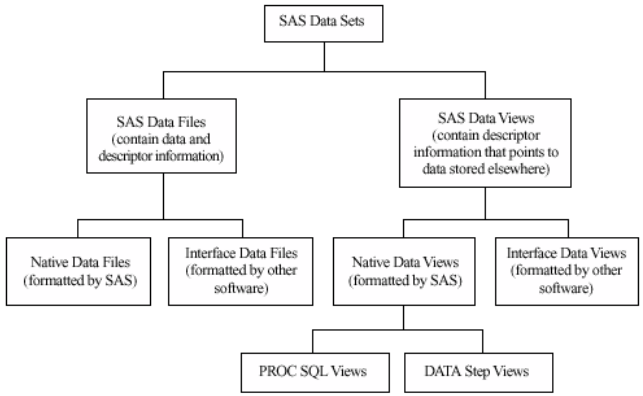
Scrittura dati e risultati (OUTPUT):
•report
•SAS Log
•SAS Data set

Componenti del linguaggio SAS

- **SAS Files**: risiedono in librerie e sono di vario tipo (data sets, catalogs, programmi compilati);
- **SAS Data sets**: data file, “viste”;
- **Files esterni**: dati “grezzi”, istruzioni, output delle procedure;
- **Linguaggio**: istruzioni, espressioni, opzioni, formati, funzioni;
- **Macro facility**: estensione programmi SAS, riduzione istruzioni.

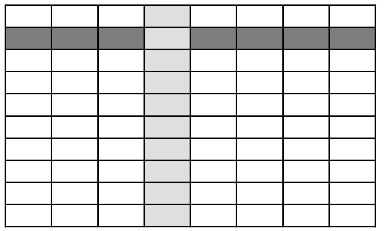
Tipi di SAS data sets

Figure 28.1 Types of SAS Data Sets



SAS Data set

- Files binari generati dal SAS.
- Dati organizzati in *forma rettangolare*



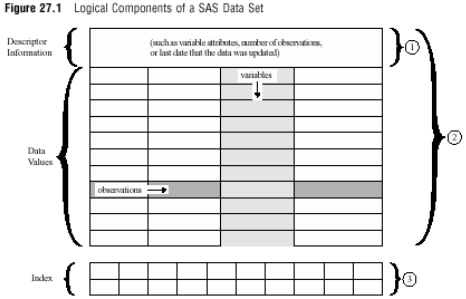
ogni colonna rappresenta una variabile

ogni riga rappresenta una osservazione;

•tutte le osservazioni possiedono le stesse variabili➔le variabili non presenti in una data osservazione sono registrate come mancanti (*missing*).

SAS Data set

Figure 27.1 Logical Components of a SAS Data Set



Componenti logiche del SAS data set:

1. DIA Descriptor Information Area
2. Dati
3. Indici

Ogni SAS data set è auto-descrittivo.

Creazione SAS data set da dati esterni

- Istruzioni SAS nel passo di DATA;
- Funzioni SAS;
- External File Interface (EFI);
- Import Wizard

Lettura di dati da programma

- Istruzione DATALINES (o CARDS)

```
DATALINES;
...
...
;
```

Ogni linea del programma compresa tra DATALINES e ; è considerata una riga di dati.

vedi `arance.sas`, sedutaSAS.pdf

Librerie SAS

libname *libref* '*percorso*';

- Consente di associare il riferimento ad una libreria SAS (*libref*) con il nome fisico di una libreria SAS permanente. Il nome fisico è il *percorso* di una directory.

Data set permanenti

DATA *lib*.*nome*;

...

RUN;

lib è il nome della libreria

nome è il nome dell'archivio

Creazione di SAS Data set da file ASCII

- Istruzione INFILE

```
DATA dataset;
INFILE 'percorso e nome file' [opzioni]
INPUT lista;
....
```

vedi `valuta1.sas`

Osservare la finestra LOG

Esempio lettura a lista da file esterno

Il file ASCII esempio_lista.txt contiene i dati relativi a un'indagine sui laureati dell'ateneo fiorentino a circa 2 anni dalla laurea:

