Febbraio 2018

LE REGOLE PER IL RILASCIO DEI RISULTATI DELLE ELABORAZIONI

Estratto da:
Il Laboratorio per l'Analisi dei Dati ELEmentari (ADELE)
Guida all'utenza

Per qualsiasi comunicazione o richiesta di informazioni: adele@istat.it.

Scopi del laboratorio

Il Laboratorio ADELE risponde alle esigenze di analisi statistica per finalità di ricerca scientifica che necessitano dell'utilizzo dei dati elementari, laddove questi non siano disponibili in altre forme.
All'interno del laboratorio, la sicurezza dei dati e il segreto statistico sono garantiti dal controllo sia dell'ambiente di lavoro che dei risultati delle analisi condotte dagli utenti.

Una volta concluse tutte le elaborazioni relative al progetto di ricerca, l'output di cui si chiede il rilascio sarà valutato sotto il profilo della riservatezza statistica dallo staff del Laboratorio ADELE.

Cosa è possibile ottenere e cosa non è possibile ottenere

Si riportano di seguito alcune regole specifiche per i tipi di output più frequenti:

1) Statistiche descrittive e tabelle a supporto di modelli statistici

Ciascun valore riportato nelle tabelle o nelle statistiche descrittive in genere, deve essere riferito ad almeno 10 unità statistiche.
In particolare:

- statistiche descrittive che riportino dati puntuali sulle singole unità (ad esempio massimo e minimo per variabili continue) non possono essere rilasciate;
- moda, mimino e massimo: possono essere rilasciati se le modalità che individuano sono assunte da almeno 10 unità;
- quantili: la mediana è considerata rilasciabile se riferita ad una distribuzione di almeno 50 unità; gli altri quantili non sono rilasciati salvo casi particolari da concordare;
- medie, rapporti e indicatori: questi output devono essere presentati nella loro forma disaggregata (ad esempio per le medie e i rapporti: separare numeratore e denominatore; medie di variabili dicotomiche: presentare anche il complemento, etc.); ciascun elemento deve essere corredato dal numero di unità (almeno 10) che concorrono a determinarne il valore; ciò vale anche per eventuali complementi, anch'essi da presentare in forma esplicita (ad esempio, se un indicatore riporta il valore del 95%, si deve poter verificare che anche il 5% corrisponda ad almeno 10 unità; stessa cosa per le medie delle variabili dicotomiche, etc.);
- tabelle di intensità: gli utenti devono specificare il numero di unità (almeno 10) che concorrono a determinare il valore di ciascuna cella (abbinare la relativa tabella di frequenza);
- tabelle di frequenza: non sono in ogni caso rilasciate tabelle con numerosità di cella inferiore alle 10 unità non pesate.

2) Grafici sulle variabili

I grafici su variabili non continue devono essere corredati dalla corrispondente tabella di valori che rappresentano; questa sarà valutata secondo quanto specificato al punto precedente. I grafici su variabili continue devono essere salvati come immagini e privati dei valori in ascissa.

3) Regressioni

Possono essere rilasciati i seguenti output:

a) (p-1) *parametri stimati*, dove p è il numero di regressori, quando siano verificate tutte le condizioni appresso specificate:

- ✓ il numero complessivo di osservazioni deve eccedere il numero di variabili esplicative di almeno 100 unità;
- ✓ tra le variabili esplicative occorre la presenza di almeno una variabile per la quale abbiano senso le operazioni di somma, differenza, prodotto e quoziente;
- ✓ le osservazioni su tutti i dati debbono essere riferite ad almeno 100 unità di analisi differenti.

b) Diagrammi sulla corretta specificazione del modello:

1. l'*istogramma dei residui*, privato dei valori in ascissa;
2. il diagramma della *densità dei residui*, privato dei valori in ascissa;
3. il Q-Q plot dei residui, privato dei valori di ascisse ed ordinate;
4. il P-P plot dei residui;
5. il diagramma dei ranghi dei residui contro i ranghi dei valori predetti della variabile esplicanda;
6. il diagramma dei ranghi dei residui contro i ranghi di una variabile esplicativa;

c) Statistiche sull'adattamento e la corretta specificazione del modello:

1. le statistiche espresse da uno scalare;
2. le statistiche espresse da un vettore, avente dimensione non superiore al numero di parametri stimati, ossia (p-1). Del regressore oscurato viene rilasciato soltanto il livello convenzionale di significatività (0.005, 0.01, 0.025, 0.05, 0.1).

