

Salvatore Ingrassia
Carmela Senatore

Laboratorio di Statistica I

Guida alle Attività

Facoltà di Economia, Università della Calabria
Corso di Laurea in Statistica
Anno Accademico 2002-2003

Indice

1	L'Analisi dei dati con SPSS	1
1.1	Introduzione	1
1.2	Statistiche descrittive	1
1.2.1	Frequenze	5
1.2.2	Descrittive	10
1.2.3	Esplora	10
1.2.4	Tavole di Contingenza	14
1.3	Grafici	17
1.3.1	Grafico a barre	17
1.3.2	Grafico a torta	19
1.3.3	Grafico a scatola	21
1.3.4	Grafico a dispersione	22
1.3.5	Istogramma	24
1.4	Sintassi: trasformazione dei dati	25
1.5	Sintassi: procedure	27

Capitolo 1

L'Analisi dei dati con SPSS

1.1 Introduzione

Nel capitolo precedente sono state trattate le procedure per l'importazione e la preparazione di un file di dati con il software SPSS , si è quindi visto come costruire un file da utilizzare durante l'analisi statistica di un fenomeno. In questa parte, si provvederà a condurre un'analisi di tipo descrittivo univariata e bivariata, cioè si mostrerà come ottenere statistiche, distribuzioni di frequenza e grafici attraverso l'utilizzo di SPSS; sarà condotta lo stesso tipo di analisi trattata nei capitoli 1-2 con il software EXCEL.

Per comodità, durante la trattazione saranno utilizzate solo due variabili del file di dati german.dat descritto nel capitolo 5 e precisamente:

- Variabile X_1 : Saldo attuale del conto corrente, variabile categorica ordinale a 4 modalità:
 - A_{11} : $X_1 < 0DM$
 - A_{12} : $0 \leq X_1 < 200DM$
 - A_{13} : $X_1 \geq 200DM$ / salario ricevuto per almeno un anno
 - A_{14} : nessun conto corrente
- Variabile X_2 : Durata del prestito in mesi, variabile quantitativa

1.2 Statistiche descrittive

Alle procedure per l'analisi di tipo descrittivo, in SPSS, si accede dal menù analizza alla voce statistiche descrittive, da quest'ultimo sottomenù si possono selezionare diversi tipi di analisi (vedi figura 1.1), in particolare:

- frequenze;
- descrittive;
- esplora;
- tavole di contingenza;

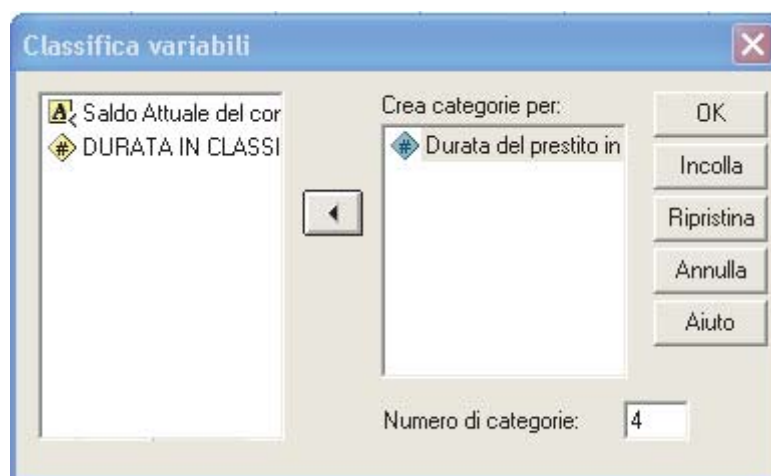
Figura 1.1: *Menù statistiche descrittive.*

le prime due funzioni permettono di condurre analisi univariate, mentre le ultime due analisi bi-variate; ovviamente ogni tipo di analisi può essere condotta contemporaneamente per più variabili.

Prima di procedere con l'analisi dettagliata dei suddetti comandi è utile introdurre il problema relativo alla ricodifica delle variabili; si è visto, infatti, nel capitolo 1 che in presenza di variabili con valori numerici continui per la costruzione della distribuzione di frequenza conviene costruire un numero discreto di categorie. Le funzioni che permettono di fare questo tipo di operazione sono più di una e sono previste nel menù trasforma sotto le voci:

- classifica variabili;
- ricodifica variabili.

Classifica variabili consente di convertire i dati numerici continui in un numero discreto di categorie. La procedura crea nuove variabili contenenti dati categoriali. I dati sono classificati in base a gruppi percentili, ciascuno dei quali contiene approssimativamente lo stesso numero di casi (vedi figura 1.2).

Figura 1.2: *Classificazione delle variabili.*

L'opzione ricodifica prevede le seguenti voci:

- ricodifica nelle stesse variabili: consente di riassegnare i valori delle variabili esistenti oppure di comprimere gli intervalli di valori esistenti in nuovi valori; quindi, si effettuano le ricodifiche richieste all'interno della variabile originale, modificandola in modo definitivo (i valori originari non possono essere recuperati); (vedi figura 1.3);

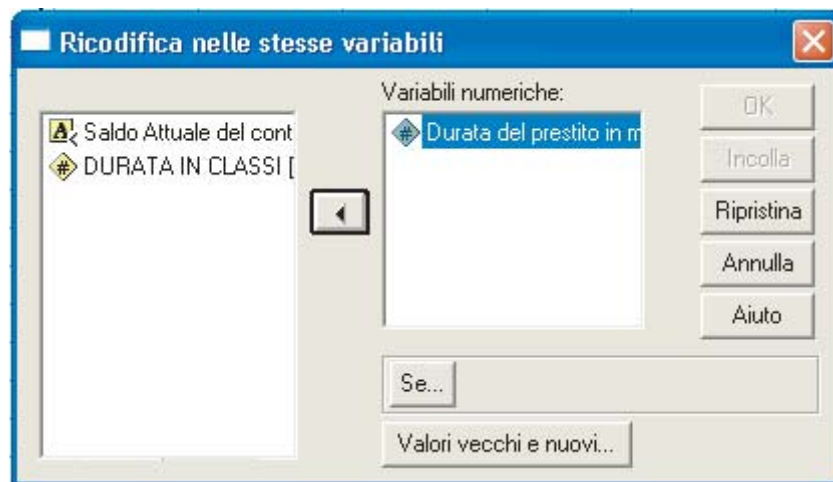


Figura 1.3: Ricodifica nella stesse variabili.

- ricodifica in variabili differenti: consente di riassegnare i valori delle variabili esistenti o di comprimere intervalli di valori esistenti nei nuovi valori di una nuova variabile (vedi figura 1.4); quindi, si creano nuove variabili contenenti le ricodifiche effettuate, lasciando intatte le variabili originali (questa procedura è molto più prudente e consigliabile rispetto alla prima).



Figura 1.4: Ricodifica in variabili differenti.

i passi per la ricodifica non cambiano nei due casi l'unica differenza consiste nel fatto che quando si crea una nuova variabile è necessario assegnargli un nome ed una etichetta. Nelle finestre di ricodifica nella parte bassa troviamo due pulsanti, SE e VALORI VECCHI E NUOVI; cliccando su questi si aprono le relative finestre di dialogo. La finestra relativa a SE permette di applicare trasformazioni di dati al sottoinsieme di casi selezionato utilizzando le espressioni logiche, mentre la finestra relativa a VALORI VECCHI E NUOVI permette di definire i valori da ricodificare (vedi figura 1.5).

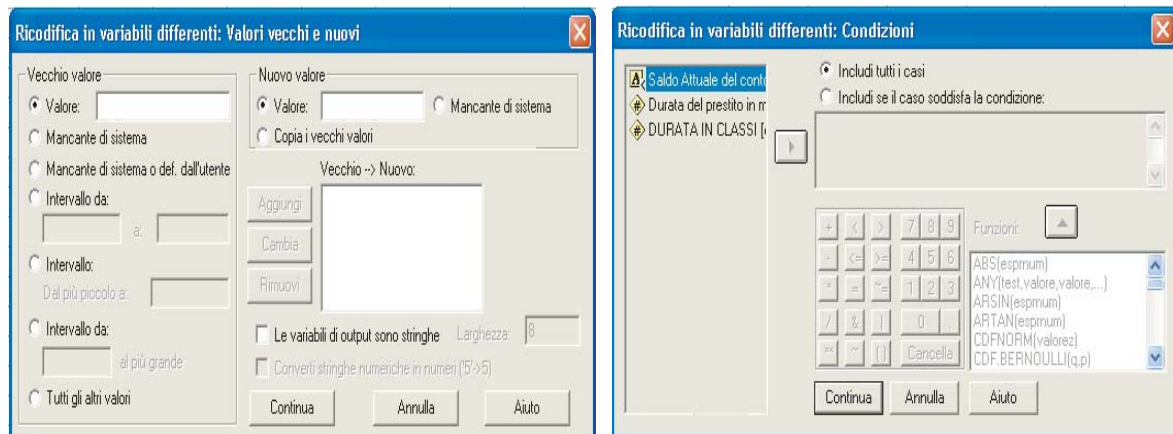


Figura 1.5: Ricodifica variabili.