- Sesso (1=maschio, 0=femmina)
- Voto di maturità
- Voto medio agli esami
- Età alla laurea
- Codice corso di laurea
- Lavora stabilmente (1=sì, 0=no)

esempio_lista.sas

Lettura file tipo .xls

Import wizard consente di leggere in maniera quasi automatica file di vario tipo (.csv, delimited, .txt, .xls)

- Dal menu `FILE` scegliere `IMPORT DATA`
- Nel campo data source scegliere `Miscrosoft Excel Workbook`
- Selezionare il file di interesse
- Assegnare un nome al sas data set su cui scrivere
- [creare file contenente istruzioni PROC IMPORT]
- Premere `finish`

Regole per l'importazione di file CSV

(1) I file EXCEL o altri files possono essere salvati in formato CSV (Comma Separated Values)

(2) Assicurarsi che nelle impostazioni della macchina raggiungibili tramite:

Menù Start➔Impostazioni➔Pannello di controllo

Opzioni internazionali➔Numeri

come separatore di elenco sia impostata la virgola [**,**], mentre come separatore decimale sia impostato il punto [**.**].

Lettura file tipo .csv

- Dal menu `FILE` scegliere `IMPORT DATA`
- Nel campo data source scegliere `comma separated values (.csv)`
- Selezionare il file di interesse
- Assegnare un nome al sas data set su cui scrivere
- Premere `finish`

Input a colonna

INPUT *var* [$] *col1*[-*col2*][.*dec*][@|@@];

INPUT nome $ 1-12 peso 13-15 eta 16-17;

- Indicazione colonna iniziale e finale del campo occupato dalla variabile;
- Le variabili devono essere memorizzate nelle stesse posizioni per tutte le osservazioni
- Variabili in forma numerica o carattere standard

Vantaggi input a colonna

- non occorre che i campi corrispondenti alle variabili siano separati da *blanks*;
- non è necessario indicare esplicitamente i valori mancanti (*missing*): un campo vuoto è interpreto come *missing*;
- non si è vincolati a leggere le variabili sequenzialmente:

INPUT nome $ 1-12 eta 16-17 peso 13-15

Istruzioni di assegnazione

variabile=*espressione*;

L'*espressione* è una sequenza di:

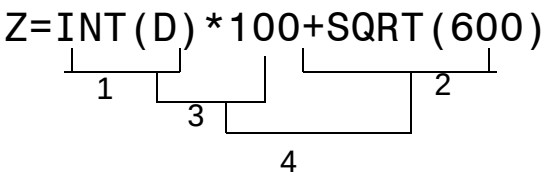
- *operandi* (variabili, costanti)
- *operatori* (caratteri speciali, funzioni, parentesi)

Tipo di operatori

- aritmetici
- di comparazione
- logici
- carattere

Regole ordine di esecuzione

- Funzioni
- espressioni entro parentesi;
- 7 livelli di priorità (1=max)
- se stessa priorità:
  - da dx a sin livello1
  - da sin a dx gli altri



FUNZIONI SAS

FUNZIONE (argomento,...,);

- esempio: SQRT(600) radice quadrata di 600
- Vedere HELP per elenco completo!
- Le funzioni vengono calcolate prima di qualunque altro operatore!

```
Z=INT(D)*100+SQRT(600)
```

1 2 3 4

LAG.SAS

Regole di conversione automatica variabile numerica/carattere

A variabile carattere, X variabile numerica

- Se si usa A con un operatore che richiede operandi numerici Y=A+X, SAS converte A in numerica;
- Se si usa un operatore di comparazione, Y=A<X, la variabile carattere è convertita in numerica;

...segue conversione automatica

- Se si usa X con un operatore che richiede operandi carattere, B=A|X, X è convertita in carattere, formato BEST12. Con allineamento a dx (si può ottenere allineamento a sin con la funzione LEFT);
- Se si usa X alla sinistra di una istr. di assegnazione e A a dx, X=A, A è convertita in numerica;
- Viceversa, A alla sinistra e X a dx, A = X, X è convertita carattere, BEST*n*., *n* lunghezza di A

...segue conversione automatica

- SAS segnala sempre nel LOG la conversione fatta;

```
1    data pippo;                                  Bianchi    1  25
2    input nome $ sesso $ voto;                   Rossi      1  29
3    x=sesso+voto;                                Neri       2  24
4    cards;                                       Verdi      2  28

NOTE: Character values have been converted to numeric values at the places given
    by:      (Line):(Column).      3:3
```