In ogni caso restano esclusi dal rilascio:

1. i residui della regressione;
2. i valori "predetti" della variabile esplicanda.

4) Analisi fattoriale e modelli ad equazioni strutturali

Possono essere rilasciati i seguenti output:

1. i parametri del modello,
2. la (eventuale) matrice di correlazione tra i fattori,
3. gli *standard errors* e le statistiche sulla significatività dei parametri del modello,
4. comunalità e specificità per ciascuna variabile,
5. i punteggi fattoriali riferiti ad unità di analisi che non siano individui, famiglie o imprese,
6. le statistiche sulla bontà del modello, espresse da uno scalare,
7. gli *scree plot* relativi agli autovalori delle matrici di covarianze/correlazioni osservate,
8. i diagrammi dei modelli relazionali tra variabili manifeste e latenti.

5) Analisi in componenti principali

Possono essere rilasciati i seguenti output:

1. autovalori,
2. le seguenti statistiche:
   a) varianza spiegata dagli assi fattoriali,
   b) matrice (p x k) dei contributi relativi (quadrati dei coseni) dei punti-variabile,
   c) matrice (p x k) dei contributi assoluti dei punti-variabile,
   d) matrice (p x k) delle coordinate dei punti-variabile,

dove p è il numero di variabili e k è il numero degli autovalori che, ordinati in successione non decrescente, cumulano una frazione della variabilità totale non superiore all'85%,

3. scree plot degli autovalori,
4. diagrammi relativi alla proiezione dei punti-variabile sui piani fattoriali.

6) Analisi delle corrispondenze

Possono essere rilasciati i seguenti output:

1. autovalori,
2. le seguenti statistiche:
   a) inerzia spiegata dagli assi fattoriali,
   b) matrice (p x k) dei contributi relativi (quadrati dei coseni) dei punti-modalità (colonna e/o riga),
   c) matrice (p x k) dei contributi assoluti dei punti-modalità (colonna e/o riga),
   d) matrice (p x k) delle coordinate dei punti-modalità (colonna e/o riga),

   dove p non eccede il numero complessivo di modalità e k è il numero degli autovalori che, ordinati in successione non decrescente, cumulano una frazione dell'inerzia totale non superiore all'85%,
   e) valori test, espressi da scalari, sulla significatività di ciascuna modalità supplementare (nell'analisi delle corrispondenze multiple),
3. scree plot degli autovalori,
4. diagrammi relativi alla proiezione dei punti-modalità riga e/o colonna sui piani fattoriali.

Relativamente alle unità di analisi, per qualunque tipo di elaborazione, restano esclusi dal rilascio i valori osservati e le statistiche non conformi alle regole su "Statistiche descrittive e tabelle".

Regole per la presentazione dei risultati delle elaborazioni

- È fortemente sconsigliato produrre risultati senza l'impiego dei pesi di riporto all'universo; tuttavia, ai fini della valutazione, gli utenti devono presentare (anche) le frequenze non pesate delle analisi; l'utente è invitato ad indicare se le proprie elaborazioni fanno uso di pesi standardizzati (normalizzati) e in che modo (se la normalizzazione è rispetto al totale della popolazione o a sottopopolazioni specifiche);
- Il **volume** dell'output può essere considerato esso stesso una ragione di rifiuto al rilascio: l'output di cui si chiede il rilascio deve essere minimale e corrispondere a quanto sarà incluso nel lavoro che si intende divulgare; a titolo indicativo, viene suggerito un numero massimo di 30 pagine (~ 60Kb in ASCII text format);
- L'output deve essere preferibilmente fornito in file di testo, oppure in file Word o Excel, ma non nel formato proprietario delle applicazioni statistiche utilizzate; eventuali statistiche descrittive e tabelle devono essere fornite in formato Excel;
- L'output deve essere redatto in modo da poter essere rilasciato così com'è, senza necessità di modifiche da parte del personale che ne effettua la valutazione: in caso di output non rilasciabile sarà necessario proseguire le elaborazioni per rendere l'output rilasciabile;
- L'output deve essere chiaramente ed estesamente documentato secondo la "Scheda per la descrizione dell'output" (*cfr.* Allegato 1), nella quale va specificato: lo scopo e le modalità

