**Premessa**

Il SAS, acronimo per Statistical Analysis System, è uno dei software più utilizzati per condurre analisi che vanno da semplici statistiche descrittive e inferenziali alle più complesse tecniche multivariate. La sua popolarità è imputabile innanzitutto alla possibilità di poter essere utilizzato su una estesa gamma di ambienti differenti, dal grande calcolatore al personal computer, ma soprattutto alla sua capacità di gestire grossi volumi di dati utilizzando procedure già implementate, disponibili in numerosi **moduli** che rispondono a esigenze diverse.

Esiste un modulo generale, denominato BASE, che consente di effettuare analisi statistiche elementari, a cui si affiancano vari moduli per applicazioni più specifiche quali ad esempio:

- STAT per applicare metodologie statistiche complesse
- QC per il controllo qualità
- OR per la ricerca operativa
- TSA per l'analisi di serie temporali
- IML per la manipolazione di matrici
- GRAPH per applicare tecniche grafiche complesse

In ciascun modulo il SAS mette a disposizione un elevato numero di programmi preconfezionati (denominati **procedure** ed abbreviati con PROC) in grado di condurre tutte le analisi statistiche comunemente utilizzate nei diversi ambiti.

È inoltre possibile programmare personalmente le analisi che si desidera effettuare mediante un linguaggio di programmazione di alto livello

# 1. ANALISI STATISTICHE MULTIVARIATE

## 1.1 Fonti e tipologie di dati

Le più comuni fonti dei dati da analizzare con i metodi statistici multivariati sono

- questionari provenienti da rilevazioni dirette
- dati statistici raccolti da enti di varia natura
- risultati di osservazioni sperimentali

I dati da elaborare sono spesso di tipo eterogeneo, nel senso che possono coesistere variabili qualitative (sia sconnesse, sia ordinabili) e variabili quantitative (sia discrete, sia continue), anche se in alcune circostanze si può avere a che fare con variabili solo qualitative oppure solo quantitative.

In generale i dati possono essere organizzati in modi diversi, come nel caso della tabella 1.1, di dimensioni $n \times h$, dove le $n$ righe rappresentano le unità statistiche considerate mentre le $h$ colonne rappresentano le variabili oggetto di rilevazione.

Tabella 1.1

Esempio di matrice dei dati

| Unità statistiche\ Variabili | $X_1$ | $X_2$ | ... | $X_i$ | ... | $X_h$ |
|---|---|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{21}$ | ... | $x_{i1}$ | ... | $x_{h1}$ |
| 2 | $x_{12}$ | $x_{22}$ | ... | $x_{i2}$ | ... | $x_{h2}$ |
| . | ... | ... | ... | ... | ... | ... |
| $n$ | $x_{1n}$ | $x_{2n}$ | ... | $x_{in}$ | ... | $x_{hn}$ |

Se le variabili sono qualitative, i dati raccolti vengono spesso organizzati in tavole di contingenza multiple di cui, nella successiva tabella 1.2, è riportato un esempio relativo a due sole variabili X e Y

Tabella 1.2

Esempio di tabella di contingenza

| X\Y | $y_1$ | $y_2$ | ... | $y_l$ | ... | $y_h$ | |
|---|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1l}$ | ... | $n_{1h}$ | $n_{1.}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2l}$ | ... | $n_{2h}$ | $n_{2.}$ |
| . | . | . | . | . | . | . | . |
| $x_j$ | $n_{j1}$ | $n_{j2}$ | ... | $n_{jl}$ | ... | $n_{jh}$ | $n_{j.}$ |
| . | . | . | . | . | . | . | . |
| $x_k$ | $n_{k1}$ | $n_{k2}$ | ... | $n_{kl}$ | .... | $n_{kh}$ | $n_{k.}$ |
| | $n_{.1}$ | $n_{.2}$ | ... | $n_{.l}$ | ... | $n_{.h}$ | $n$ |

In altri casi, infine, i dati da elaborare sono organizzati in una tabella di correlazione o di covarianza