A questo punto utilizzando la procedura ricodifica in variabili differenti si costruisca la variabile DURATA5 la quale è ottenuta dalla variabile DURATA ed è costituita di 5 classi di frequenza con ampiezza pari a 14, la procedura da seguire è la seguente: nel riquadro di sinistra (Vecchio valore) si spunta la voce **intervallo da a** e si riempiono i campi, in questo caso i valori sono 4 e 18 per la prima classe, nel riquadro di destra (Nuovo valore), si attribuisce un nuovo valore (ad esempio 1) a questo intervallo; si clicca poi su aggiungi. Allo stesso modo si definiscono gli altri quattro intervalli. Per quanto riguarda i valori mancanti di sistema o definiti dall'utente, si può richiedere che vengano copiati i vecchi valori o che vengano considerati come mancanti di sistema, cliccando successivamente su aggiungi; cliccando a questo punto sul pulsante continua si torna alla finestra precedente e cliccato su cambia, e su Ok, si manda in esecuzione il comando. La nuova variabile creata compare così nella matrice dei dati. A questo punto è possibile (sul foglio visualizzazione variabili) definire le etichette dei valori, la lunghezza, il numero dei decimali e tutti gli altri attributi.

Spesso, si ha la necessità, partendo da una o più variabili di costruire delle nuove variabili effettuando dei calcoli sulle variabili originarie. A questo scopo SPSS mette a disposizione la procedura Calcola. Dalla barra dei menù si clicca su Trasforma, quindi su calcola; selezionando la procedura calcola, si apre la finestra riportata in figura 1.6. In essa, in alto a sinistra, occorre indicare il nome della variabile che si desidera creare, variabile destinazione, mentre nel riquadro di destra, espressione numerica, occorre registrare l'espressione che si desidera applicare al fine di ottenere la nuova variabile. Sulla pulsantiera della calcolatrice compaiono:

- operatori aritmetici (addizione, sottrazione, moltiplicazione, divisione, elevazione a potenza);
- operatori relazionali (uguale, non uguale a, minore di, maggiore di, minore o uguale a, maggiore o uguale a);

- operatori logici (AND, OR, NOT);

questi ultimi sono molto importanti nel caso si voglia procedere al calcolo di nuove variabili non su tutti i casi ma solo su alcuni.

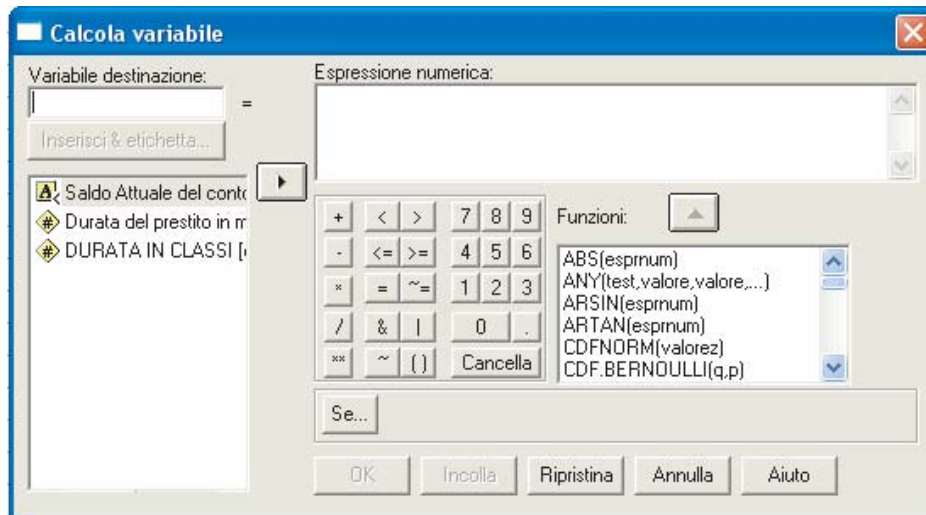


Figura 1.6: *Calcola variabile.*

1.2.1 Frequenze

Optando per la voce frequenze, dal sottomenù statistiche, sarà visualizzata la finestra in figura 1.7.

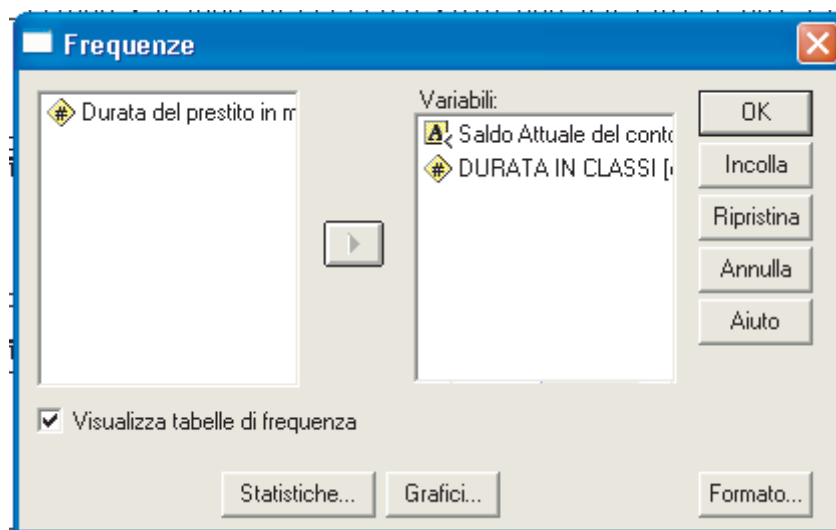


Figura 1.7: *Statistiche descrittive: frequenze.*

Nel riquadro a sinistra si trova l'elenco alfabetico di tutte le variabili del file, selezionando una variabile e selezionando quelle di cui si interessa conoscere le distribuzioni di frequenza e poi clic-

cando sulla freccia esse sono trasferite nella casella a destra, selezionando la voce visualizza tabelle di frequenza e cliccando su ok i risultati dell'elaborazione richiesta saranno visualizzati sulla finestra di output. Inoltre, sempre da questa finestra e come in tutte le finestre di dialogo proposte da SPSS cliccando con il pulsante destro del mouse, si può accedere al menù informazioni sulle variabili tramite il quale si ha la possibilità di visualizzare le impostazioni relative agli attributi relativi alla variabile quali: nome, misura ed etichette di valori (vedi figura 1.8).

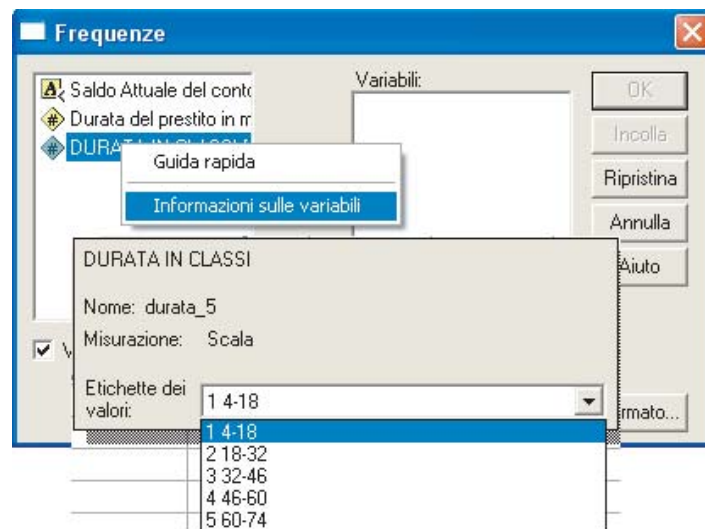


Figura 1.8: Informazioni sulla variabile selezionata.

Nella finestra frequenza troviamo nella parte bassa tre pulsanti: STATISTICHE, GRAFICI, FORMATO. Cliccando su questi pulsanti si aprono le relative finestre di dialogo. La finestra relativa alle statistiche contiene quattro riquadri, nel primo si può scegliere tra i quartili i percentili o la divisione di n gruppi uguali. Nel secondo relativo alle misure di posizione cioè alla situazione del valore centrale si può scegliere tra moda, media e mediana. Nel terzo, relativo alle misure di dispersione, cioè alla posizione dei valori della distribuzione rispetto alla misura centrale si trovano alcuni indici di dispersione come deviazione standard, varianza, errore standard, range e valore minimo e massimo. Nel quarto riquadro si trovano gli indicatori di simmetria e di curtosi della distribuzione per la verifica della normalità; la curva normale, a forma di campana, è basata su una distribuzione teorica di un infinito numero di osservazioni. La forma a campana implica che la maggior parte dei valori venga a trovarsi attorno alla media, questa curva ha delle caratteristiche importanti: è perfettamente simmetrica, il 50 % dei casi si trova sul lato sinistro rispetto alla media e l'altro 50 % si trova sul lato destro, media, mediana e moda coincidono. Tuttavia molte distribuzioni di variabile non hanno la forma della curva normale, molto spesso le distribuzioni sono asimmetriche e presentano una coda di valori piuttosto pronunciata a destra o a sinistra: quando ciò avviene la moda e la mediana non coincidono più:

- se l'asimmetria è positiva allora la curva presenterà una coda pronunciata di valori verso destra e la media si troverà a destra rispetto alla mediana;
- se l'asimmetria è negativa allora la curva presenterà una coda pronunciata di valori verso sinistra e la media si troverà a sinistra rispetto alla mediana.

Inoltre, spesso, rispetto alla curva normale, la forma della distribuzione di certe variabili può essere: più alta o più stretta: in questo caso la curva è leptocurtica, cioè la varianza attorno alla media è molto piccola; oppure può essere più bassa e più piatta: in questo caso la curva è platicurtica, cioè la varianza attorno alla media è piuttosto grande. La curtosi è l'indicatore di questi tipi di forma:

- se la curtosi > 0 allora la curva è leptocurtica;
- se la curtosi < 0 allora la curva è platicurtica;
- se la curtosi è $= 0$ allora la curva è mesocurtica (molto simile alla curva normale teorica).

Quindi selezionando le voci asimmetria e curtosi dalla finestra statistiche si possono conoscere interessanti caratteristiche della distribuzione.

Una volta selezionate le statistiche desiderate e possibili si dovrà cliccare sul pulsante continua per tornare alla finestra principale (vedi figura 1.9).

