SAS
STATISTICAL ANALYSIS SYSTEM

Parte II: procedure per l'analisi statistica dei dati

Analisi dei dati: procedure preliminari

- operazioni di spoglio e controllo dei dati
- operazioni di ponderazioni dei dati

Operazioni di spoglio e controllo dei dati:
la Proc FREQ per analisi preliminari

L'istruzione PROC FREQ, è l'unica istruzione obbligatoria per utilizzare la procedura; se si specificano le istruzioni che seguono, la procedura da luogo a una tabella di frequenza semplice, per ogni variabile presente nel data set creato più di recente:

```
proc freq;
run;
```

Se si vuole analizzare uno specifico data set, è necessario utilizzare l'opzione data dell'istruzione proc freq:

```
proc freq data=library.pippo;
run;
```

Se invece si desidera analizzare solo alcune variabili è necessario utilizzare l'istruzione tables, come segue:

```
proc freq;
tables var1 var2 ….;
run;
```

Sintassi:

PROC FREQ <option(s)>;
TABLES request(s) </ option(s)>;
OUTPUT statistic-keyword(s) <OUT=SAS-data-set>;
TEST statistic-keyword(s);
FORMAT;
WEIGHT variable;

RUN;

Options TABLES:

- NOFREQ
- NOPERCENT
- NOCUM
- NOROW
- NOCOL
- MISSING

Esempio

L'esempio che segue riporta il codice della *proc freq* eseguita su 3 variabili del data set sui consumi delle famiglie:

```
data prova;set istat_md.consumi2;
proc freq data=prova;
tables regione numcomp sesso1;
run;
```

Segue l'output di default per la prima variabile analizzata, con le seguenti caratteristiche:

- nome della variabile e modalità;
- frequenze assolute;
- frequenze percentuali;
- frequenze cumulate assolute;
- frequenze cumulate percentuali;

-
- segnalazione dati mancanti “missing”, non conteggiati tra le frequenze

The FREQ Procedure

regione

regione	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2072	9.26	2072	9.26
3	2468	11.03	4540	20.30
4	1104	4.94	5644	25.23
5	1513	6.76	7157	32.00
6	643	2.87	7800	34.87
7	806	3.60	8606	38.47
8	1239	5.54	9845	44.01
9	1284	5.74	11129	49.75
10	610	2.73	11739	52.48
11	713	3.19	12452	55.67
12	1707	7.63	14159	63.30
13	811	3.63	14970	66.93
14	512	2.29	15482	69.21
15	1721	7.69	17203	76.91
16	1430	6.39	18633	83.30
17	430	1.92	19063	85.22
18	911	4.07	19974	89.30
19	1675	7.49	21649	96.79
20	719	3.21	22368	100.00

Frequency Missing = 1550

Analisi dei dati: procedure fondamentali

- ☞ Tabelle di frequenza e di contingenza
- ☞ Sintesi delle distribuzioni semplici e multivariate

Tabelle di frequenza e di contingenza

Istruzioni base

Come già introdotto nella sezione precedente, la procedura PROC FREQ consente di produrre tabelle di frequenza semplici, ma anche tabelle a doppia entrata, a tre variabili e così via fino a tabelle a n-entrate. Lo schema che segue riassume la sintassi di base della procedura.

```
PROC FREQ <option(s)>;

    TABLES request(s) </ option(s)>;
    OUTPUT statistic-keyword(s) <OUT=SAS-data-set>;
    TEST statistic-keyword(s);
    FORMAT;
    WEIGHT variable;
RUN;
```

Con l'istruzione TABLES si selezionano la\le variabili per le tabelle di frequenza semplici o multiple, nel secondo caso le variabili coinvolte sono incrociate con l'operatore "*".

Creazione di un file SAS di output

- Attraverso l'opzione out dell'istruzione table, si possono salvare le frequenze (assolute e percentuali) di output su un file tipo data set.

```
proc freq data=prova;
tables sesso1*regione/out=tabella;
run;
```

Sintesi delle distribuzioni semplici e multivariate

In questa sezione si analizzano le principali procedure che producono sintesi delle distribuzioni:

- **PROC MEANS**
- **PROC UNIVARIATE**

PROC MEANS

La Proc Means fornisce gli strumenti di sintesi per computare statistiche descrittive per variabili numeriche; le elaborazioni possono essere compiute su tutte le osservazioni oppure per gruppi di osservazioni omogenee classificate rispetto a una o più variabili. Per esempio la Proc Means:

- calcola statistiche descrittive basate sui momenti;
- calcola la mediana e tutti gli altri quantili;
- identifica i valori estremi;
- calcola l'intervallo di confidenza per la media.