Tabella 1.3

Esempio di matrice di covarianza

| X\Y | $X_1$ | $X_2$ | ... | $X_j$ | ... | $X_h$ |
|---|---|---|---|---|---|---|
| $X_1$ | $s_1^2$ | $s_{12}$ | ... | $s_{1j}$ | ... | $s_{1h}$ |
| $X_2$ | $s_{21}$ | $s_2^2$ | ... | $s_{2j}$ | ... | $s_{2h}$ |
| . | . | . | . | . | . | . |
| $X_j$ | $s_{j1}$ | $s_{j2}$ | ... | $s_j^2$ | ... | $s_{jh}$ |
| . | . | . | . | . | . | . |
| $X_h$ | $s_{h1}$ | $s_{h2}$ | ... | $s_{hj}$ | .... | $s_h^2$ |

Tabella 1.4

Esempio di matrice di correlazione

| X\Y | $X_1$ | $X_2$ | ... | $X_j$ | ... | $X_h$ |
|---|---|---|---|---|---|---|
| $X_1$ | 1 | $r_{12}$ | ... | $r_{1j}$ | ... | $r_{1h}$ |
| $X_2$ | $r_{21}$ | 1 | ... | $r_{2j}$ | ... | $r_{2h}$ |
| . | . | . | . | . | . | . |
| $X_j$ | $r_{j1}$ | $r_{j2}$ | ... | 1 | ... | $r_{jh}$ |
| . | . | . | . | . | . | . |
| $X_h$ | $r_{h1}$ | $r_{h2}$ | ... | $r_{hj}$ | .... | 1 |

Nelle pagine seguenti verranno descritti metodi statistici diversi che variano a seconda

- della natura delle variabili
- degli obiettivi dell'indagine

### 1.2 Tipi di analisi

Gli obiettivi di un'analisi statistica multivariata consistono generalmente:

- nello studio delle relazioni e dei legami esistenti fra le variabili rilevate
- nel tentativo di sintetizzare le informazioni, ossia di semplificare la struttura dei dati riducendo il numero di variabili considerate in modo da riuscire a interpretare più facilmente le caratteristiche del fenomeno oggetto di indagine
- nell'individuazione di eventuali sottogruppi di unità statistiche che possano essere ritenuti omogenei fra di loro
- nella formulazione e verifica di ipotesi operative

Nel seguito verranno considerate alcune delle più comuni procedure SAS che consentono di effettuare:

**1)** l'**analisi di regressione (Proc REG)**. In particolare, verranno esaminate le istruzioni che consentono di

- verificare le ipotesi sottostanti il modello
- valutare il grado di collinearità fra regressori
- individuare eventuali outliers

**2)** l'**analisi della varianza**, o **Anova (Proc GLM)**. Questo tipo di analisi viene effettuata quando si vuole verificare se i valori di una o più variabili quantitative (dette **variabili dipendenti**) risentono o meno delle determinazioni assunte da una o più variabili qualitative o quantitative (dette **fattori**).

Per effettuare questa valutazione si confrontano i valori medi di una o più variabili dipendenti all'interno di gruppi omogenei rispetto ai fattori considerati per verificare se tali valori medi risultano significativamente diversi fra loro oppure se le differenze rilevate sono da imputarsi all'effetto dei soli fattori casuali (ad esempio si può verificare se la resa produttiva di diversi fertilizzanti è significativamente diversa, se la durata di conservazione di un alimento varia in relazione al materiale utilizzato per l'imballaggio, alla temperatura utilizzata nella sua produzione, al grado di umidità).

Nell'analisi della varianza univariata si ha una sola variabile dipendente e una o più variabili esplicative, mentre in quella multivariata le variabili dipendenti sono più di una per cui si confronteranno non singoli valori medi, ma vettori di medie. Come nel caso univariato, ci possono essere uno o più fattori, sia di natura qualitativa sia quantitativa (per esempio si potrebbe voler verificare se due diversi tipi di messaggio pubblicitario abbiano causato effetti differenti significativi su una serie di aspetti quali il giudizio complessivo sul prodotto, la credibilità del messaggio, l'interesse suscitato).