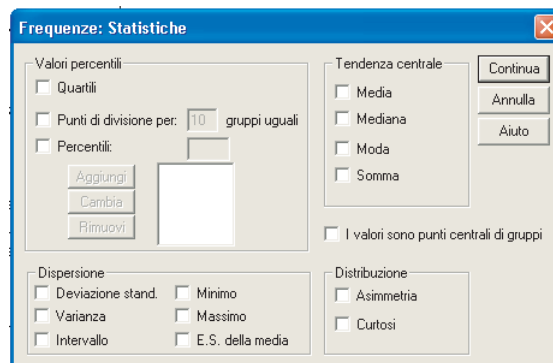


Figura 1.9: *Statistiche descrittive: frequenza.*

Sempre dalla finestra frequenze si può scegliere la voce grafici che introduce nella relativa sotto finestra (vedi figura 1.10), qui si può scegliere tra il grafico a barre, grafici a torta e l'istogramma, quest'ultimo può riportare sovrapposta la curva normale. Si può inoltre scegliere se utilizzare le frequenze assolute o le percentuali per rappresentare le etichette dei valori. Cliccando sul grafico ottenuto è possibile modificarlo e copiarlo.

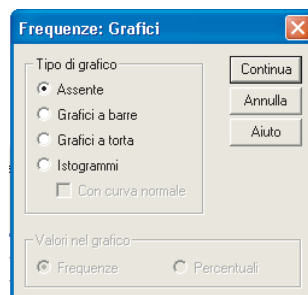


Figura 1.10: *Grafici delle frequenze.*

Nelle figure 1.11, 1.12, 1.13 sono riportati rispettivamente degli esempi di output di statistiche e grafici.

Statistiche		Saldo Attuale del conto corrente	DURATA IN CLASSI
N	Validi	1000	1000
	Mancanti	0	0
Media			1.69
Mediana			1.00
Moda			1
Varianza			.826
Asimmetria			1.189
Errore std dell'asimmetria			.077
Curtosi			.502
Errore std della curtosi			.155
Minimo			1
Massimo			5
Percentili	25		1.00
	50		1.00
	75		2.00

Tabella di frequenza

Saldo Attuale del conto corrente				
		Frequenza	Percentuale	Percentuale valida
Validi	<0 dm	274	27.4	27.4
	0-200 dm	269	26.9	26.9
	>=200	63	6.3	6.3
	no conto	394	39.4	39.4
Totale		1000	100.0	100.0

DURATA IN CLASSI				
		Frequenza	Percentuale	Percentuale valida
Validi	4-18	546	54.6	54.6
	18-32	281	28.1	28.1
	32-46	108	10.8	10.8
	46-60	64	6.4	6.4
	60-74	1	.1	.1
Totale		1000	100.0	100.0

Figura 1.11: Output delle frequenze.

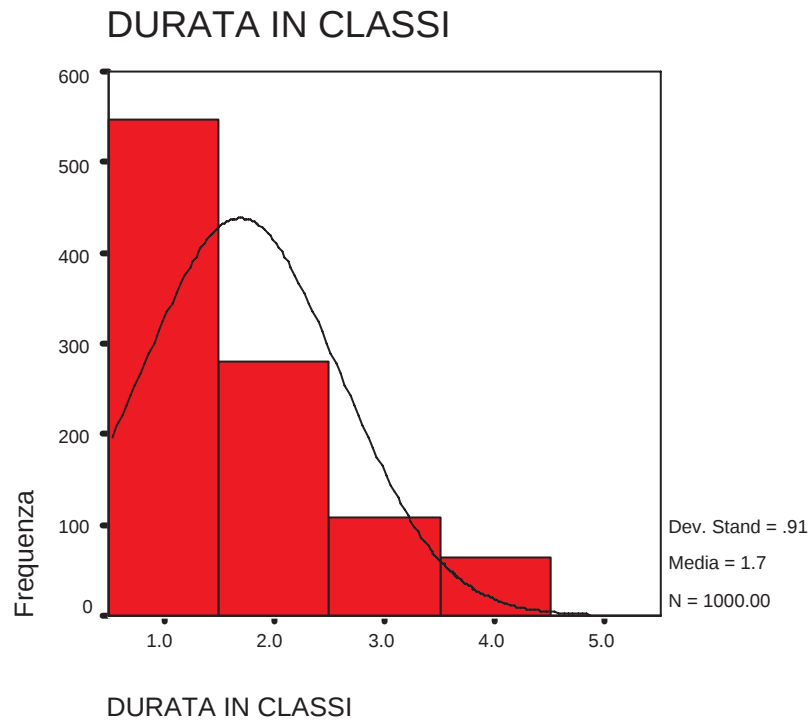


Figura 1.12: *Istogramma.*

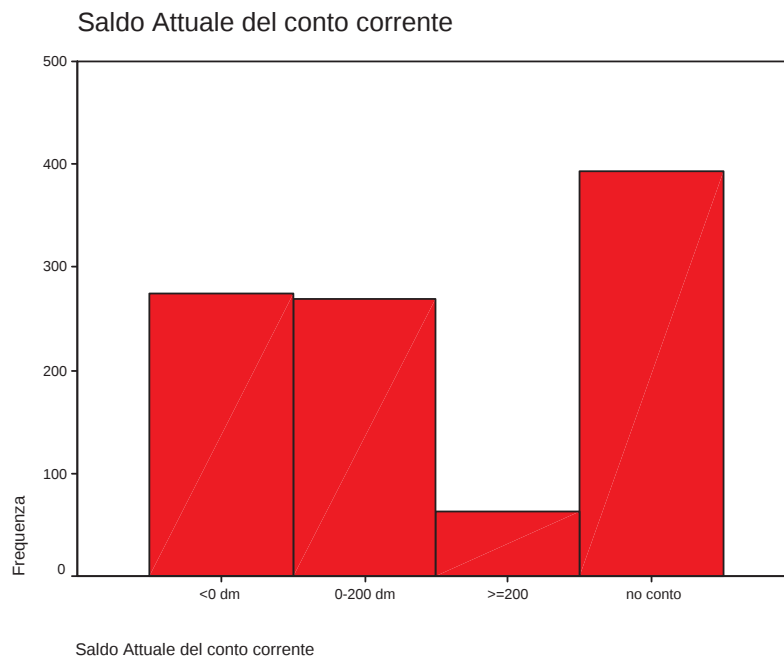


Figura 1.13: *Grafico a barre.*

1.2.2 Descrittive

Selezionando la voce descrittive sarà visualizzata una finestra strutturata nello stesso modo della finestra frequenze, quindi la procedura della selezione delle variabili da analizzare non cambia; inoltre in questa finestra si può decidere se salvare i valori standardizzati della variabile analizzata. Nella finestra descrittive nella parte bassa si trova il pulsante opzioni al quale è collegata una finestra di dialogo strutturata in tre riquadri: nel primo si selezionano misure di dispersione, nel secondo misure relative alla distribuzione e nel terzo l'ordine di visualizzazione che può essere relativo alle variabili nel file, all'ordine alfabetico, alle medie crescenti e decrescenti (questa opzione è molto utile poiché SPSS permette di analizzare contemporaneamente più variabili). Questa funzione può essere applicata solo a variabili numeriche (vedi figura

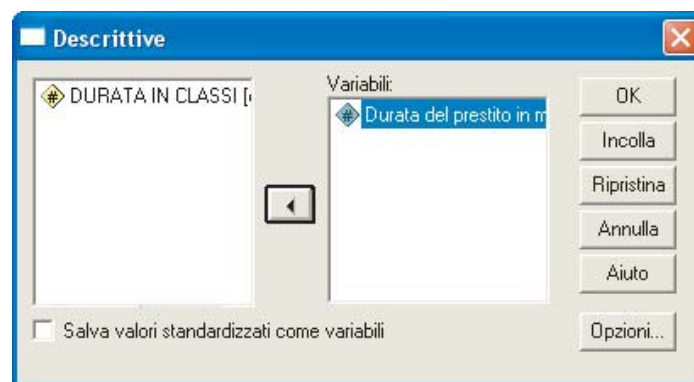
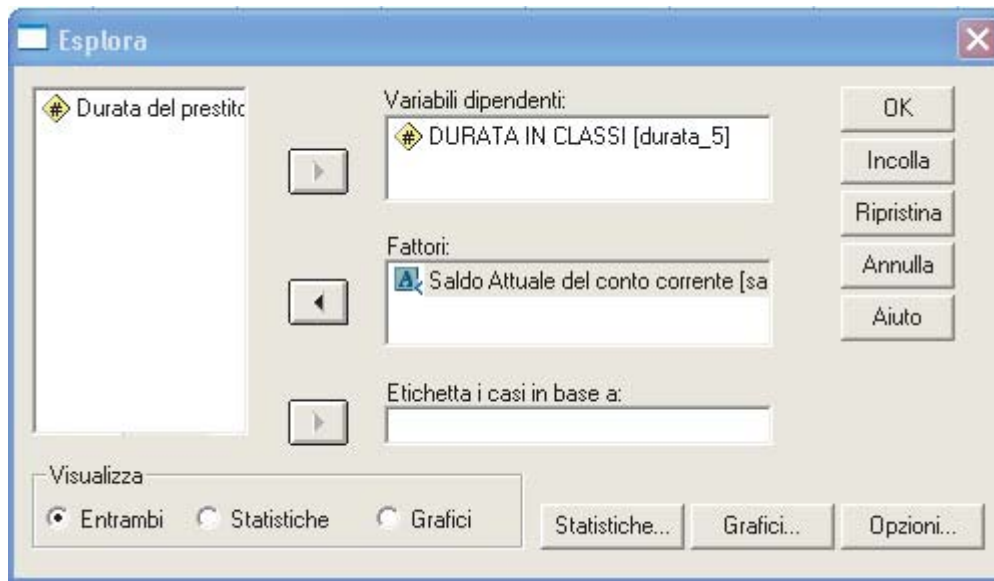