- Se la conversione non è valida mette missing

```
NOTE: Character values have been converted to numeric values at the places given
    by:
      (Line):(Column).
      13:3
NOTE: Invalid numeric data, nome='Bianchi' , at line 13 column 3.
RULE:----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+--
    --8----+----9
15   Bianchi    1  25
nome=Bianchi sesso=1 voto=25 x=. _ERROR_=1 _N_=1
```

CONVERTI.SAS

Controllo scrittura delle osservazioni nel SAS data set

OUTPUT consente di scrivere le osservazioni in un momento qualsiasi del data step;

DELETE consente di eliminare osservazioni

`output1.sas-output3.sas`

SUBSETTING IF: comando IF come selezionatore

IF *espressione*;

- Il SAS data set contiene un sottoinsieme (subset) dei dati letti: tutti e soli i records che rispettano la condizione espressa nel IF statement
  - se espressione VERA (valore NON 0 o missing), SAS include l'osservazione nel data set.
  - se espressione FALSA (valore 0 o missing), nessun altro statement viene eseguito e l'osservazione non viene scritta sul data set creato
- Il subsetting IF è equivalente all'istruzione

  if not (expression) then delete;

Controllo del valore iniziale delle variabili

- Il valore iniziale delle variabili create con INPUT o istr. di assegnazione è missing per default;
- L'istruzione RETAIN consente di assegnare un valore iniziale diverso

```
RETAIN [[var ...[valore]]...;
```

var missing alla prima iterazione del data step, poi mantiene valore iterazione precedente fino a nuova assegnazione

valore assegna a var il valore specificato solo alla prima iterazione (al posto di missing)

retain.sas

Istruzione SUM

variabile+ espressione

- Sostituisce

```
RETAIN VAR 0;
VAR=VAR+ESPRESSIONE;
```

- Non sensibile a propagazione dei missing

`sum.sas retain2.sas`

Trattamento di SAS data set

- SAS data set auto-descrittivo➔il compilatore legge DIA per creare PDV
- Inizializzazione delle variabili nel PDV dipende da MERGE, SET e BY;
- non è possibile definire inizializzazione diversa da quella prevista da SAS

Lettura SAS data set

DATA data_set2;
SET data_set1;

Variabili di data data_set1:

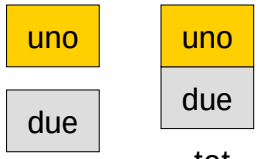
- missing alla prima iterazione;
- Nelle iterazioni successive mantengono il loro valore fino a che non sono disponibili nuovi dati.

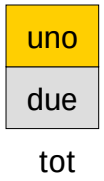
indice.sas

Concatenamento: SET

- uno sull'altro

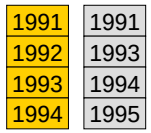
```
Data tot;
set uno due;
run;
```

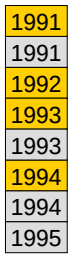




- interleaving

```
Data tot;
set uno due;
by anno;
run;
```





set1.sas, set2.sas

Concatenamento SAS data set

1. 1ª iter: tutte le var missing nel PDV
2. Lettura 1ª obs dal 1° data set
3. elaborazione e scrittura nel nuovo data set
4. Ad ogni iter le var mantengono il valore nel PDV fino a lettura di una nuova obs
5. Lettura nuova obs e torna a 3. fino a fine archivio
6. Pone di nuovo a missing tutte le var nel PDV e inizia lettura 1ª obs dal 2° archivio, ripetendo i passi 3-5

torna p. 111 lucidi SAS

Concatenamento SAS data set

- CONCATENAMENTO UNO A MOLTI (merge2.sas)
- FUSIONE UNO A UNO CON BY E IN= (merge3.sas)
- FUSIONE UNO A MOLTI CON BY E IN= (merge3.sas)
- FUSIONE MOLTI A MOLTI CON BY E IN= (merge3.sas)

lucidi SAS 113-119

Opzione IN=

SET dati1 (IN=var) dati2;