**3)** l'**analisi in componenti principali**, o **ACP (Proc PRINCOMP)**. Lo scopo di una ACP consiste nel ridurre il numero delle *h* variabili quantitative rilevate sulle *n* unità statistiche minimizzando la perdita di informazioni che ne consegue.

Considerate *h* variabili, per avere informazioni su di esse e sulle loro relazioni si potrebbero calcolare *h* valori medi, *h* varianze e $h(h-1)/2$ covarianze. Il numero di indici necessari cresce al crescere di *h*.

Nelle situazioni reali le *h* variabili sono sempre più o meno correlate fra loro, ma un insieme di *h* variabili incorrelate può essere sempre ottenuto utilizzando opportune trasformazioni lineari delle variabili originarie.

Nell'ACP viene effettuata una sintesi delle informazioni utilizzando non tutte le *h* variabili incorrelate ottenute mediante trasformazione lineare (le cosiddette **componenti principali**), ma un numero $k < h$, con l'obiettivo di descrivere in modo semplificato la struttura dei dati senza perdere però troppe informazioni.

Lo scopo di questo tipo di analisi consiste quindi nel semplificare la descrizione di un fenomeno *h*-dimensionale tramite un numero inferiore di variabili incorrelate, ottenute mediante trasformazioni lineari delle variabili rilevate che siano incorrelate tra di loro, ma che contengano al loro interno una quantità soddisfacente delle informazioni originarie.

**4)** l'**analisi fattoriale**, o **AF (Proc FACTOR**). Questo metodo viene utilizzato quando si vogliono analizzare le interrelazioni esistenti fra un gran numero di variabili quantitative e spiegare queste interrelazioni in termini di **fattori comuni** sottostanti che non sono direttamente osservabili. In pratica, analizzando il comportamento delle variabili rilevate, si ipotizza che quelle che variano allo stesso modo siano legate fra loro perché si riferiscono a un qualche "fattore" non direttamente osservabile né misurabile. Lo scopo dell'AF consiste nell'individuare questo fattore.

In questo caso ognuna delle variabili quantitative rilevate è considerata dipendente, nel senso che è funzione di uno o più fattori. Nel caso di due o più fattori, questi, per ipotesi, sono incorrelati fra di loro. Le variabili indipendenti sono quindi i singoli fattori.

Lo scopo delle tecniche di analisi fattoriale è quello di condensare il set di informazioni molto esteso (dato dai valori assunti dalle variabili originali) in un set più piccolo (i valori dei fattori) senza perdere troppe informazioni. I fattori devono quindi comportarsi in modo molto simile alle variabili originali che li compongono, cioè devono essere molto correlati ad esse.

Per chiarire il concetto (e per fare riferimento al contesto nel quale è nata questa metodologia di analisi) si può considerare la situazione in cui si voglia misurare l'intelligenza di un gruppo di individui. Questa caratteristica non è direttamente osservabile, ma si può fare riferimento ad un insieme di indicatori (rilevabili e misurabili) che si ritengono correlati con il fattore "intelligenza". Gli indicatori dovranno essere correlati con la variabile di interesse. Tipicamente le variabili che vengono considerate in casi come questi sono i punteggi riportati in test attitudinali che misurano l'abilità linguistica, di calcolo e di logica.

Per valutare invece la condizione socio-economica di un gruppo di famiglie ci si potrebbe basare invece, per esempio, sulle informazioni circa il reddito, l'occupazione, il livello di istruzione dei suoi componenti.

**5)** l'**analisi discriminante** o **AD (Proc DISCRIM**). Questa tecnica viene utilizzata nei casi in cui le *n* unità sono naturalmente suddivise in gruppi, come quando le unità presentano o meno una certa caratteristica (sano-malato, occupato-disoccupato) o appartengono a gruppi ben identificati (maschio-femmina, celibe-coniugato-vedovo-separato-divorziato).