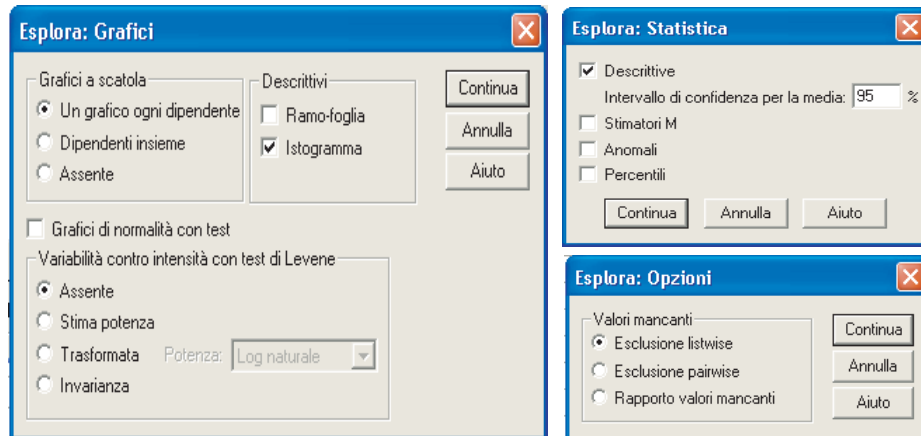
Figura 1.14: *Statistiche descrittive: descrittive.*

1.2.3 Esplora

Selezionando la procedura Esplora si producono statistiche riassuntive e visualizzazioni grafiche per tutti i casi o per singoli gruppi di casi. Risulta inoltre utile per numerose operazioni, ovvero screening dei dati, identificazione dei valori anomali, descrizione, verifica delle ipotesi e caratterizzazione delle differenze tra sottopopolazioni (gruppi di casi). Lo screening dei dati può evidenziare la presenza di valori insoliti, intervalli vuoti tra i dati o altri elementi specifici. L'esplorazione dei dati può consentire di determinare l'idoneità delle tecniche statistiche selezionate per l'analisi dei dati. L'esplorazione può evidenziare la necessità di eseguire una trasformazione dei dati se una particolare tecnica richiede una distribuzione normale. In figura 1.15 è visualizzata la finestra relativa alla funzione esplora, la struttura ancora una volta è identica a quella della finestra sopra descritta. Nel campo variabili dipendenti possono essere inserite solo variabili numeriche mentre nel campo fattori devono essere inserite variabili in modo da produrre analisi separate per gruppi di casi in base ai valori di una o più variabili; selezionando più variabili vengono prodotti riepiloghi separati per ciascuna variabile dipendente e per ciascun fattore.

Figura 1.15: *Statistiche descrittive:Esplora.*

In basso alla finestra esplora c'è un riquadro in cui scegliere se visualizzare solo le statistiche solo i grafici o entrambi; inoltre si trovano i seguenti pulsanti: statistiche, grafici, opzioni (vedi figura 1.16).

Figura 1.16: *Statistiche, grafici, opzioni.*

La finestra statistiche prevede la selezione delle voci:

- descrittive, e selezionandola saranno visualizzate la media, la mediana, la moda, l'errore standard, la varianza, la deviazione standard, il minimo, il massimo, l'intervallo, la distanza interquartilica, l'asimmetria, l'errore standard dell'asimmetria. La curtosi e l'errore standard della curtosi;
- stimatori m, sono stimatori robusti della tendenza centrale, basati sul criterio di massima verosimiglianza. Ai casi vengono assegnati pesi diversi in base a criteri che variano da uno

stimatore all'altro. I valori estremi ricevono un peso inferiore rispetto a quello dei valori più vicini al centro. Quando i dati provengono da una distribuzione simmetrica con code lunghe, o quando i dati hanno valori estremi, gli stimatori M forniscono stime migliori della tendenza centrale rispetto a quelle fornite dalla media o dalla mediana;

- anomali, visualizza i casi con i cinque valori più alti e quelli con i cinque valori più bassi. Vengono etichettati come estremi nell'output;
- percentili, visualizza i seguenti percentili: 5, 10, 25, 50, 75, 90, 95 e Vengono, inoltre, visualizzati i cardini di Tukey, che definiscono i quartili della distribuzione.

La finestra grafici è struttura in tre riquadri: nel primo è possibile selezionare i grafici a scatola, nel secondo grafici descrittivi e nel terzo grafici relativi al test di Levene.

Infine la finestra opzioni valori mancanti, consente di controllare la modalità di elaborazione dei valori mancanti. In particolare possono essere selezionate le seguenti voci:

- esclusione listwise: i casi con valori mancanti per qualsiasi variabile dipendente o fattore verranno esclusi da tutte le analisi. Impostazione di default;
- esclusione pairwise: i casi che non contengono valori mancanti per le variabili di un gruppo (cella) verranno inclusi nell'analisi per tale gruppo. Il caso può includere valori mancanti per le variabili utilizzate in altri gruppi;
- rapporto valori mancanti: i valori mancanti per le variabili fattore vengono trattati come categoria distinta. Tutto l'output viene prodotto per questa categoria aggiuntiva. Le tabelle di frequenza includono categorie per i valori mancanti. I valori mancanti per una variabile fattore vengono inclusi, ma etichettati come mancanti.

In figura 1.17 e 1.17 sono riportate due tabelle ottenute dalla procedura esplora per le variabili saldo attuale del conto corrente e durata del prestito.

Riepilogo dei casi							
Durata del prestito in mesi	Saldo Attuale del						
	<0 dm	274	100,0%	0	,0%	274	100,0%
	0-200 dm	269	100,0%	0	,0%	269	100,0%
	>=200	63	100,0%	0	,0%	63	100,0%
	no conto	394	100,0%	0	,0%	394	100,0%

Figura 1.17: Output prodotto dalla funzione esplora.

Descrittive				
Durata del prestito in mesi	<0 dm	Media	21,34	,704
		Intervallo di confidenza per la media al 95%	19,95	
		Limite inferiore		
		Limite superiore	22,73	
		Media 5% trim	20,61	
		Mediana	18,00	
		Varianza	135,976	
		Deviazione std.	11,661	
		Minimo	6	
		Massimo	60	
		Intervallo	54	
		Distanza interquartilica	12,00	
		Asimmetria	,934	
		Curtosi	,382	
	0-200 dm	Media	22,68	,834
		Intervallo di confidenza per la media al 95%	21,04	
		Limite inferiore		
		Limite superiore	24,32	
		Media 5% trim	21,78	
		Mediana	18,00	
		Varianza	187,300	
		Deviazione std.	13,686	
		Minimo	6	
		Massimo	72	
		Intervallo	66	
		Distanza interquartilica	18,00	
		Asimmetria	1,035	
		Curtosi	,453	
	>=200	Media	17,35	1,200
		Intervallo di confidenza per la media al 95%	14,95	
		Limite inferiore		
		Limite superiore	19,75	
		Media 5% trim	16,77	
		Mediana	15,00	
		Varianza	90,650	
		Deviazione std.	9,521	
		Minimo	4	
		Massimo	42	
		Intervallo	38	
		Distanza interquartilica	14,00	
		Asimmetria	,898	
		Curtosi	,119	
no conto		Media	19,95	,570
		Intervallo di confidenza per la media al 95%	18,83	
		Limite inferiore		
		Limite superiore	21,07	
		Media 5% trim	19,03	
		Mediana	18,00	
		Varianza	127,937	
		Deviazione std.	11,311	
		Minimo	4	
		Massimo	60	
		Intervallo	56	
		Distanza interquartilica	12,00	
		Asimmetria	1,145	
		Curtosi	1,360	

Figura 1.18: Output prodotto dalla funzione esplora.

1.2.4 Tavole di Contingenza

La procedura Tavole di contingenza consente di formare tabelle bivariate e a più dimensioni e fornisce una serie di test e misure di associazione per le tabelle bivariate. Il test o la misura da utilizzare vengono determinati in base alla struttura della tabella e al fatto che le categorie siano ordinate o meno. Le statistiche e le misure delle tavole di contingenza vengono calcolate solo per le tabelle bivariate. Se si specifica una riga, una colonna o uno strato (variabile di controllo), verrà visualizzato un riquadro contenente le statistiche associate e le misurazioni per ciascun valore dello strato (o una combinazione di valori per due o più variabili di controllo).

Quindi, è chiaro che questo comando è un ottimo strumento per l'analisi dei legami tra variabili. Se ad esempio si vuole analizzare l'influenza di una variabile (dipendente) su un'altra (indipendente) ponendo la variabile dipendente in riga e la variabile indipendente in colonna e selezionando le percentuali per colonna, uno dei metodi di lettura della tavola di contingenza consiste nel confrontare ogni percentuale di colonna (se si percentualizza all'interno delle modalità della variabile posta in colonna) con il totale marginale di riga corrispondente. Se la percentuale di colonna risulta essere maggiore del totale di riga (scarto positivo) si parla di *attrazione positiva* tra le modalità, se risulta essere inferiore (scarto negativo) si parla di *attrazione negativa*. Qualora le due percentuali fossero uguali ciò indicherebbe che non vi è alcuna relazione fra le due variabili.

Se invece si vuole misurare l'intensità di una relazione tra due variabili occorre determinare misure di associazione (quando la relazione fra due variabili nominali), di cograduazione (quando la relazione fra due variabili ordinali) o di correlazione (quando la relazione fra due variabili cardinali). Se le due variabili sono invece miste (ad esempio una nominale e una ordinale) si utilizza la tecnica di analisi statistica applicabile alla variabile di livello inferiore.