La variabile var consente di individuare il data set da cui provengono i dati presenti nel PDV ad ogni iterazione del data step:

se l'osservazione proviene da dati1 var=1, altrimenti var=0

Ordinamento degli archivi

- Procedura SORT

PROC SORT [DATA=dataset] [OUT=dataset2];
BY [DESCENDING] var1[[DESCENDING] var2];

Esercizio 1: utilizzo di IN=

- Una ditta produttrice di articoli sportivi vuole inviare un depliant pubblicitario a tutti i clienti che non hanno effettuato ordini durante l'anno in corso.
- L'azienda dispone di due archivi di dati, uno contenente l'elenco dei clienti (CLIENTI.DAT) e l'altro l'elenco degli ordini (ORDINI.DAT).
- L'archivio clienti contiene le seguenti variabili: CODICE CLIENTE, NOME, INDIRIZZO
- L'archivio ordini contiene una osservazione per ogni ordine effettuato nell'anno in corso e, per ciascun ordine, le variabili: CODICE CLIENTE, IMPORTO ORDINE

INT: unire i due archivi utilizzando l'opzione IN= e selezionare i clienti che non hanno osservazioni nell'archivio ordini. (soluzione in **clienti.sas**)

Riorganizzazione SAS data set

- Variabili in osservazioni: reshape1.sas
- Valori di sintesi in osservazioni: reshape2.sas
- Valori di sintesi per gruppo in osservazioni: reshape3.sas

Riorganizzazione SAS data set

- Valori di sintesi per gruppo in osservazioni: reshape3.sas
  - sas-data set REGIONI (info per regione)
  - calcola media di alcune variabili per ripartizione con PROC MEANS+BY;
  - Unione archivio originale con valori di sintesi per ripartizione: MERGE+BY

reshape3.sas

Fusione con variabili comuni

Se ci sono valori ripetuti rispetto alle variabili comuni

nel nuovo data set si considera l'osservazione ripetuta in tutti gli archivi

esercizi in merge.pdf, mergedoc.sas

Esercizio 2

- Considerare i due file VARESE_ANAGR.CSV e FRAZIONI_VA_ANAGR.CSV.
- Importare i due file e si crei un archivio in cui si considerano i due archivi concatenati (tramite SET).
- Assegnare le etichette (LABEL) alle variabili "cod_com" (Codice del comune) e a Cod_fam (Codice della famiglia).
- Ordinare i dati in base al numero di componenti (ncomp).
- Stampare i risultati.

Esercizio 3

Importare i files in formato .txt dal nome “Squadre” e “Ruolo”con le seguenti variabili:

- Squadre: Cognome Nome eta anni_gioco cambio_squadre
- Ruolo: Cognome ruolo squadra_attuale
- Unire i due file (fusione) tramite il comando MERGE.
- Costruire una nuova variabile chiamata “mobilita”con i valori seguenti:
  - 1 se le squadre cambiate sono <3
  - 2 se le squadre cambiate sono >=3
- Ordinare le osservazioni sulla base della variabile appena creata e stampare i risultati.

Esercizio 4

Partendo dai dati contenuti nei files COLLEZIONE_INT.CSV e COLLEZIONE_NAZ.CSV in formato CSV:

- Si importino i due files e si crei un unico archivio tramite il comando SET.
- Si stampino le osservazioni separandole per negozio di appartenenza.
- Si importino i due files e si crei un unico archivio tramite il comando MERGE utilizzando la variabile “Negozio” come chiave di unione.
- Si stampino le osservazioni del nuovo archivio.