PROC MEANS <*option(s)*> <*statistic-keyword(s)*>;
BY <DESCENDING> *variable-1* <... <DESCENDING> *variable-n*><NOTSORTED>;
CLASS *variable(s)* </ *option(s)*>;
OUTPUT <OUT=*SAS-data-set*> <*output-statistic-specification(s)*>
<*id-group-specification(s)*> <*maximum-id-specification(s)*>
<*minimum-id-specification(s)*> </ *option(s)*> ;
VAR *variable(s)* < / WEIGHT=*weight-variable*>;
WEIGHT *variable*;
RUN;

Procedura base

Il codice riportato nel riquadro mostra una procedura di tipo Means di base, applicata alle variabili relative alla spesa per elettricità e all'uso del telefono, presenti nel data set sui consumi delle famiglie.

```
data prova;set istat_md.consumi;
label c_elettr='consumi elettricità' c_tel='consumi telefonici';

proc means;
var c_elettr c_tel;
run;
```

Nel riquadro successivo è riportato l'output stampato nella relativa finestra.

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
c_elettr	consumi elettricità	23877	65440.41	51613.88	6000.00	750000.00
c_tel	consumi telefonici	22150	75583.69	50734.91	10000.00	1000000.00

Commento all'output:
Notare che la terza colonna N riporta la numerosità delle osservazioni sulle quali sono state calcolate le misure. In entrambi i casi non si ritrova la numerosità totale del campione, pari a 23918 unità; questo perché la procedura effettua le elaborazioni escludendo i dati mancanti "missing".

Istruzione proc means: statistiche opzionali

Le statistiche di default della procedura, come si può vedere negli esempi di output, sono la media semplice, la deviazione standard, il valore minimo e il valore massimo. Altre statistiche possono essere richieste opzionalmente utilizzando specifiche parole-chiave. La tabella seguente riporta una selezione delle parole chiave per le principali statistiche descrittive:

Statistiche descrittive		**Indici di posizione**	
Parola-chiave	*Indice*	*Parola-chiave*	*Indice*
CSS	Devianza	MEDIAN|P50	Mediana
CV	Coeff. variazione	P1	1° percentile
KURTOSIS|KURT	Indice di Curtosi	P5	5° percentile
MAX	Valore massimo	P10	10° percentile
MEAN	Media aritmetica	Q1|P25	1° quartile
MIN	Valore minimo	Q3|P75	3° quartile
N	Num. osser. valide	P90	90° percentile
NMISS	Num. osser.manc.	P95	95° percentile
RANGE	Interv. di variaz.	P99	99° percentile
SKEWNESS|SKEW	Indice di simmetr.	QRANGE	Dist. interquartile
STDDEV|STD	Dev. Standard		
SUM	Somma dei valori		
SUMWGT	Somma dei pesi		
USS	Somma val. quadr.		
VAR	Varianza		

Esempio

Si vogliono analizzare ancora una volta la spesa per i consumi di elettricità e per quelli telefonici. Stavolta però si vuole produrre un output che contenga il numero di osservazioni valide, la somma dei valori di spesa individuali, la media aritmetica e il coefficiente di variazione.
I riquadri che seguono riportano il codice predisposto e il relativo output.

```
proc means data=prova N sum mean cv;
var c_elettr c_tel;
run;
```

The MEANS Procedure

Variable	Label	N	Sum	Mean	Coeff of Variation
c_elettr	consumi elettricità	23877	1562520749	65440.41	78.8715644
c_tel	consumi telefonici	22150	1674178756	75583.69	67.1241510

Analisi per gruppi di osservazioni omogenee

i. Istruzione BY

Produce le statistiche relative alla variabile d'analisi, separatamente per ogni modalità della variabile di classificazione (una o più) specificata con l'istruzione BY.

Per eseguire l'istruzione BY è necessario, prima di eseguire la proc, ordinare il data set in maniera crescente secondo la variabile di classificazione.

Segue un programma esemplificativo e l'output relativo.

```
proc sort data=prova;
by sesso1;
proc means data=prova N sum mean cv;
var c_elettr c_tel;
by sesso1;
run;
```

```
---------------------------------------- genere=. ----------------------------------------

                                                                              Coeff of
Variable    Label                          N            Sum            Mean  Variation
------------------------------------------------------------------------------------------
c_elettr    consumi elettricità          425     28239582.00        66446.08  101.3807219
c_tel       consumi telefonici           382     28674750.00        75064.79   66.7099223
------------------------------------------------------------------------------------------

---------------------------------------- genere=M ----------------------------------------

                                                                              Coeff of
Variable    Label                          N            Sum            Mean  Variation
------------------------------------------------------------------------------------------
c_elettr    consumi elettricità        17644      1237280015        70124.69   74.6000230
c_tel       consumi telefonici         16603      1305037762        78602.53   66.0839965
------------------------------------------------------------------------------------------

---------------------------------------- genere=F ----------------------------------------

                                                                              Coeff of
Variable    Label                          N            Sum            Mean  Variation
------------------------------------------------------------------------------------------
c_elettr    consumi elettricità         5808       297001152        51136.56   88.2372498
c_tel       consumi telefonici          5165       340466244        65917.96   68.8230171
------------------------------------------------------------------------------------------
```