In figura 1.19 è visualizzata la finestra relativa al comando tabelle di contingenza; in essa bisogna specificare la formattazione della tabella cioè le variabili che avranno modalità su riga e su colonna, inoltre selezionando le relative voci si può decidere se visualizzare le tabelle e i grafici.

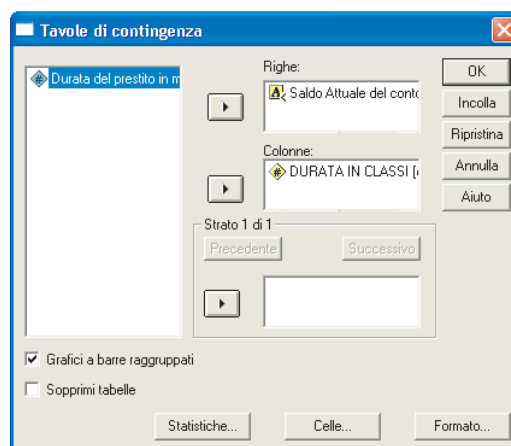


Figura 1.19: *Tabelle di contingenza*.

Come le finestre dei comandi descritti nei precedenti paragrafi, anche la finestra tavole di contingenza ha nella parte bassa tre pulsanti: statistiche, celle, formato (vedi figura 1.20).

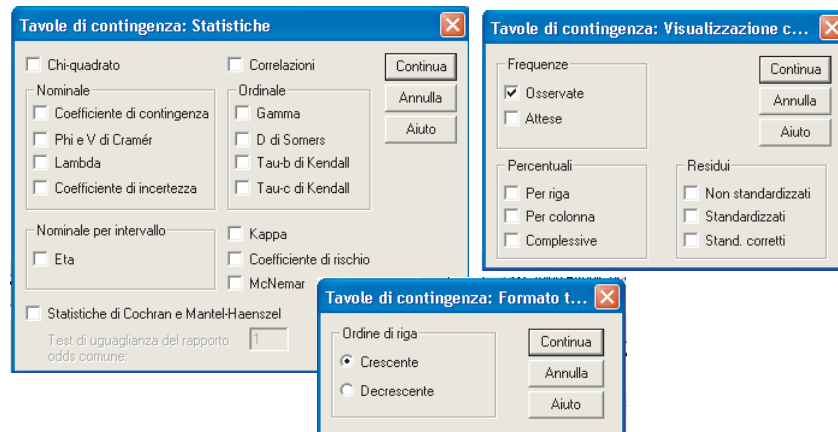


Figura 1.20: Opzioni tabelle di contingenza.

Nella finestra statistiche è possibile selezionare le seguenti voci:

- **chi-quadrato:** per tabelle con due righe e due colonne, scegliendo Chi-quadrato si calcolano il chi-quadrato di Pearson, il chi-quadrato del rapporto di verosimiglianza, il test esatto di Fisher e il chi-quadrato corretto di Yates (correzione di continuità). Per le tabelle 2x2, il test esatto di Fisher viene calcolato quando una tabella non creata in base a righe o colonne mancanti in una tabella di dimensioni maggiori contiene una cella con una frequenza attesa minore di 5. Per tutte le altre tabelle 2x2 viene calcolato il chi-quadrato corretto di Yates. Per tabelle con un numero qualsiasi di righe e colonne, si può selezionare Chi-quadrato per calcolare il chi-quadrato di Pearson e il chi-quadrato del rapporto di verosimiglianza. Se entrambe le variabili delle tabelle sono quantitative, l'opzione Chi-quadrato restituisce il test dell'associazione lineare;
- **correlazioni:** per tabelle in cui sia le righe che le colonne contengono valori ordinati, l'opzione Correlazioni restituisce il coefficiente di correlazione di Spearman, ρ (solo per dati numerici). Il coefficiente ρ di Spearman è una misura di associazione tra punteggi di rango. Se entrambe le variabili delle tabelle (fattori) sono quantitative. Correlazioni restituisce il coefficiente di correlazione di Pearson, r , una misura dell'associazione lineare tra le variabili;
- **nominale:** per i dati nominali, è possibile selezionare il coefficiente ϕ e V di Cramér, il Coefficiente di contingenza, λ (λ simmetrico e asimmetrico e τ di Goodman e Kruskal), nonché il Coefficiente di incertezza;
- **ordinale:** per tabelle in cui sia le righe che le colonne contengono valori ordinati, è possibile selezionare Gamma (gamma di ordine zero per tabelle a 2 vie e gamma condizionali per tabelle da 3 a 10 vie), Tau-b di Kendall e Tau-c di Kendall. Per desumere le categorie delle colonne delle righe, invece, si può selezionare D di Somers;
- **nominale per intervallo:** se una variabile è categoriale e l'altra quantitativa, selezionare Eta; in questo caso è necessario che la variabile categoriale sia codificata numericamente;
- **kappa:** per tabelle che contengono le stesse categorie sia nelle righe che nelle colonne;

- coefficiente di rischio: per tabelle con due righe e due colonne, è possibile selezionare Rischio per ottenere la valutazione del coefficiente di rischio e il rapporto odd;
- McNemar: il test di McNemar è un test non parametrico per due variabili dicotomiche correlate. Consente di verificare le variazioni della risposta utilizzando la distribuzione del chi-quadrato;
- statistiche di Cochran e Mantel-Haenszel: è possibile utilizzare le statistiche di Cochran e Mantel-Haenszel per valutare l'indipendenza tra una variabile fattore dicotomica e una variabile risposta dicotomica, sulla base di modelli di covariata definiti da uno o più variabili (di controllo) di strato. Vengono inoltre calcolati il rapporto odds comune di Mantel-Haenszel e le statistiche di Breslow-Day e Tarone per il test di omogeneità del rapporto odds comune;

Dalla finestra visualizzazione celle si può decidere il tipo di frequenze da visualizzare, osservate o attese; il tipo di percentuali, per riga, per colonna, complessive; il tipo di residui non standardizzati, standardizzati, standardizzati corretti.

Infine nella finestra formato si seleziona l'ordine di riga.

In figura 1.21 è riportato un esempio di output ittenuto con la funzione tavole di contingenza.

Riepilogo dei casi						
Saldo Attuale del conto corrente * DURATA IN CLASSI	1000	100,0%	0	,0%	1000	100,0%

Tavola di contingenza Saldo Attuale del conto corrente * DURATA IN CLASSI

Saldo Attuale del conto corrente	<0 dm	142	83	32	17	0	274
	0-200 dm	142	67	31	28	1	269
	>=200	41	16	6	0	0	63
	no conto	221	115	39	19	0	394
Totale		546	281	108	64	1	1000

Chi-quadrato

			Sig. asint.
Chi-quadrato di Pearson	19,521 ^a	12	,077
Rapporto di verosimiglianza	22,488	12	,032
N. di casi validi	1000		

a.

5 celle (25,0%) hanno un conteggio atteso inferiore a 5.

Figura 1.21: Output prodotto dalla funzione tavole di contingenza.

1.3 Grafici

Sebbene ogni procedura proposta da SPSS prevede la visualizzazione dei grafici più adatti per il tipo di analisi che si sta conducendo, alle volte è necessario produrre dei grafici non previsti di default dalla funzione selezionate in questi casi, o quando si desidera visualizzare solo dei grafici è possibile costruirli utilizzando le funzioni del menu Grafici.

I grafici che analizzeremo in questo capitolo sono i seguenti:

- a barre;
- torta;
- grafico a scatola;
- dispersione;
- istogramma;

1.3.1 Grafico a barre

Optando per la voce a barre, dal menù grafici, sarà visualizzata la finestra iniziale riportata in figura 1.22.



Figura 1.22: Menù del grafico a barre.

In questa finestra si può selezionare il tipo di grafico che si vuole ottenere, semplice sovrapposto, raggrupato e come rappresentare i dati nel grafico:

- riepilogo per gruppi di casi: il grafico rappresenta i valori di una variabile raggruppati in base alle categorie di un'altra variabile; nei grafici raggruppati e sovrapposti i sottogruppi sono definiti dalle combinazioni fra le categorie di due variabili;
- riepilogo di variabili distinte: il grafico rappresenta più variabili. I grafici semplici riassumono tutti i casi di ciascuna variabile, mentre quelli complessi rappresentano i valori di ciascuna variabile attraverso le categorie di una variabile di raggruppamento;

- valori singoli casi: il grafico rappresenta i singoli valori di una o più variabili;

la finestra che si visualizza dopo la scelta del grafico è la seguente:



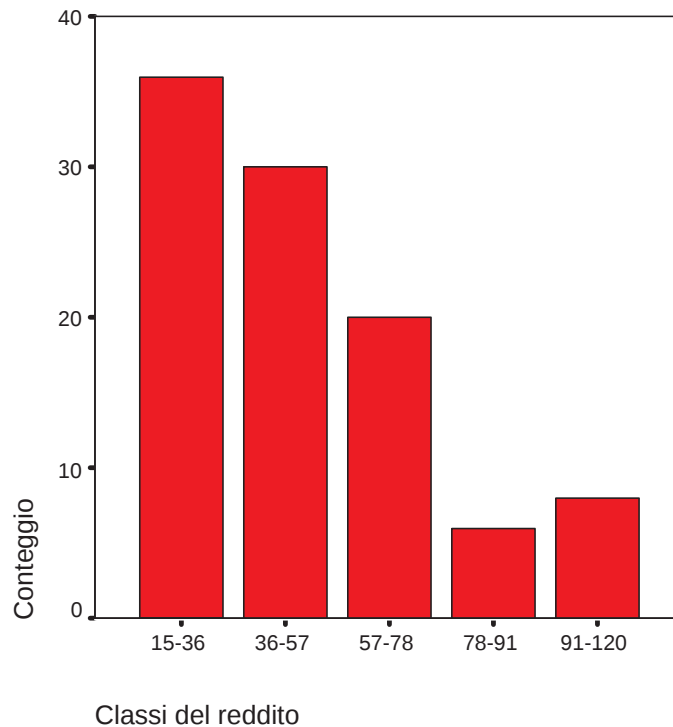
Figura 1.23: Definizione dati delle barre.

in questa finestra si seleziona il tipo di dato che devono rappresentare le barre e l'asse delle categorie. Inoltre, in questa finestra, come in tutte le finestre di dialogo relative alla costruzione di grafici, in basso a destra si trovano i pulsanti: opzioni e titoli, il primo fa scegliere come trattare i dati mancanti il secondo permettere di dare i titoli al grafico (vedi Figura 1.24).