dell'analisi, nome e contenuto dei file di output, i trattamenti effettuati sul data set originario e le eventuali (sub)popolazioni oggetto d'analisi, il significato di ciascuna variabile (per quelle derivate anche la definizione), ed ogni altra informazione utile ad una corretta interpretazione dei file di output. La descrizione dell'output deve essere sufficiente a comprenderlo (non è consentito il riferimento ad altre fonti quali, ad esempio, i file di sintassi utilizzati);

- **Non è consentito il rilascio di output intermedi** (ovvero di elaborazioni che non concludano il progetto);

Al termine di ciascun progetto agli utenti viene chiesto di rispondere facoltativamente a un breve questionario finalizzato a valutare gli aspetti del servizio dal punto di vista dell'utente. Il modulo non fa riferimento a dati personali; le informazioni raccolte sono utilizzate esclusivamente per produrre dei report sulla qualità del servizio e non sono in alcun modo diffuse associandole a dati personali sugli utenti. Tale questionario verrà inviato via email all'utente al termine del progetto.

ALLEGATO 1

SCHEDA PER LA DESCRIZIONE DELL'OUTPUT

DATI UTILIZZATI

Specificare, tra i dati forniti, quelli effettivamente utilizzati nelle elaborazioni di cui si chiede il rilascio: indicare il nome ed il periodo di riferimento della/e rilevazione/i utilizzate e specificare eventuali file di dati esterni impiegati nell'elaborazione.

DESCRIZIONE DELLE VARIABILI / INDICATORI

Riportare il nome ed una breve descrizione delle variabili utilizzate. Nel caso di variabili non presenti nelle basi di dati originarie (riclassificazioni effettuate dall'utente, variabili esterne etc.) oltre al nome ed alla descrizione, riportare il significato delle modalità assunte(o il procedimento di costruzione, soprattutto nel caso in cui la variabile assuma valori in funzione di altre variabili).

DESCRIZIONE DELLE TRASFORMAZIONI OPERATE SULLE VARIABILI

Per ciascuna variabile fornita dal laboratorio e sottoposta a trasformazione, indicare la funzione utilizzata per ottenerne la trasformazione. Per ciascuna variabile creata ex-novo dall'utente indicare in modo dettagliato il procedimento di costruzione.

FILE DI OUTPUT

Riportare il nome e la struttura (esempio: file excel con un foglio per anno considerato) dei file di output dei quali si richiede il rilascio, fornendo una descrizione sintetica del contenuto.

ELABORAZIONI EFFETTUATE

Descrivere le singole elaborazioni effettuate, fornendone una descrizione breve ma esauriente.

È utile associare una denominazione a ciascuna elaborazione e riportarla nel file di output, così da poterne garantire una non equivoca identificazione ed interpretazione.

FILTRI SULLE UNITÀ

Per ciascuna elaborazione (o gruppo di elaborazioni) specificare i filtri applicati alla popolazione di partenza e la numerosità delle osservazioni coinvolte.

Notare che è necessario specificare esattamente la numerosità effettiva in ogni elaborazione, anche nel caso di riduzioni della numerosità dovute alla presenza di valori mancanti in una o più delle variabili adoperate.

SISTEMA DI PESI

Specificare il sistema di pesi eventualmente utilizzato e se questo varia tra le diverse elaborazioni.

Nel caso si faccia uso di pesi standardizzati (normalizzati), specificare se la normalizzazione è rispetto al totale della popolazione o a sottopopolazioni specifiche.

Notare che nel caso si richieda il rilascio di output pesato, lo stesso deve essere presentato anche in versione non pesata per consentirne la valutazione.

NOTE

Riportare ogni altra informazione si ritenesse utile ad una corretta interpretazione dei file di output.

Il richiedente: _____ Data: ___/___/_____

N.B.: la descrizione dell'output deve essere sufficiente a comprenderlo; non è consentito il riferimento ad altre fonti (quali, ad esempio, i file di sintassi utilizzati).