La variabile dipendente è quella che definisce i gruppi, mentre esistono più variabili esplicative che si ritiene siano correlate con la variabile dipendente. Sulla base delle informazioni raccolte sulle *n* unità, si costruisce una **regola di classificazione**, basata sui valori assunti dalle variabili esplicative, che consente di assegnare un'unità a uno dei gruppi. In pratica si misurano le differenze fra gli elementi appartenenti ai singoli gruppi e le differenze esistenti fra i gruppi stessi e si tenta di individuare quelle particolari variabili che contribuiscono alla differenziazione fra i gruppi.

In genere l'analisi viene svolta in **due fasi diverse**:

- nella prima vengono identificate le variabili esplicative che consentono la migliore attribuzione delle unità ai gruppi e viene stabilita la regola di classificazione;
- nella seconda vengono considerate delle nuove unità sulle quali viene applicata la regola di classificazione stabilita in precedenza e si misura la proporzione di unità che viene classificata correttamente e in maniera errata, ottenendo una valutazione della bontà del modello.

Esempi di applicazione di questo metodo ai casi reali si hanno

- quando una banca vuole individuare le variabili in grado di distinguere fra clienti affidabili e non,
- quando si vuole determinare la probabilità di insolvenza/bancarotta delle società in funzione dei valori assunti da variabili economiche e finanziarie estratte dai bilanci

- quando si vuole determinare la specie di una pianta sulla base di alcune caratteristiche delle foglie e dei fiori (il primo a utilizzare l'AD fu Fisher, che nel 1936 la utilizzò per attribuire dei reperti fossili alle scimmie o agli umanoidi).

**6)** l'**analisi dei gruppi** o **cluster analysis (Proc CLUSTER)**. Anche in questo caso si utilizzano tecniche per ridurre in qualche modo la quantità dei dati, costruendo gruppi di unità (detti **cluster**) in base a una qualche "somiglianza" o "vicinanza" fra le unità che li formano. Lo scopo di un'analisi dei gruppi consiste infatti nel ridurre il numero delle righe di una matrice dei dati analoga a quella riportata nella tabella 1.1, sostituendo a tutte le righe che contengono i dati relativi alle unità confluite in un singolo cluster, una riga (in genere fittizia) rappresentativa del cluster stesso. Lo scopo è quello di formare gruppi "omogenei" di unità che devono differire fra loro per almeno una caratteristica.

La costruzione dei cluster si può effettuare in molti modi:

- in funzione della scelta del criterio di "misura della somiglianza" (o della "differenza") tra i dati
- in funzione delle diverse strategie di raggruppamento

Ogni scelta tra questi criteri porta, in genere, a classificazioni più o meno differenti (due unità potrebbero appartenere allo stesso gruppo utilizzando un certo criterio di classificazione, ma a gruppi diversi con un'altra classificazione).

Un elemento fondamentale per la determinazione dell'algoritmo di costruzione dei cluster è la misura che si intende adottare per valutare la "somiglianza" o la "dissomiglianza" tra due unità: per variabili quantitative uno dei metodi maggiormente usati è la usuale "distanza euclidea".

Una volta deciso come misurare la dissomiglianza o distanza tra i dati, si deve scegliere il metodo di classificazione. Questi vengono distinti in:

- metodi di **classificazione gerarchica**
- metodi di **classificazione non gerarchica**.

I primi consentono di ottenere un insieme di gruppi ordinabili secondo livelli crescenti, con un numero di gruppi che va da *n* (ciascuna unità costituisce un gruppo) a 1 (**metodi aggregativi o agglomerativi**) oppure da 1 ad *n* (**metodi partitivi, divisivi o scissori**). I metodi aggregativi procedono attraverso una serie di successive fusioni delle singole unità in gruppi sempre più ampi, fino a quando tutte le unità risultano inserite in un unico gruppo (costruzione ascendente delle gerarchie). I metodi scissori, invece, suddividono il complesso delle unità in sottoinsiemi sempre più ristretti, fino a ottenere le singole unità (costruzione discendente delle gerarchie).