ii. Istruzione CLASS

Le statistiche relative a gruppi di osservazioni omogenee, originate da una variabile di classificazione (una o più), sono prodotte con un'unica tavola di output.

Segue un programma esemplificativo e l'output relativo.

```
proc means data=prova1 N sum mean cv;
var c_elettr c_tel;
class sesso1;
run;
```

The MEANS Procedure

sesso1	N Obs	Variable	Label	N	Sum	Mean	Coeff of Variation
M	17669	c_elettr	c_elettr	17644	1237280015	70124.69	74.6000230
		c_tel	c_tel	16603	1305037762	78602.53	66.0839965
F	5819	c_elettr	c_elettr	5808	297001152	51136.56	88.2372498
		c_tel	c_tel	5165	340466244	65917.96	68.8230171

iii. Più variabili di classificazione

Le istruzioni by e class possono essere applicate con più variabili di classificazione contemporaneamente; i nomi delle variabili devono seguire l'istruzione e devono essere separate da spazi.

PROC UNIVARIATE

La Proc Univariate produce statistiche descrittive semplici per variabili numeriche. Per esempio la Proc univariate:

- calcola statistiche descrittive basate sui momenti;
- calcola la mediana, la moda, il *range* e i quantili;
- crea il plot della distribuzione di una variabile;
- realizza plot di normalità;
- crea data set di output contenente gli indici richiesti.

Sintassi di base:

PROC UNIVARIATE <*option(s)*>;
BY <DESCENDING> *variable-1* <...<DESCENDING> *variable-n*><NOTSORTED>;
CLASS *variable-1*<(*variable-option(s)*)> <*variable-2*<(*variable-option(s)*)>>
</ **KEYLEVEL=**'*value1*'|('*value1*' '*value2*')>;
OUTPUT <OUT=*SAS-data-set*>*statistic-keyword-1=name(s)*<... *statistic-keyword-n=name(s)*> <*percentiles-specification*>;
ID *variable;*
VAR *variable(s)*;
WEIGHT *variable*;
RUN;

Procedura base

Il codice riportato nel riquadro mostra una procedura di tipo Univariate di base, applicata alla variabili relative alla spesa per acqua, presente nel data set sui consumi delle famiglie.

```
proc univariate data=prova;
var c_acqua;
run;
```

Nei riquadri che seguono sono state riportate le due pagine di output che la procedura produce di default.

La procedura UNIVARIATE
Variabile: c_acqua (c_acqua)

Momenti

N	14569	Somma dei pesi	14569
Media	35606.5039	Somma delle osservazioni	518751155
Deviazione std	33261.5719	Varianza	1106332168
Skewness	2.95468355	Kurtosis	13.5039107
SS non corretta	3.4588E13	SS corretta	1.6117E13
Coeff variaz	93.4143157	Errore std media	275.567438

Misure statistiche di base

Posizione		Variabilità	
Media	35606.50	Deviazione std	33262
Mediana	25067.00	Varianza	1106332168
Moda	16667.00	Intervallo	330000
		Intervallo interquartile	28333

Test di posizione: Mu0=0

Test	-Statistica-		-----Valore p-----
T di Student	t	129.2116	Pr > |t| <.0001
dei segni	M	7284.5	Pr >= |M| <.0001
dei segni per ranghi	S	53067583	Pr >= |S| <.0001

Quantili (Definizione 5)

Quantile	Stima
100% Max	333333
99%	166667
95%	100000
90%	70667
75% Q3	43333
50% Mediana	25067
25% Q1	15000
10%	10000
5%	7333
1%	4667
0% Min	3333

La procedura UNIVARIATE
Variabile: c_acqua (c_acqua)

Osservazioni estreme

---Inferiori---		----Superiori---	
Valore	Oss	Valore	Oss
3333	23038	333333	5732
3333	22279	333333	8794
3333	21910	333333	13862
3333	21757	333333	20399
3333	21665	333333	21410