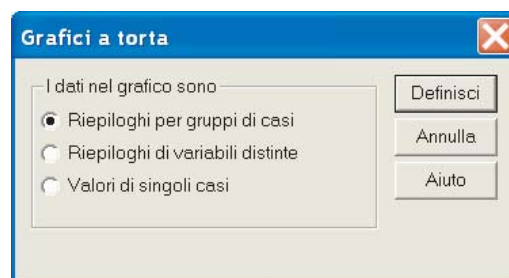
Figura 1.24: Titoli e opzioni del grafico.

In figura 1.25 è riportato un esempio di grafico a barre per la variabile reddito, raggruppati in classi, del set di dati datinordsud.dat.

Figura 1.25: *Grafico a barre.*

1.3.2 Grafico a torta

Selezionando la voce torta, dal menù grafici, sarà visualizzata la finestra iniziale riportata in figura 1.26.

Figura 1.26: *Menù del grafico a torta.*

In questa finestra si può selezionare come rappresentare i dati nel grafico, e ancora una volta le voci che si possono selezionare sono identiche a quelle previste nella finestra di dialogo del grafico a barra, cioè: riepilogo per gruppi di casi, riepilogo di variabili distinte e valori singoli casi.

La finestra che si visualizza dopo la selezione è la seguente:

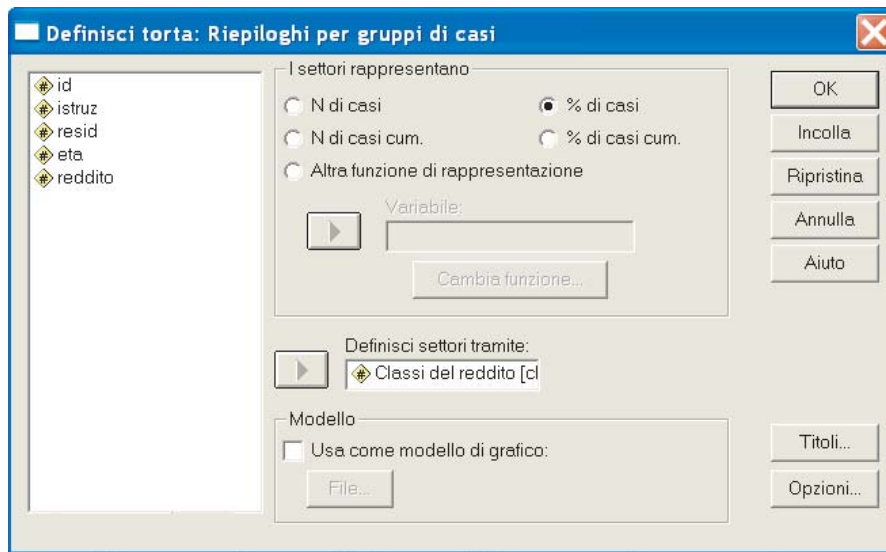


Figura 1.27: Definizione dati del grafico a torta.

in questa finestra si seleziona il tipo di dato che devono rappresentare i settori e la variabile tramite la quale definire i settori. In figura 1.28 è riportato un esempio di grafico a torta per la variabile reddito, raggruppata in classi, del set di dati *datinordsud.dat*.

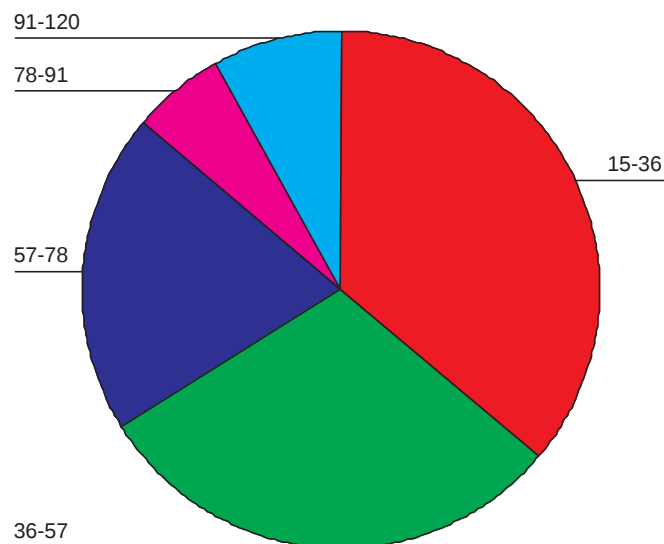


Figura 1.28: Grafico a torta.

1.3.3 Grafico a scatola

Optando per la voce grafico a scatola, dal menù grafici, sarà visualizzata la finestra iniziale riportata in figura ??.



Figura 1.29: Menù del grafico a scatola.

In questa finestra si può selezionare il tipo di grafico che si vuole ottenere semplice o raggrupato. Nel primo caso sarà creato un grafico per ogni variabile, mentre, nel secondo caso in un unico grafico saranno rappresentate le scatole delle diverse variabili, inoltre si deve selezionare come rappresentare i dati nel grafico. Questa volta le voci disponibili sono solo due: riepilogo per gruppi di casi, riepilogo di variabili distinte.

La finestra a cui si accede dopo le selezioni è la seguente:

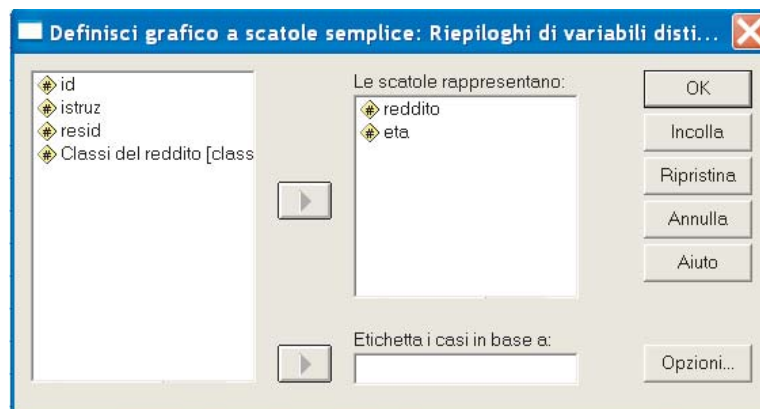


Figura 1.30: Definizione dati del grafico a scatole.

in questa finestra si seleziona il tipo di dato che devono rappresentare le scatole. In figura 1.31 è riportato un esempio di grafico a scatola per la variabile reddito, del set di dati datinordsud.dat.

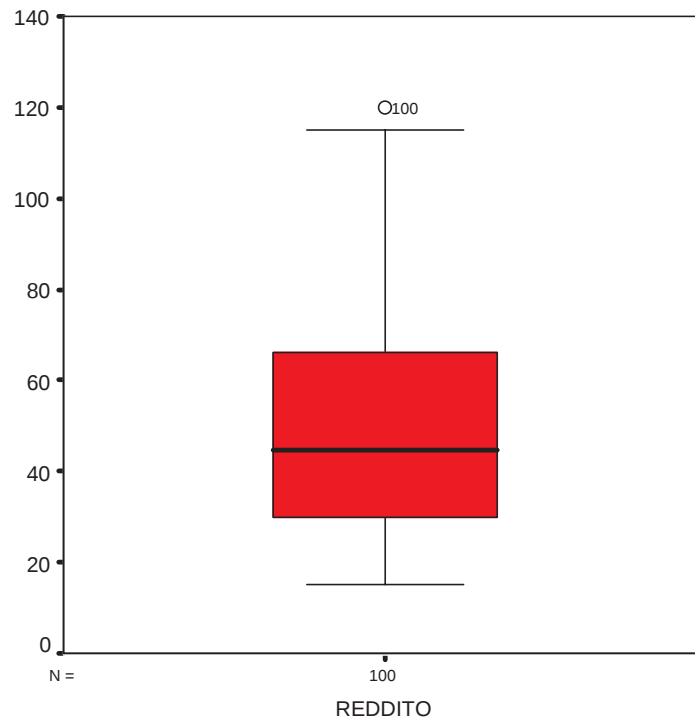


Figura 1.31: Esempio di grafico a scatola.

1.3.4 Grafico a dispersione

Optando per la voce a barre, dal menù grafici, sarà visualizzata la finestra iniziale riportata in figura 1.32.

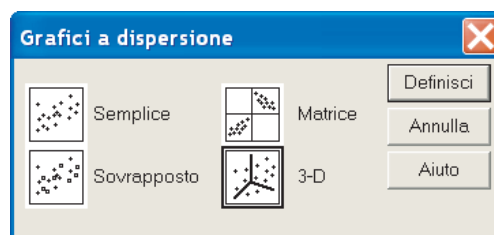


Figura 1.32: Menù del grafico a dispersione.