Entrambi questi metodi vengono utilizzati quando il numero di variabili rilevate non è particolarmente elevato, per la complessità dei calcoli necessari.

I metodi non gerarchici creano invece un'unica partizione delle *n* unità in *g* gruppi, dove *g* deve essere specificato a priori, e possono essere usati in ogni situazione.

Nella cluster analysis, a differenza dell'analisi discriminante, non viene fatta alcuna assunzione "a priori" sui gruppi. Nella maggior parte dei casi, infatti, questa tecnica viene utilizzata a fini esplorativi, al fine di ottenere la partizione più probabile. Alcuni tipici esempi di applicazione dell'analisi dei gruppi si ha

- in ricerche di mercato, per identificare gruppi omogenei di consumatori di un bene,
- in ambito assicurativo, per identificare gruppi di assicurati con caratteristiche comuni,
- in ricerche geografiche, per identificare aree terrestri con caratteristiche simili.
- per la ricerca degli outlier: in alcune situazioni gli outlier sono più importanti dei valori comuni, come nell'individuazione di frodi nelle carte di credito oppure o nel commercio elettronico

**7)** l'**analisi delle corrispondenze** o **AC** (**Proc CORRESP**). Questa tecnica venne utilizzata la prima volta per descrivere le relazioni fra due variabili qualitative, con lo scopo di confrontare le distribuzioni condizionate di una variabile al variare delle determinazioni assunte dall'altra. Per esempio, sulla base della distribuzione dei condannati in Italia a seconda del tipo di delitto commesso e della posizione nella professione, si potrebbero analizzare le distribuzioni del tipo di delitto commesso per gruppi omogenei rispetto alla posizione nella professione o le distribuzioni della posizione nella professione per gruppi omogenei rispetto al tipo di delitto.

Successivamente questo tipo di analisi è stata estesa al caso di più variabili (qualitative o quantitative) e, quindi, alle tabelle di contingenza multiple (**ACM**). Per ottenere risultati facilmente interpretabili si utilizza un'opportuna rappresentazione grafica delle modalità dei caratteri in uno spazio di dimensionalità minima (in quasi tutte le applicazioni si utilizza semplicemente il piano cartesiano).

L'analisi delle corrispondenze trova ampia applicazione nello studio di matrici di contingenza e per l'analisi di dati raccolti tramite inchieste, sondaggi e ricerche di mercato. Lo scopo è quello di spiegare perché la matrice dei dati si discosta dalla situazione di indipendenza (che si avrebbe se le righe o le colonne fossero tutte proporzionali) e deve evidenziare l'intreccio di legami, le corrispondenze appunto, tra le righe, tra le colonne e tra righe e colonne della matrice dei dati. Questi legami vengono messi in evidenza mediante opportune rappresentazioni grafiche.

Il procedimento utilizzato consiste nel rappresentare graficamente i dati raccolti, per cui le righe e le colonne della matrice, opportunamente ricodificate, vengono interpretate come punti geometrici in due diversi spazi multidimensionali, nei quali è definita una distanza. Si considerano quindi le due "nuvole" di punti (che derivano dalle righe e dalle colonne della matrice) e ciascuna nuvola viene proiettata in un sottospazio a due dimensioni, costruito in modo da rappresentare la nuvola originaria nel miglior modo possibile. Grazie a opportune trasformazioni operate sulle righe e sulle colonne della matrice, si possono far coincidere i due piani, ottenendo una sorta di "mappa" unica sulla quale le righe e le colonne della matrice vengono ad essere rappresentate dalle proiezioni dei loro punti rappresentativi.
Per valutare la dispersione dei profili, riga e colonna, rispetto al loro "centro di gravità" viene utilizzata la metrica del Chi-quadrato.
La prossimità tra proiezioni sulla mappa serve per individuare le prossimità tra punti delle nuvole nel loro spazio multidimensionale ossia serve per individuare gli eventuali legami tra le caratteristiche rilevate sulle *n* unità statistiche.
Nell'ACM possiamo distinguere due tipologie di variabili:

- **attive**: utilizzate per la formazione degli assi fattoriali
- **passive** o **supplementari**: si considera la loro posizione sugli assi fattoriali per aiutare l'interpretazione degli assi stessi.