Valori mancanti

Valore mancante	Conteggio	---Percentuale di--- Tutte le oss	Oss mancanti
.	9349	39.09	100.00

Contenuto dell'output

All'interno dell'output di default si possono distinguere 6 gruppi di statistiche (alcune ricorrono in più di un gruppo):

a) *moments, statistiche basate sui momenti;*
b) *misure statistiche di base;*
c) *test di locazione (μ_0=0 sotto H_0);*
d) *quantili;*
e) *valori estremi;*
f) *missing.*

Come si può notare le statistiche computate sono varie e numerose, l'output stampato è quindi molto ingombrante. Se si desidera un output più compatto è consigliabile dare l'opzione noprint (all'istruzione iniziale) e creare un data set di output con le statistiche desiderate (vedere più avanti la sezione Produzione di data-set di output*).*

Produzione di data set di output

Le istruzioni per produrre un data set di output con la Proc Univariate, risultano simili a quelle già viste in precedenza con la Proc Means (pag. 51). Anche le parole chiave da utilizzare per selezionare gli indici da inserire nel file di output sono le stesse esaminate per la precedente procedura (per visualizzare ulteriori indici e relative parole chiave si consulti l'help del programma SAS).

Segue il programma completo di istruzione output e la stampa della finestra di output.

```
proc univariate data=prova noprint;
var eta1;
class sesso1;
output out=indici mode=modeta  q1=q1eta median=medeta
          q3=q3eta;
format sesso1 sexfmt.;
run;
proc print;
run;
```

Obs	sesso1	q3eta	medeta	q1eta	modeta
1	M	64	52	42	55
2	F	75	65	48	71

Si noti che l'output in finestra è stato stampato solo perché il programma termina con l'istruzione proc print; in caso contrario l'opzione noprint, presente nell'istruzione principale, agisce sopprimendo la stampa dell'output nella relativa finestra.

Produzione di data set di output per il calcolo dei percentili

Come già accennato in precedenza, la procedura Univariate può produrre una vasta gamma di statistiche opzionali (si veda ancora l'elenco delle parole chiave a pagina 51) attraverso l'istruzione output, contenute ovviamente in un file di output. I percentili disponibili attraverso parole chiave sono il 1°, 5°, 10°, 90°, 95° e il 99° percentile. Se si desiderano degli altri percentili bisogna opportunamente programmare l'istruzione output attraverso le seguenti opzioni:

pctlpts= indica quali percentili devono essere calcolati (esempio il 20° e il 40°);
pctlpre= indica il prefisso da utilizzare per la denominazione dei percentili nel file di output;
pctlname= indica il nome da utilizzare per la denominazione dei percentili nel file di output.

Segue il programma completo di istruzione output e la stampa della finestra di output.

```
proc univariate data=prova noprint;
      var eta1 c_elettr;
      output out=Perc pctlpts=20 40 pctlpre=eta elettr
      pctlname=P20 P40;run;
proc print;
   run;
```

Analisi dell'associazione tra variabili

- ☞ studio dell'associazione
- ☞ variabile risposta categorica

Studio dell'associazione tra variabili:

- PROC CORR
- PROC REG

PROC CORR

La Proc Corr computa il coefficiente di correlazione di Pearson, tre misure di associazione non parametrica (opzionali) e le probabilità associate a queste statistiche.

Sintassi di base

PROC CORR <*option(s)*>;

BY <DESCENDING> *variable-1*<...<DESCENDING> *variable-n*> <NOTSORTED>;

PARTIAL *variable(s)*;

VAR *variable(s)*;

WEIGHT *weight-variable*;

RUN;

Procedura base

Nella procedura base è sufficiente, oltre all'istruzione obbligatoria proc corr, elencare le variabili (separate da spazi) sulle quali calcolare il coefficiente di correlazione lineare mediante l'istruzione var. Il programma che segue produce la matrice di correlazione per alcune variabili di spesa del data set sui consumi delle famiglie.

```
data prova;set istat_md.consumi2;run;
proc corr;
var c_elettr c_acqua c_gas_r numcomp eta1 stanze;
run;
```