CICLI

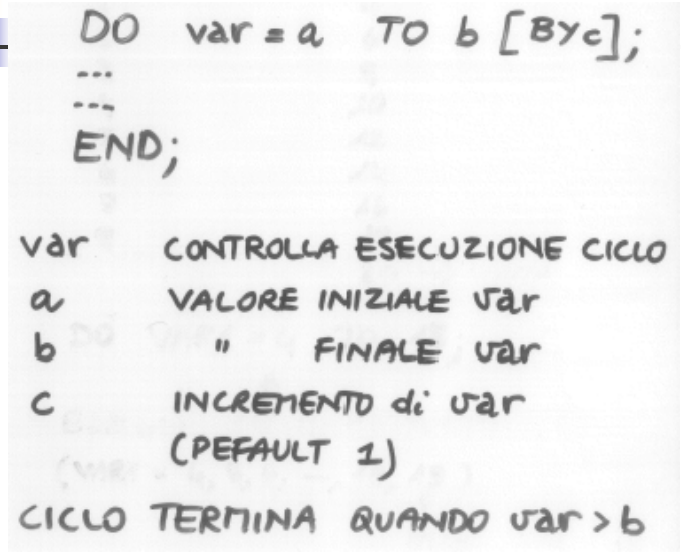
- DO iterativo
- DO WHILE
- DO UNTIL

- DO iterativo

```
DO var=a TO b [BY c];
…;
END;
```

var controlla esecuzione ciclo

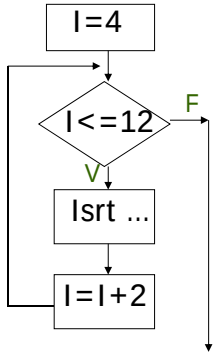
DO ITERATIVO



DO ITERATIVO: esempio

```
DO I=4 TO 12 BY 2;
...;
END;
```

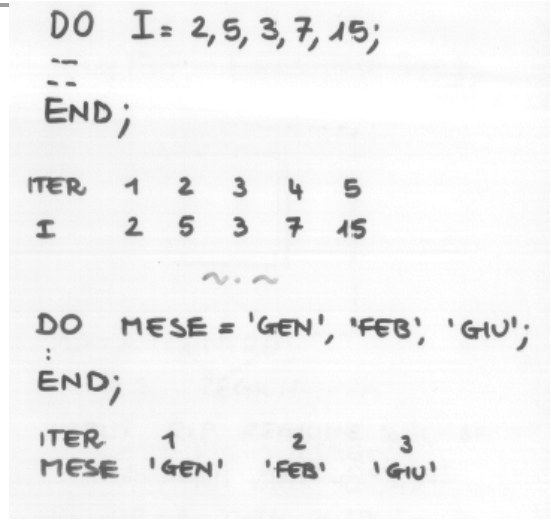
Iter	I
1	4
2	6
3	8
4	10
5	12
----------	*uscita*
	14



DO ITERATIVO con elenco valori

DO var=a,b,c, ...;

La variabile che controlla il ciclo può essere definita da un elenco di valori (numerici o carattere) separati da virgole



GENERAZIONE DEI DATI DA PROGRAMMA

- Utilizzare DO ITERATIVO;
- Utilizzare OUTPUT per aggiungere una osservazione al data set ad ogni iterazione del ciclo.

Esempi:
doiter1.sas
doiter2.sas

Tabella di contingenza

Tab. 1 – Tabella a doppia entrata

Sesso	*Titolo di studio*					**Totale**
	1	2	3	4	5	
Maschi	11	28	11	1	5	56
Femmine	12	30	13	14	13	82
Totale	23	58	24	15	18	138

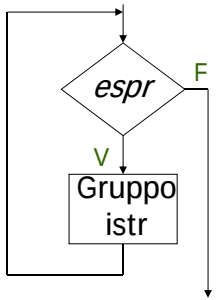
Scrivere un programma SAS per:
- Input da tabella a lista
- Input da tabella a doppia entrata

doiter4.sas

Tab. 2 – Tabella a lista

Sesso	*Titolo di studio*	*Frequenze*
Maschi	1	11
Maschi	2	28
Maschi	3	11
Maschi	4	1
Maschi	5	5
Femmine	1	12
Femmine	2	30
Femmine	3	13
Femmine	4	14
Femmine	5	13

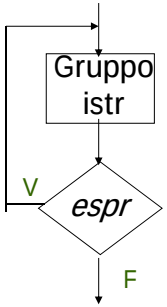
DO WHILE




```
DO WHILE (espr);
...
...
END;
```

- n. Iterazioni del ciclo non noto a priori;
- Se espr falsa a inizio ciclo, il ciclo non viene MAI eseguito!