## 2. BREVI RICHIAMI ALLE REGOLE DI SCRITTURA DI UN PROGRAMMA SAS

### 2.1 Interfaccia

L'interfaccia SAS è costituita da tre finestre fisse:

• **Editor**

**•Visualizzatore dei risultati**

**• Log**

La scrittura del programma viene effettuata nella finestra Editor, mentre le segnalazioni inviate dal sistema sono contenute nella finestra LOG, che contiene una copia annotata del programma in cui, per ogni istruzione eseguita, viene evidenziato il numero di osservazioni lette, il tempo impiegato e ogni eventuale anomalia che si è venuta a verificare durante l'esecuzione.
Oltre a queste finestre possono ne comparire altre, opzionali, che dipendono dalle procedure richieste.

Come per ogni applicazione Windows compaiono i soliti menu a tendina, che consentono di eseguire diverse istruzioni, che possono essere sia di carattere generale, come per esempio

- File (apri, salva, stampa…)
- Modifica (taglia, copia, incolla…)
- …

sia tipiche di un ambiente SAS, come per esempio

- Finestra (che consente di spostarsi fra le varie finestre SAS)
- Guida (per leggere i manuali SAS disponibili sul PC in uso)
- …

## 2.2 Struttura generale di un programma SAS

Il primo passo di un qualunque programma SAS consiste nella creazione, gestione ed eventuale manipolazione dei dati da elaborare. Le istruzioni che consentono di effettuare tali operazioni costituiscono il cosiddetto **DATA STEP**.
Il passo successivo consiste nell'analisi di questi dati mediante l'uso di procedure SAS costituiscono il cosiddetto **PROC STEP**.

Un programma SAS può essere costituito da un solo DATA STEP, da un PROC STEP oppure da più DATA STEP e/o più PROC STEP.
Per eseguire correttamente gli STEP, è necessario rispettare alcune regole fondamentali.
Le istruzioni che compongono il programma sono usualmente composte da una successione di parole scritte su una o più righe di cui la prima è la cosiddetta **keyword** o parola chiave, che viene automaticamente riconosciuta dal sistema e colorata in blu.
Ciascuna istruzione termina sempre con il punto e virgola

;

Le istruzioni digitate possono essere scritte in lettere minuscole o maiuscole, possono iniziare in una qualsiasi colonna della riga e proseguire su più righe. Su una stessa riga possono essere anche scritte più istruzioni diverse, ma ciascuna di esse deve comunque essere separata dalle altre mediante il punto e virgola.

In un qualsiasi punto di un programma SAS è possibile inserire commenti, digitando la stringa

/*

Il testo, riconosciuto dal SAS, verrà automaticamente colorato in verde. Per chiudere il commento va digitata la stringa inversa

*/

L'elaborazione dei dati viene eseguita mediante l'esecuzione di procedure (PROC) che indicano al sistema il tipo di analisi che si desidera effettuare. Con i PROC STEP si possono creare (o leggere) dataset SAS che a loro volta possono essere analizzati da altri DATA STEP o da PROC STEP.
Un passo di PROC inizia con l'istruzione PROC seguita dal nome della procedura che si vuole eseguire e termina con un passo di DATA, con un passo di una nuova PROC o con l'istruzione finale

```
run;
```

Ogni procedura SAS opera agisce sull'ultimo Data Set SAS (DSS) creato, a meno che non venga diversamente specificato.

Nelle pagine successive si elencano le procedure necessarie per eseguire le analisi statistiche multivariate considerate nel capitolo 1.