Lettura dell'output

La prima parte dell'output (si veda il riquadro seguente) riporta le principali statistiche descrittive per ogni variabile. Segue la vera e propria matrice di correlazione sulla quale, per ogni incrocio, sono stampati: il coefficiente di correlazione di Pearson, la probabilità che r=0 sotto ipotesi nulla e il numero di osservazioni valide.

```
                                The CORR Procedure

             6  Variables:    c_elettr c_acqua  c_gas_r  numcomp  eta1     stanze


                                Simple Statistics

Variable          N         Mean      Std Dev          Sum      Minimum      Maximum   Label

c_elettr      23877        65440        51614   1562520749         6000       750000   c_elettr
c_acqua       14569        35607        33262    518751155         3333       333333   c_acqua
c_gas_r       15749       106400       110978   1675691786         5000      1000000   c_gas_r
numcomp       23918      2.72719      1.30323        65229      1.00000     10.00000   numcomp
eta1          23751     55.42356     15.58377      1316365     16.00000    101.00000   eta1
stanze        23877      4.40524      1.47062       105184      1.00000     20.00000   stanze


                           Pearson Correlation Coefficients
                              Prob > |r| under H0: Rho=0
                                 Number of Observations

               c_elettr       c_acqua       c_gas_r       numcomp          eta1        stanze
```

c_elettr	1.00000	0.21560	0.24634	0.34204	-0.10554	0.26246
c_elettr		<.0001	<.0001	<.0001	<.0001	<.0001
	23877	14569	15749	23877	23710	23877
c_acqua	0.21560	1.00000	0.14119	0.24319	-0.04299	0.15420
c_acqua	<.0001		<.0001	<.0001	<.0001	<.0001
	14569	14569	10472	14569	14454	14569
c_gas_r	0.24634	0.14119	1.00000	0.13848	-0.00203	0.25092
c_gas_r	<.0001	<.0001		<.0001	0.7996	<.0001
	15749	10472	15749	15749	15656	15749
numcomp	0.34204	0.24319	0.13848	1.00000	-0.33261	0.29005
numcomp	<.0001	<.0001	<.0001		<.0001	<.0001
	23877	14569	15749	23918	23751	23877
eta1	-0.10554	-0.04299	-0.00203	-0.33261	1.00000	-0.02447
eta1	<.0001	<.0001	0.7996	<.0001		0.0002
	23710	14454	15656	23751	23751	23710
stanze	0.26246	0.15420	0.25092	0.29005	-0.02447	1.00000
stanze	<.0001	<.0001	<.0001	<.0001	0.0002	
	23877	14569	15749	23877	23710	23877

PROC REG

La procedura REG è uno dei numerosi strumenti per l'analisi della regressione disponibili nel sistema SAS. La Proc Reg si propone obiettivi piuttosto generici, mentre altre procedure finalizzate alla regressione forniscono applicazioni più specialistiche.

Sintassi di base

PROC REG < options > **;**
< label: > **MODEL** dependents=<regressors> < / options > **;**
BY variables **;**
WEIGHT variable **;**
OUTPUT < **OUT=**SAS-data-set > keyword=names
< ... keyword=names > **;**
PLOT <yvariable*xvariable> <=symbol>
< ...yvariable*xvariable> <=symbol> < / options > **;**
PRINT < options > < **ANOVA** > < **MODELDATA** > **;**
RUN;

Procedura base

Il programma che segue esegue una Proc Reg di base che analizza la regressione lineare tra la spesa per il consumo di elettricità e la numerosità familiare.

```
proc reg data=prova;
label numcomp='numerosità familiare';
model c_elettr=numcomp;
run;
```

Opzionalmente può essere inserita l'istruzione plot che produce un diagramma scatter con retta di regressione e relativa equazione. Segue il programma completo di istruzione plot.

```
proc reg data=prova;
model c_elettr=numcomp;
plot c_elettr*numcomp;
run;
```

Lettura dell'output

Nel riquadro seguente è riportato l'output della Proc Reg di default. Come si può osservare la prima parte è dedicata all'analisi della varianza mentre nella seconda sono riportate le stime dei parametri con relativo test t.

La procedura REG
Modello: MODEL1
Variabile dipendente: c_elettr consumi elettricità

Numero di osservazioni lette	23918
Numero di osservazioni usate	23877
Numero di osservazioni con valori mancanti	41

Analisi della varianza

Origine	DF	Somma dei quadrati	Media dei quadrati	Valore F	Pr > F
Modello	1	7.441194E12	7.441194E12	3163.19	<.0001
Errore	23875	5.616429E13	2352430897		
Totale corretto	23876	6.360548E13			

Radice MSE	48502	R-quadro	0.1170
Media dipendente	65440	R-quadr corr	0.1170
Coeff var	74.11607		

Stime dei parametri

Variabile	Etichetta	DF	Stima dei parametri	Errore standard	Valore t	Pr > |t|
Interc	Interc	1	28491	728.09347	39.13	<.0001
numcomp	numerosità familiare	1	13546	240.84890	56.24	<.0001

```
proc reg data=prova;
label c_elettr='consumi elettricità' c_tel='consumi telefonici';
model c_elettr=c_tel;
plot c_elettr*c_tel;
run;
```