Anche nella procedura per la costruzione dei grafici a dispersione nel primo passo è prevista la selezione del tipo di grafico che si vuole ottenere, semplice, sovrapposto, matrice e 3-D.

La finestra che si visualizza dopo la scelta del grafico è la seguente:

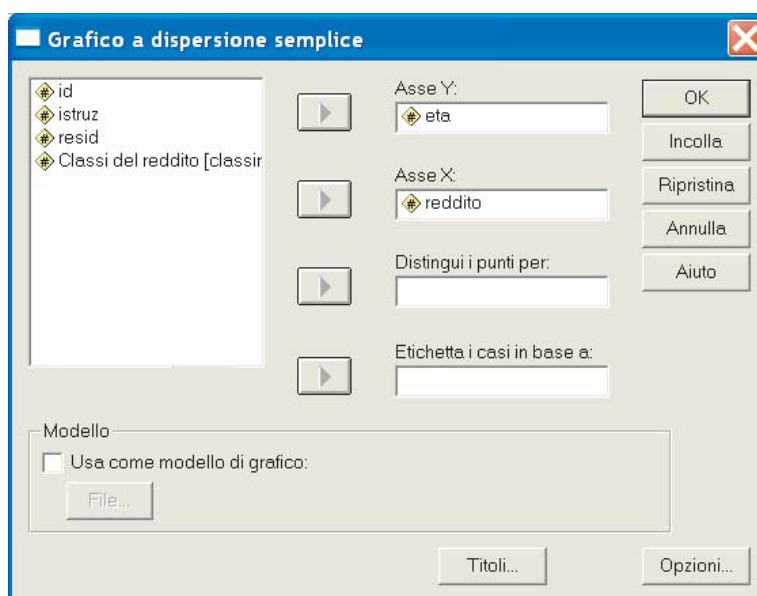
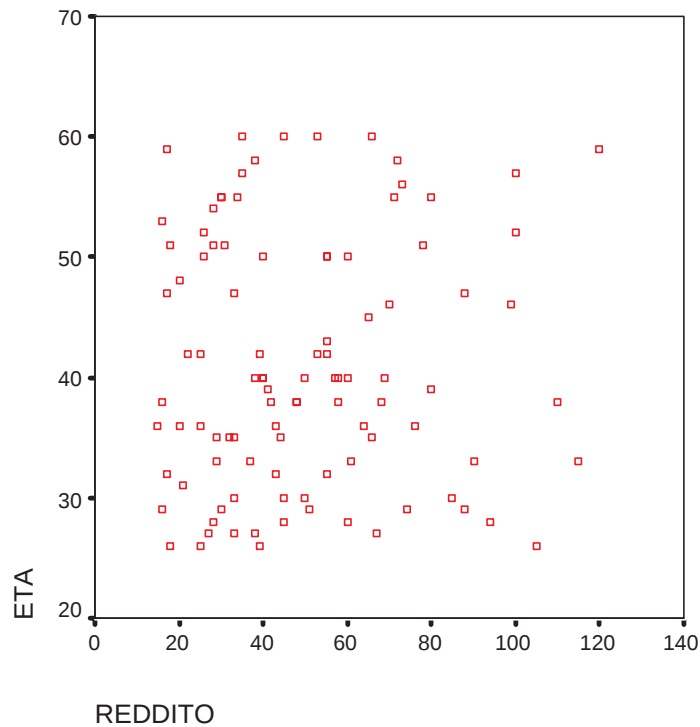


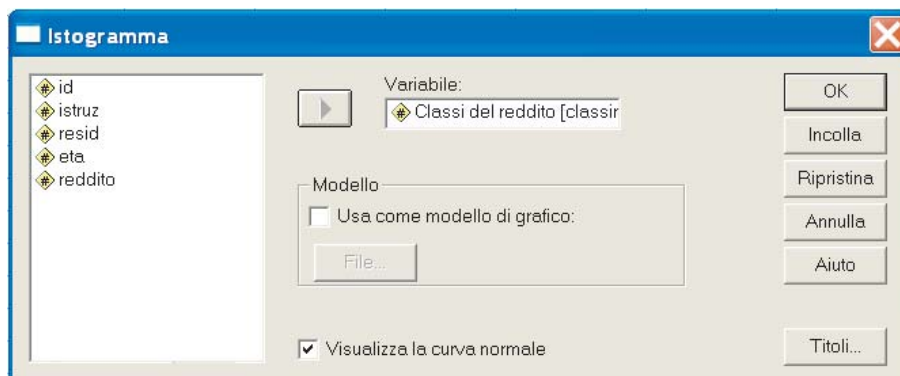
Figura 1.33: *Definizione dei dati del grafico a dispersione.*

in questa finestra si selezionano le variabili da posizionare sugli assi ed eventualmente una variabile che rappresenta gruppi distinti di valori. In figura 1.34 è riportato un esempio di grafico a dispersione per le variabili età e reddito relative al set di dati *datinordsud.dat*.

Figura 1.34: *Grafico a dispersione.*

1.3.5 Istogramma

Optando per la voce istogramma, dal menù grafici, sarà visualizzata la finestra riportata in figura 1.35, in questa finestra si specifica la variabile per la quale si vuole costruire l'istogramma ed inoltre si può decidere se sovrapporre o meno la curva normale all'istogramma.

Figura 1.35: *Definizione dei dati per l'istogramma.*

In figura 1.25 è riportato un esempio di istogramma per la variabile reddito, relativa al set di dati datinordsud.dat.

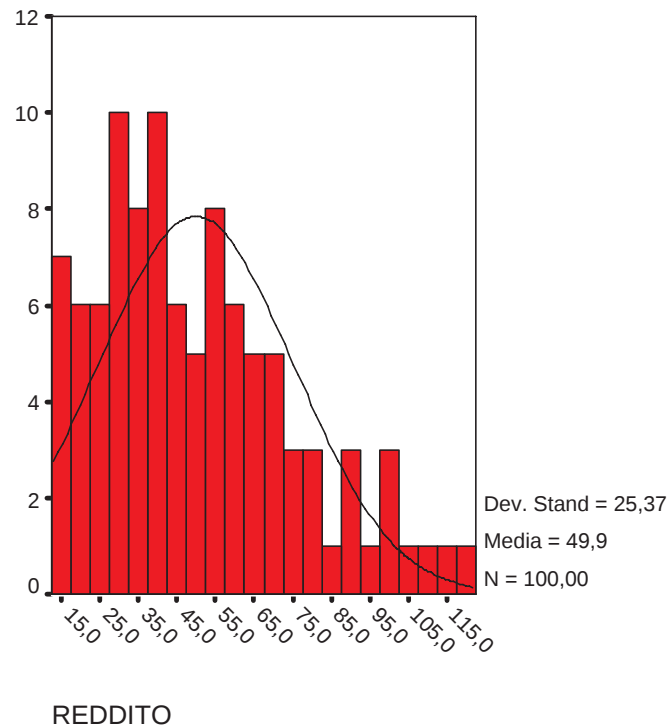


Figura 1.36: Istogramma.

1.4 Sintassi: trasformazione dei dati

Nel capitolo precedente si è visto come definire dei dati attraverso la sintassi di spss, adesso si introdurranno i comandi utilizzati, per creare nuove variabili, modificare quelle esistenti o per di selezione di un sottocampione per l'analisi.

In una situazione ideale i dati sono perfettamente adatti al tipo di analisi desiderata, nella reale situazione di ricerca si è costretti a fare delle trasformazioni dei dati affinché essi rivelino più chiaramente la struttura sottostante la realtà indagata. I comandi che la sintassi di spss offre per fare questo tipo di trasformazione sono i seguenti:

- compute;
- recode;
- count;
- if;
- select if;

Compute:

Questo comando si utilizza per creare nuove variabili, attraverso trasformazioni di variabili esistenti. Crea variabili nuove o di modificare delle variabili vecchie utilizzando delle formule matematiche e lacune funzioni. Queste variabi sono conosciute come variabili calcolate.

Oltre agli operatori aritmetici sono utilizzabili le funzioni che rientrano nelle categorie seguenti:

- funzioni sui valori mancanti;
- funzioni sui record;
- funzioni per numeri casuali;
- funzioni per la gestione di date;

Un esempio dell'uso di questo comando è riportato in figura 1.37.

Recode:

L'istruzione record serve per trasformare i valori di una variabile in altri valori. Ad esempio, per trasformare alcuni valori in valori mancanti, per passare da un certo livello di misurazione ad uno più basso, per raggruppare fra loro valori diversi che si desidera elaborare assieme, cioè, permette di ricodificare i valori di una variabile riducendone le modalità.

All'interno del comando si può utilizzare la convezione *trhu* oppure le keyword MISSING, SYS-MIS O ELSE.

Un esempio dell'uso di questo comando è riportato in figura 1.37.

Count:

Si usa per creare una variabile numerica, che per ogni caso, conta quante volte lo stesso valore o una lista di valori compare attraverso una lista di variabili. In genere alla variabile si attribuisce un nuovo nome, che verrà aggiunta alla struttura dei dati; è possibile, però, indicare una variabile già esistente che verrà ovviamente riscritta e il cui vecchio contenuto andrà perso. Per quanto riguarda la sintassi è identica a quella dei comandi precedentemente descritti, cioè, si scrive il comando (COUNT), quindi il nome della varibile e poi le condizioni per il conteggio.

If:

Questo comando permette di fare trasformazioni simili al COMPUTE su casi che soddisfano una determinata condizione logica. La sinatssi è molto simile a quella del comando compute, infatti è possibile usare tutte le funzioni applicabili al comando compute; si scrive prima il nome del comando seguito dall'espressione logica, quindi il nome della variabile a cui applicare la trasfrmazione e per finire si digita uguale seguito dal valore che si intende assegnare. Se il nome della variabile è quello di una variabile non presente nella struttura dei dati sarà aggiunta in coda ai dati è quando la condizione per attribuire il valore non è sodisfatta verrà attribuito un valore mancante.