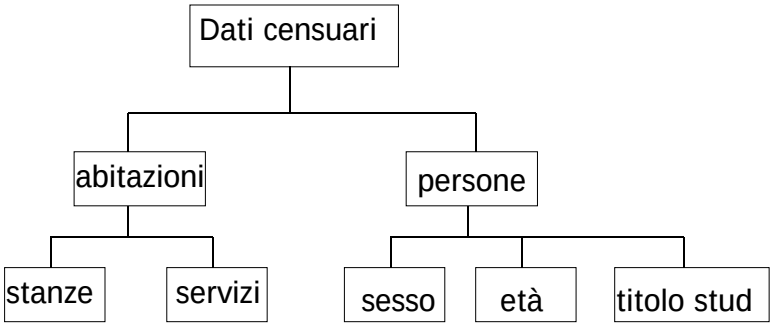
DO UNTIL




```
DO UNTIL (espr);
...
...
END;
```

Le istruzioni del ciclo vengono eseguite almeno una volta

Lettura archivi gerarchici



File gerarchico: caso speciale di file con record misti

censimento2.sas

Lettura file con record a lunghezza variabile

- File originale contiene un record per ogni studente;
- Il numero di esami sostenuto da ogni studente è registrato nel campo NESA;
- Vogliamo leggere voto e codice esame per tutti gli esami sostenuti da ogni studente e creare un record per ogni esame.

input7.sas

Introduzione a SAS

ESERCIZI LETTURA DEI DATI

Esercizio 1: lettura dati da programma

- Scrivere un programma SAS che generi un SAS data set temporaneo di nome DATI e che contenga i dati relativi a anno di nascita, sesso e titolo di studio di un gruppo di 10 studenti (vedi lucido successivo);
- Creare la variabile età, sapendo che i dati sono stati rilevati nel 2003;
- Stampare il contenuto dell'archivio;
- Stampare il contenuto della DIA.

Dati esercizio 1

- Anno di nascita
- Sesso (1=M, 2=F)
- Titolo di studio
  - 1=liceo classico
  - 2=liceo scientifico
  - 3=istituto tecnico
  - 4=istituto professionale
  - 5=altro

1982	2	5
1983	2	3
1983	1	3
1981	2	2
1981	1	2
1983	1	1
1984	2	2
1983	1	2
1983	1	2
1983	1	3
1984	1	5
1983	2	1
1983	1	2
1983	1	3
1976	1	2
1975	1	2
1983	2	3
1981	1	2
1980	2	5

Esercizio 2: lettura a lista da file ASCII

- Creare un archivio permanente SAS dell'archivio in formato testo v:\carla\sas\studenti.txt;
- L'archivio contiene 4 variabili numeriche: leggere solo le prime 3;
- Stampare il contenuto dell'archivio.

Esercizio 3

- Importare i dati contenuti nel file STUDENTI.XLS, salvandoli nell'archivio temporaneo STUDENTI
- Scrivere un passo di DATA che generi la variabile ETA, contenente l'età degli studenti, sapendo che i dati sono stati rilevati nel 2003.

Esercizio 4

- Importare i dati contenuti nel file V:\CARLA\SAS\KAZUAKI.CSV, salvandoli nell'archivio temporaneo KAZUAKI
- Stampare il contenuto della DIA;
- Stampare i dati.

Esercizio 5: lettura a colonna

- Punteggi assegnati da 5 degustatori ai biscotti della ditta X. Il file contiene: nome prodotto anno di inizio produzione, tipo (cioccolato si/no), data del giudizio, punteggi.
- Creare un archivio SAS permanente a partire dai dati contenuti nel file ASCII 'v:\didat\carla\sas\biscotti.txt'.
- Creare una variabile che contiene, per ciascuna osservazione, la media dei punteggi.

Variabili e posizione nel file v:\carla\sas\biscotti.txt

- `nome    1-13 nome dei biscotti`
- `anno    14-15 anno inizio produzione`
- `tipo    16-17 al cioccolato (sì o no)`
- `data    18-27 giorno della rilevazione`
- `punti1  29-31 punteggio 1° assaggiatore`
- `punti2  33-35 punteggio 2° assaggiatore`
- `punti3  37-39 punteggio 3° assaggiatore`
- `punti4  41-43 punteggio 4° assaggiatore`
- `punti5  45-47 punteggio 5° assaggiatore`