Select If:

Con questo comando è possibile selezionare dei casi sulla condizione logica, la sintassi è molto semplice, basta scrivere il comando seguito dall'espressione logica. Il select if seleziona i comandi in modo irreversibile, quindi per rendere temporanea la selezione dei casi è necessario premettere al comando select if il comando TEMPORARY, quest'ultimo seleziona i casi solo per il comando di procedura successivo, ad esempio una frequenza.

1.5 Sintassi: procedure

Precedentemente si è visto come costruire delle procedure attraverso il menù *analizza*, ovviamente sono tutte procedure attuabili anche attraverso i comandi di sintassi. I comandi che analizzeremo in questo paragrafo sono i seguenti:

- frequencies;
- descriptive;
- crosstabs;

Questa procedura permette di calcolare le distribuzioni di frequenza. La stessa procedura può produrre delle grossolane rappresentazioni grafiche sul tipo degli istogrammi a barre, può calcolare i percentili e i ranghi percentili e può essere utilizzata per ottenere alcune statistiche descrittive standard.

La sintassi è molto semplice poiché basta scrivere il nome del comando, il nome delle variabili coinvolte e definire tutte le opzioni relative al comando quest'ultima sono seguite dai seguenti comandi:

- **FORMAT** permette di controllare il formato di stampa dei risultati: i valori riscontrati nella variabile sono generalmente elencati in ordine crescente di valore, ma possibile richiederli in ordine decrescente di valore (**DVALUE**) oppure in ordine crescente (**AFREQ**) o decrescente (**DFREQ**) di frequenze osservate. Se i valori della variabile sono molti possibile richiedere l'uso di un formato **CONDENSE** che visualizza le percentuali in una tabella più compatta anche se meno precisa; oppure possibile specificare che ogni variabile deve occupare una sola pagina (**ONEPAGE**), in tal caso SPSS passa automaticamente al formato compresso quando le variabili presentano molti valori. Se si desiderano le percentuali solo delle variabili con pochi valori, possibile usare l'opzione **LIMIT(n)** che stampa i risultati solo per le variabili che presentano meno di *n* valori diversi;
- **NOLABELS** sopprime la stampa delle etichette associate ai valori e **NOTABLE** sopprime la stampa della tavola delle frequenze e viene usata in associazione con le opzioni grafiche. **DOUBLE** chiede l'uso dell'interlinea doppia e **NEWPAGE** chiede un salto pagina per ogni variabile;
- **MISSING=INCLUDE** chiede che i valori mancanti delle variabili siano compresi nella tavola delle frequenze come valore utili;
- **BARCHART** e **HISTOGRAM** svolgono la stessa funzione con alcune piccole differenze grafiche: propongono una rappresentazione grafica con diagrammi a barre orizzontali delle

frequenze associate ai singoli valori di una variabile. L'asse y del diagramma corrisponde ai valori e l'asse x alle frequenze che vengono scalate in modo automatico per rappresentare tutte le frequenze osservate;

- MINIMUM(n) e MAXIMUM(n) agiscono sull'asse y (dei valori) e precisano che il diagramma deve escludere i valori inferiori al minimo indicato e/o superiori al massimo indicato. INCREMENT(n) permette di specificare il valore dell'ampiezza dell'intervallo da usare nella rappresentazione;
- FREQ(n) e PERCENT(n) agiscono invece sull'asse x (delle frequenze) e rappresentano il valore massimo della scala da utilizzare, indicato rispettivamente in frequenza o in percentuale;
- NORMAL e NONNORMAL permettono di sovrapporre all'istogramma (HISTOGRAM) la curva gaussiana per poter stimare graficamente l'approssimazione alla distribuzione normale;
- NTILES=n chiede di stampare i valori che cadono sui punti di divisione in n gruppi. Il valore 4 corrisponderà ai valori dei quartili, 10 ai decili e 100 ai centili;
- PERCENTILES chiede invece di stampare i valori corrispondenti a determinati valori percentili. I valori 25 50 75 corrispondono ai quartili;
- STATISTICS seguito da una delle keyword indicate permette di ottenere alcuni dei principali indici statistici. MEAN la media aritmetica, STDDEV la deviazione standard, MINIMUM il minimo, MAXIMUM il massimo, SEMEAN l'errore standard della media, VARIANCE la varianza, SKEWNESS l'asimmetria, SESKEW l'errore standard della simmetria, RANGE la differenza fra il valore massimo e quello minimo e corrisponde all'estensione dei valori, MODE la moda, KURTOSIS la curtosi o "piattezza" della curva, SEKURT l'errore standard della curtosi, MEDIAN la mediana e SUM la sommatoria dei valori.

Un esempio dell'uso di questo comando è riportato in figura 1.37.

Descriptives

Questa procedura è appropriata con variabili di tipo quantitativo (scale a intervallo o a rapporto). Calcola un buon numero di indici statistici descrittivi della tendenza centrale, della dispersione e della forma della distribuzione di una variabile singola: media, varianza, deviazione standard, curtosi, asimmetria...

La sintassi è come quella del comando frequencies. Se si indicano variabili di tipo carattere, SPSS visualizza un messaggio d'avvertimento e non effettua calcoli, mentre se si includono variabili di tipo dicotomico o nominale, SPSS effettua lo stesso i calcoli.

Le opzioni si riferiscono al formato dei risultati, al modo di trattare i valori mancanti, alla possibilità di registrare i valori standardizzati (punti Z).

Le statistiche corrispondono in larga parte a quelle disponibili nel comando FREQUENCIES (media, errore standard della media, deviazione standard, varianza, curtosi, asimmetria, estensione dei valori (gamma), minimo, massimo, somma).

Un esempio dell'uso di questo comando è riportato in figura 1.37.

Crosstabs

Con la procedura CROSSTABS si costruiscono tabelle di contingenza a doppia, tripla o multipla entrata con eventuale calcolo delle percentuali per riga, colonna o sul totale generale della tabella. Si possono calcolare le frequenze attese in ciascuna cella e i residui rispetto alle frequenze osservate nonché altri indici statistici di associazione come il chi quadrato o il coefficiente di contingenza.

Pu essere considerata la procedura corrispondente a FREQUENCIES, ma dedicata all'analisi contemporanea di più variabili.

La sintassi è come quella descritta per i comandi precedenti, l'unica differenza è che i nomi delle variabili che sono introdotte nell'analisi devono essere separate da BY. I nomi delle variabili prima di BY indicano le variabili che devono essere utilizzate come righe della tabella da generare, mentre le variabili indicate dopo il primo BY sono le variabili da usare per le colonne, ovviamente, per ogni variabile indicata viene generata una tabella specifica. I BY successivi al primo, permettono di generare tabelle a 3 o più entrate, secondo il numero di BY utilizzati.

Anche con questo comando è possibile definire delle opzioni che servono per selezionare il contenuto che dev'essere visualizzato in ogni cella. Senza alcuna indicazione di OPTIONS viene visualizzata solo la frequenza dei casi che presentano contemporaneamente entrambi i valori indicati in quella riga e in quella colonna, le opzioni previste sono le seguenti: inclusione dei valori mancanti; eliminazione di tutte le etichette; visualizzazione della percentuale di riga, di colonna e totale; eliminazione delle etichette dei valori; ordine delle righe in modo decrescente di valore; eliminazione di tabelle e frequenze; visualizzazione delle frequenze teoriche, dei residui del chi quadro, dei residui stand. del chi quadro, dei residui stand ponderati del chi quadro.

L'indicazione delle statistiche tramite il comando STATISTICS permette di ottenere alcune delle principali statistiche specifiche delle tabelle di contingenza come: il Chi quadro o il Coefficiente di contingenza.

La sintassi relativa a tutti i comandi descritti in questo capitolo è disponibile sulla guida in linea di SPSS.

Un esempio dell'uso di questo comando è riportato in figura 1.37.

```
DATA LIST FILE = 'C:\DATINORDSUD.DAT' FREE
/ ID (F2)
ISTRUZ (F3)
RESID (F1)
ETA (F2)
REDDITO (F3).

VARIABLES LABELS ID "Identificatore"
/ISTRUZ "Livello di istruzione"
RESID "Residenza"
ETA "Età"
REDDITO "Livello del reddito".

VALUE LABELS
RESID
0 "Sud"
1 "Nord".

COMPUTE REDEUR = REDDITO/1.936.

COMPUTE CLREDEUR=REDEUR.

RECODE CLREDEUR ( LO THRU 21.31 = 1) (21.31 THRU 34.87 =2) (34.87 THRU 48.42 = 3) (48.42 THRU HI =4).

FORMATS REDEUR (F2.2) CLREDEUR (F1).

VARIABLES LABELS REDEUR "Reddito in euro"
/CLREDEUR "Reddito raggruppato in classi".

FREQUENCIES CLREDEUR ISTRUZ
/MISSING=INCLUDE
/STATISTICS= MEAN STDDEV MINIMUM MAXIMUM SEMEAN VARIANCE SKEWNESS MODE KURTOSIS MEDIAN
/HISTOGRAM=NORMAL.

DESCRIPTIVES REDEUR ETA
/STATISTICS MEAN STDDEV MINIMUM MAXIMUM SEMEAN VARIANCE SKEWNESS KURTOSIS.

CROSSTABS RESID BY CLREDEUR
/STATISTICS CHISQ.

CROSSTABS RESID ISTRUZ BY CLREDEUR.

CROSSTABS ISTRUZ BY CLREDEUR BY RESID.
```

Figura 1.37: *Programma.*